



PROCESSAMENTO DE LINGUAGEM NATURAL

Conceitos, Técnicas
e Aplicações em
Português

Organizado por:
Helena de Medeiros Caseli
Maria das Graças Volpe Nunes

ISBN: 978-65-02-00178-3

4ª Edição | 2026

Volume 3: Aplicações e Domínios

ISBN do Volume 3: 978-65-02-00167-7

*Este livro é o produto de um sonho coletivo do grupo
Brasileiras em Processamento de Linguagem Natural (BPLN)*

© 2026 by Helena de Medeiros Caseli, Maria das Graças Volpe Nunes, organizadoras.
Editoração eletrônica: Helena de Medeiros Caseli

Licença CC BY-NC-ND 3.0 BR

Ficha Catalográfica: Marina P. Freitas CRB-08/ 6069

Dados internacionais de Catalogação-na-Publicação (CIP)

Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português / organização de Helena de Medeiros Caseli e Maria das Graças Volpe Nunes. — 4. ed. - São Carlos : BPLN, 2026. Internet: brasileiraspln.com/livro-pln-4ed/volume3.html 518p.

v. 3: Aplicações e Domínios.

ISBN: 978-65-02-00167-7

1. Processamento de Linguagem Natural 2. Inteligência Artificial. 3. I. Título.

Processamento de Linguagem Natural

Conceitos, Técnicas e Aplicações em Português

Organizado por:

Helena de Medeiros Caseli

Maria das Graças Volpe Nunes

4^a Edição

Volume 3: Aplicações e Domínios

Abril 2026



<https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/>

Conteúdo

Volume 3: Aplicações e Domínios	1
Sobre o Volume 3	1
Como citar	1
 I Aplicações e Domínios	 3
1 Classificação de texto	4
1.1 Taxonomia	5
1.2 Tipos de problemas	6
1.2.1 Classificação binária	6
1.2.2 Classificação multiclasse	7
1.2.3 Classificação multirrótulo	7
1.2.4 Classificação hierárquica	7
1.3 Representações computacionais de texto	8
1.3.1 Representações distributivas (esparsas)	8
1.3.2 Representações distribuídas (densas)	10
1.3.3 Representações contextuais (modelos de linguagem profundos)	11
1.4 Métodos de classificação	12
1.4.1 Métodos probabilísticos: Naïve Bayes	13
1.4.2 Métodos baseados em distância: K-vizinhos mais próximos	14
1.4.3 Métodos de otimização: Regressão Logística e SVM	15
1.4.4 Métodos baseados em árvores e <i>ensembles</i>	17
1.4.5 Métodos conexionistas: Redes neurais e <i>Deep learning</i>	18
1.4.6 Modelos de linguagem	19
1.4.7 Redes com atenção: Transformers	19
1.4.8 Classificação via <i>prompt</i> e modelos de linguagem de larga escala (LLMs)	20
1.4.9 Síntese e Recomendação	22
1.5 Avaliação de modelos	22
1.5.1 Metodologias de validação	22
1.5.2 Métricas de desempenho	23
1.5.3 Análise via Curva ROC e AUC	25
1.6 Recursos em português	26
1.6.1 Conjuntos de dados para categorização	26
1.6.2 Modelos de linguagem pré-treinados	27
1.6.3 Ferramentas e Bibliotecas	27
1.7 Desafios e Tendências	27
1.7.1 Dinamismo e Desvio de Conceito (<i>Concept Drift</i>)	28



1.7.2	Escalabilidade e Eficiência energética	28
1.7.3	Explicabilidade e Interpretabilidade (XAI)	28
1.7.4	Ética, Equidade (<i>Fairness</i>) e Mitigação de vieses	28
1.7.5	Soberania digital	29
1.7.6	Categorização em baixos recursos e dialetos	29
1.8	Referências para aprofundamento	29
2	Recuperação de Informação	30
2.1	Introdução	30
2.1.1	Relação com o PLN	30
2.1.2	O Foco da Recuperação de Informação	30
2.1.3	O Conceito de Relevância	31
2.1.4	Principais Livros	31
2.1.5	Organização deste Capítulo	32
2.2	Visão Geral de um Sistema Típico de Recuperação de Informação	32
2.2.1	Pré-Processamento	32
2.2.2	Indexação	35
2.3	Modelos Clássicos de recuperação de informação	36
2.3.1	Modelo Booleano	36
2.3.2	Modelo Vetorial	36
2.3.3	Modelos Probabilísticos	40
2.4	Avaliação da Qualidade de Sistemas de recuperação de informação	42
2.4.1	Métricas	42
2.4.2	Coleções de Teste	46
2.5	Modificação Automática de Consultas	49
2.6	Ferramentas e Bibliotecas	50
2.6.1	Sistemas Comerciais	50
2.6.2	Ferramentas para Pesquisa	50
2.7	Conclusão	51
	Agradecimentos	51
3	Extração de Informação	52
3.1	Introdução	52
3.2	Um pouco de história	53
3.3	Conceituação formal: Relação e Entidade	55
3.3.1	Entidade	55
3.3.2	Relação	55
3.4	Extração de Informação (EI)	56
3.4.1	Reconhecimento de Entidades Nomeadas	56
3.4.2	Extração de Relações	57
3.4.3	Extração Conjunta de Entidades e Relações	58
3.4.4	Métodos empregados para EI na literatura	58
3.5	Extração de Informação Aberta	62
3.5.1	Formalização	63
3.5.2	Abordagens	63
3.6	Avaliação	65
3.7	Considerações finais	67
	Agradecimentos	67



4	Tradução Automática	68
4.1	Introdução	68
4.2	Abordagens	69
4.2.1	Tradução direta	69
4.2.2	Tradução Automática Baseada em Regras	70
4.2.3	Tradução por interlíngua	71
4.2.4	Tradução Automática Baseada em Exemplos	72
4.2.5	Tradução Automática Estatística	73
4.2.6	Tradução Automática Neural	76
4.3	Avaliação da Tradução Automática	79
4.3.1	O que é Avaliação da Tradução Automática?	80
4.3.2	A importância de uma avaliação replicável	83
4.3.3	Métricas Automáticas	84
4.3.4	Métricas Humanas	86
4.3.5	Avaliação dependente de contexto	90
4.4	O Futuro da Tradução Automática	92
5	Sumarização Automática	94
5.1	Introdução	94
5.2	Por trás do processo	96
5.2.1	Sobre conceitos	96
5.2.2	Sobre o processo de sumarização	98
5.3	Avaliação na SA	100
5.4	Aplicação em Sumarização Extrativa	103
5.4.1	Algoritmo baseado na frequência das palavras	104
5.4.2	Algoritmo baseado em Transformers	107
5.4.3	Avaliando os sumários	109
5.5	Aplicação em Sumarização Abstrativa	111
5.5.1	Algoritmos baseados em estrutura sintática e/ou semântica	112
5.5.2	Algoritmos baseados em Transformers	113
5.5.3	Sumarização abstrativa na prática	115
5.6	A Sumarização automática e a língua portuguesa	120
5.7	Considerações finais	121
6	Complexidade Textual e suas Tarefas Relacionadas	123
6.1	Introdução	123
6.2	Complexo para quem?	124
	Por que não ensinar a ler em vez de simplificar um texto?	124
6.3	Tarefas de PLN relacionadas à complexidade textual	125
6.3.1	Adaptação textual	125
6.3.2	Simplificação textual	126
6.3.3	Simplificação lexical	127
6.3.4	Simplificação sintática	128
6.3.5	Sumarização automática	128
6.3.6	Elaboração textual	129
6.4	Avaliação da Complexidade e suas Métricas	129
6.4.1	Fórmulas clássicas	130
6.4.2	Métricas Linguísticas	132



6.4.3	Métricas Psicolinguísticas	137
6.4.4	NILC-Metrix	139
6.4.5	Rastreamento ocular	140
6.5	Recursos para Complexidade Textual e Sentencial no Português Brasileiro	142
6.5.1	Recursos para Complexidade Textual	142
6.5.2	Recursos para Complexidade Sentencial	146
6.6	Uso de Modelos de Língua para a Simplificação Textual	146
6.7	A complexidade como parâmetro de descrição da linguagem	148
6.7.1	Questões sobre a complexidade como parâmetro	150
6.7.2	Métricas de Complexidade Baseadas em Compressibilidade	151
6.8	Considerações finais	156
7	Correção Automática de Redação	157
7.1	Introdução	157
7.1.1	O que é uma redação escolar?	158
7.1.2	O que é avaliado?	159
7.1.3	Alguns modelos brasileiros de correção	160
7.2	Detecção de desvios no texto	162
7.2.1	Tipos de desvios	163
7.2.2	Formalismos de regras	165
7.3	Atribuição de nota	168
7.3.1	Como atribuir nota a redações?	168
7.3.2	Atribuição de nota para redações do Enem	171
7.4	Feedback para o aluno	172
7.4.1	Estatísticas básicas do texto	173
7.4.2	Assistentes de escrita e ferramentas de auxílio à escrita	174
7.4.3	Identificação de pontos fortes e elogiáveis	176
7.5	Correção manual vs(?) correção automática	178
7.5.1	Avanços dos últimos anos	178
7.5.2	Vantagens da correção automática	179
7.5.3	Vantagens da correção manual	179
7.5.4	O exemplo da língua inglesa	180
7.5.5	O que defendemos	181
7.6	Considerações finais	181
8	Detecção Automática de Notícias Falsas	182
8.1	Introdução	182
8.1.1	Conteúdo enganoso: definições	183
8.1.2	Notícias falsas no tempo e no espaço	184
8.1.3	Combate ao conteúdo enganoso	186
8.2	<i>Corpora</i>	187
8.3	Métodos de detecção automática	189
8.3.1	Aprendizado de máquina com atributos linguísticos	190
8.3.2	Modelos de linguagem	191
8.3.3	Checação de conteúdo	194
8.3.4	Recuperação de informação	195
8.3.5	Análise de credibilidade	195
8.3.6	Outras iniciativas relacionadas	196



8.4	Impactos para a sociedade	197
8.5	Considerações finais	200
8.6	Exercícios de fixação	201
9	PLN na Saúde	203
9.1	Introdução	203
9.2	O texto livre em narrativas clínicas	205
9.3	Aplicações de PLN na Saúde	207
9.3.1	Predição	207
9.3.2	Desidentificação	207
9.3.3	Extração de conceitos clínicos	208
9.3.4	Relações temporais	209
9.3.5	Sumarização	210
9.4	Recursos de PLN na área clínica	211
9.4.1	<i>Corpora</i>	211
9.4.2	Modelos	212
9.5	Para onde estamos caminhando?	213
10	Deteccção de Transtornos de Saúde Mental a partir de Texto	215
10.1	Introdução	215
10.1.1	Definição do problema	216
10.2	Conjuntos de dados para deteção de depressão/ansiedade	218
10.3	Modelos computacionais de deteção de depressão/ansiedade	221
10.4	SetembroBR: deteção de depressão/ansiedade em português	223
10.4.1	<i>Corpus</i> SetembroBR	224
10.4.2	Modelos desenvolvidos no projeto SetembroBR	225
10.5	Amive: deteção de sintomas de depressão em português	229
10.5.1	Sinais associados à depressão	229
10.5.2	<i>Corpus</i> do projeto Amive e anotação manual	232
10.5.3	Identificação automática de sinais de depressão	233
10.6	Considerações finais	233
	Agradecimentos	234
11	Classificação de Áudio aplicada à Saúde	235
11.1	Introdução	235
11.1.1	Deteccção de COVID-19 através da voz e tosse	236
11.1.2	Insuficiência Respiratória e seus efeitos na voz e fala	237
11.1.3	Outros problemas respiratórios	238
11.1.4	Organização do capítulo	239
11.2	<i>Corpora</i> de áudios	239
11.2.1	<i>Corpus</i> de deteção de insuficiência respiratória	239
11.2.2	<i>Corpus</i> de deteção de COVID-19	242
11.2.3	<i>Corpora</i> para deteção de outros problemas respiratórios	244
11.3	Métodos de deteção de problemas respiratórios através de áudios	244
11.3.1	Pré-processamento de áudios	244
11.3.2	Modelos de aprendizado profundo para classificação de áudios	246
11.3.3	Avaliação dos modelos de deteção de problemas respiratórios	250
11.4	Considerações finais e impactos para a sociedade	251



12 PLN no Direito	254
12.1 Introdução	254
12.2 O Direito – uma moldura para a significação	255
12.3 Entre as terminologias e as palavras no Direito do Brasil	257
12.4 Um caso concreto: em pequeníssima escala	258
12.5 Outros casos/exemplos: Direito Ambiental	260
12.6 Aplicação: Análise de Sentimentos em Direito: desafios e exemplos	262
12.7 Considerações finais	266
13 Reconhecimento de Entidades Nomeadas no Domínio Legal	268
13.1 Introdução	268
13.2 Contextualização	268
13.3 Iniciativas de REN-Legal	270
13.4 Iniciativas de REN-Legal em língua portuguesa	273
13.4.1 <i>Corpora</i> para REN-Legal	273
13.4.2 Modelos de linguagem neurais para REN-Legal	279
13.5 Modelos de linguagem de grande escala (LLM) para REN	281
13.5.1 LLM para REN-Legal em língua portuguesa	282
13.6 Aplicação de REN em sistema de recuperação de informação legislativa	284
13.7 Considerações finais	285
Agradecimentos	286
14 Geração Automática de Ementas Jurídicas	287
14.1 Introdução	287
14.1.1 Contextualização	287
14.1.2 Organização do Capítulo	289
14.2 Conceitos Preliminares	289
14.2.1 Transferência de Aprendizado e Ajuste Fino	289
14.2.2 <i>Bilingual Evaluation Under-study</i> (BLEU)	289
14.3 Geração de verbetações	290
14.3.1 Aquisição de dados	291
14.3.2 Geração textual	291
14.4 Avaliação utilizando RI	292
14.4.1 Formulação da tarefa	293
14.4.2 Experimentos realizados	293
14.4.3 Métodos de RI avaliados	294
14.4.4 Ordenação dos documentos	295
14.4.5 Métricas avaliadas	295
14.5 Resultados e discussões	295
14.5.1 Comparação entre verbetações originais e geradas	295
14.5.2 Experimentos com RI	297
14.6 Conclusão	299
Agradecimentos	299
15 PLN e Humanidades Digitais	300
15.1 Introdução	300
15.2 Preparação das fontes	301
15.2.1 Digitalização	302



15.2.2	Metadados	302
15.2.3	Normalização	303
15.2.4	Anotação de textos	304
15.3	Transformação de textos em dados	305
15.3.1	Extração de informação	305
15.3.2	Representação de conhecimento e ontologias	306
15.3.3	Dados ligados e dados <i>FAIR</i>	306
15.4	Exemplos de projetos de HD envolvendo o processamento da língua portu- guesa	308
15.4.1	O <i>corpus</i> Tycho Brahe	308
15.4.2	As Memórias Paroquiais	308
15.4.3	O projeto MONSOON	309
15.4.4	A História do Futuro	310
15.4.5	Manuais médicos do século XVIII	311
15.5	Considerações finais	312
	Agradecimentos	313
16	PLN em Redes Sociais	314
16.1	Introdução	314
16.2	Redes Sociais	314
16.2.1	Facebook	315
16.2.2	Reddit	316
16.2.3	Youtube	316
16.2.4	Twitter/X	316
16.2.5	Whatsapp	316
16.2.6	Instagram	316
16.3	Áreas de Aplicação	317
16.3.1	Deteção de Discurso de Ódio e Linguagem Ofensiva	317
16.3.2	Análise de Sentimento	322
16.3.3	Deteção de Notícias Falsas	325
16.3.4	Deteção de Ironia/Sarcasmo/Humor	327
16.4	Considerações finais	329
17	PLN e Responsabilidade Social	330
17.1	Introdução	330
17.2	Semântica de <i>Frames</i>	331
17.3	FrameNet Brasil	333
17.3.1	As relações <i>frame-a-frame</i>	336
17.3.2	As relações qualia ternárias	339
17.4	Projeto DataToStopGBV: Semântica contra a violência de gênero	341
17.4.1	Contexto e motivação	342
17.4.2	A contribuição da FrameNet para identificação da Violência Baseada em Gênero	345
17.4.3	Avaliação da identificação da Violência Baseada em Gênero	350
17.4.4	Impacto Social e técnico	353
17.5	Multimodalidade, acessibilidade cultural e sentido compartilhado no Audi- tion: <i>dataset</i> multimodal de curtas-metragens acessíveis	355
17.5.1	O <i>dataset</i> Audition	356



17.5.2	Anotação semântica multimodal	360
17.5.3	Validação, Aplicações em PLN e Impactos na Tradução Audiovisual Acessível	364
17.6	Considerações finais	366
	Agradecimentos	367
	Glossário deste capítulo	367
18	Redes complexas no processamento de língua natural do português	370
18.1	Introdução	370
18.2	Redes complexas	371
18.2.1	Conceitos fundamentais	372
18.2.2	Propriedades	373
18.2.3	Tipos de redes	375
18.2.4	Dinâmica	378
18.3	Detecção de comunidades	379
18.4	Aplicações de redes complexas em PLN para o português	381
18.5	Exemplo de aplicação: Panorama de PLN em língua portuguesa	385
18.5.1	Redes de citações e similaridade	386
18.5.2	Identificação de tópicos em redes de pesquisa	387
18.5.3	Pesquisa em PLN para o português	388
18.6	Considerações finais	395
	Referências	397
	Apêndices	486
A	Considerações prévias à sumarização automática	486
B	Exemplo de sumarização abstrativa	490
	Sobre as/os autoras/es	493
	Helena de Medeiros Caseli	493
	Maria das Graças Volpe Nunes	493
	Adriana Silvina Pagano	493
	Adriano L. I. Oliveira	493
	Aline Macohin	494
	Amanda Pontes Rassi	494
	Ana Paula Banza	494
	Ana Sofia Ribeiro	494
	André Carlos Ponce de Leon Ferreira de Carvalho	494
	Beatriz Raposo de Medeiros	495
	Brenda Salenave Santana	495
	Cassia Trojahn dos Santos	495
	Celso Ricardo Fernandes de Carvalho	495
	Claudia Maria Cabral Moro Barra	496
	Crysttian Arantes Paixão	496
	Daniela Barreiro Claro	496
	Danilo Samuel Jodas	496



Diego Raphael Amancio	496
Douglas Rodrigues	497
Elisa Terumi Rubel Schneider	497
Ellen Souza	497
Evelin Amorim	497
Fátima Farrica	498
Felipe Ribas Serras	498
Fernanda Olival	498
Flaviane Romani Fernandes Svartman	498
Gabriel Lino Garcia	498
Helena Freire Cameron	499
Hidemberg O. Albuquerque	499
Ivandr� Paraboni	499
Jackson Wilke da Cruz Souza	499
Jo�o Paulo Papa	500
Jo�o Renato Ribeiro Manesco	500
Jo�o Vitor Mariano Correia	500
Joaquim Santos	500
Kenzo Sakiyama	500
Larissa Cristina Berti	501
Larissa A. de Freitas	501
Lilian Mie Mukai Cintho	501
L�via Vicente Dutra	501
Lucas Daniel Volpiano Figueiredo	501
Marcelo Finger	502
Marcelo Matheus Gauy	502
Marcelo Gomes de Queiroz	502
Maria Cristina Ferreira de Oliveira	502
Maria Jos� Bocorny Finatto	502
Mariane de Carvalho Pinto	503
Marlo Souza	503
Matheus Cerqueira	503
Maucha Andrade Gamonal	503
N�dia F�lix Felipe da Silva	503
Oswaldo Novais de Oliveira Jr.	504
Paula Christina Figueira Cardoso	504
Pedro Henrique Paiola	504
Priscilla de Abreu Lopes	504
Rafael Oleques Nunes	504
Raphael Montanari	505
Renata Vieira	505
Renato Moraes Silva	505
Rodrigo Nogueira	505
Roney Lira de Sales Santos	505
Roseli A. F. Romero	506
Roseval Donisete Malaquias Junior	506
Sandra Maria Al�sio	506
Sheila Castilho	506



Conteúdo

Sidney Evaldo Leal	507
Thiago Alexandre Salgueiro Pardo	507
Tiago Agostinho de Almeida	507
Tiago Timponi Torrent	507
Vitor Hugo Bridi Dei Santi	507
Viviane Pereira Moreira	508
Yohan Bonescki Gumiel	508



Volume 3: Aplicações e Domínios

Sobre o Volume 3

Dos dezoito capítulos deste volume, oito ilustram o PLN em várias aplicações: classificação de textos (Capítulo 1), recuperação (Capítulo 2) e extração (Capítulo 3) de informações, tradução (Capítulo 4) ou sumarização (Capítulo 5) automáticas, avaliação de complexidade textual (Capítulo 6) e correção automática de redações (Capítulo 7) e detecção de notícias falsas (*fake news*) (Capítulo 8).

Três capítulos ilustram o PLN aplicado ao domínio da saúde: o Capítulo 9 discorre sobre as tarefas de PLN comuns a esse domínio; o Capítulo 10 traz uma aplicação na área de saúde mental; e o Capítulo 11 mostra uma aplicação do processamento de fala para detecção de problemas respiratórios. Já no domínio jurídico ou do direito, o Capítulo 12 traz os principais conceitos e desafios para o PLN, sendo que o tratamento das entidades nomeados neste domínio é abordado no Capítulo 13, e a geração automática de ementas jurídicas é apresentada no Capítulo 14.

O Capítulo 15 discorre sobre a potencialidade do uso de PLN na ampla área de humanidades digitais, caracterizada por trabalhos derivados das ciências humanas que abordam a linguagem. O Capítulo 16 introduz o gênero especial da linguagem usada nas redes sociais, enquanto que o Capítulo 17 ilustra uma aplicação do formalismo Framenet no enfrentamento à violência de gênero e na promoção da acessibilidade cultural. Finalmente, o Capítulo 18 explora o potencial do formalismo de redes complexas para modelar textos para aplicações de PLN, ilustrando com o mapeamento das pesquisas em PLN para português.

A leitura dos capítulos que integram esse volume deve expandir o horizonte dos leitores para as variadas áreas de aplicação do PLN, e mostrar-lhes a potencialidade do emprego de PLN quando ajustado a um domínio de conhecimento, um gênero específico ou uma aplicação em particular.

Como citar

Caseli, H.M.; Nunes, M.G.V. (org.) **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. 4 ed. BPLN, 2026. v. 3. ISBN 978-65-02-00167-7. Disponível em: <https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3>.

Note 1: Bibref



```
@Book{BPLN_livro_4ed_vol3:2026,  
  title      = {Processamento de Linguagem Natural: Conceitos,  
                Técnicas e Aplicações em Português},  
  editor      = {Caseli, H. M. and Nunes, M. G. V.},  
  year        = {2026},  
  publisher    = {BPLN},  
  edition     = {4},  
  volume      = {3},  
  isbn        = {978-65-02-00167-7},  
  url         = {https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3}  
}
```



Parte I

Aplicações e Domínios



Capítulo 1

Classificação de texto

*Evelin Amorim
Tiago A. Almeida*

Publicado em: 16/04/2026

A **classificação de texto** é uma das tarefas fundamentais e mais estudadas no Processamento de Linguagem Natural (PLN). Em sua essência, a tarefa consiste em atribuir um ou mais rótulos ou categorias a um documento de texto, com base em seu conteúdo. Essa tarefa, aparentemente simples, é a base para uma vasta gama de aplicações que encontramos em nosso dia a dia como: a organização automática de e-mails em pastas (por exemplo, “Principal”, “Social”, “Promoções”), a identificação do sentimento em uma avaliação de produto (positivo, negativo, neutro) ou a classificação de notícias por tópicos (como “Esportes”, “Política”, “Tecnologia”). Como resultado desta tarefa temos uma coleção de textos organizada de acordo com critérios humanos relevantes para um propósito como o de auxiliar as pessoas a localizar rapidamente a informação de que precisam. Esta tarefa pode, portanto, ser a primeira etapa para as pessoas catalogarem informação. É por este motivo que esta tarefa também pode ser chamada de categorização de texto (Sebastiani, 2002).

A importância da categorização de texto reside em sua capacidade de estruturar a imensa quantidade de dados textuais não estruturados gerados a cada segundo. Como descrito no Capítulo **O que é PLN?**, um dos grandes desafios da área é justamente criar métodos que permitam aos computadores “entender” e processar a linguagem humana para extrair informações úteis. A categorização é um passo crucial nesse processo, transformando texto livre em informação organizada e facilmente acessível.

A separação de textos em categorias diferentes pode parecer uma tarefa simples à primeira vista; porém, estabelecer as categorias mais adequadas dentro de um conjunto, definir os nomes das categorias, entre outros critérios a serem estabelecidos para os dados, pode consumir muito tempo e demandar conhecimento específico (Freitas Araujo, 2018). Por isso, existe um campo de estudo dedicado a esta organização, chamado Ciência da Informação, que pode abrigar cursos sobre diferentes tipos de catalogação e organização da informação, como Biblioteconomia e Arquivologia, entre outros. Essa área surgiu após a Segunda Guerra Mundial, com o avanço e o desenvolvimento tecnológico (Saracevic, 1996).

Podemos considerar, portanto, a separação de textos em categorias pré-definidas como uma subárea desta grande área, a Ciência da Informação. Com o surgimento da Inteligência Artificial (IA), também por volta da década de 1950 (Haenlein; Kaplan, 2019), foi natural associá-la à crescente quantidade de informação gerada pelos avanços tecnológicos. Assim, a classificação de texto foi uma das primeiras tarefas exploradas em IA por volta de 1960. Maron (1961) propôs classificar um conjunto de resumos de artigos em 32 categorias utilizando como características preditivas as palavras presentes nos resumos do



conjunto de treino. Desde então, muitos avanços foram alcançados. Os trabalhos que se seguiram mostraram o lado multifacetado da classificação de texto, por exemplo, aplicando diferentes estratégias de IA para classificar textos de saúde ou jurídicos, ou definindo que um documento pode ser classificado em mais de uma categoria. Com relação ao domínio textual, referimos ao leitor os Capítulos 10 e 8 para a classificação de textos no domínio da saúde mental e a detecção de notícias falsas, respectivamente.

Neste capítulo, exploraremos os múltiplos aspectos da categorização de texto, desde seus fundamentos teóricos até suas aplicações práticas, com foco especial nos recursos disponíveis para o português. A seguir, definiremos a taxonomia à qual um conjunto de dados de classificação de texto pode pertencer.

1.1 Taxonomia

O termo “categorização de texto” é frequentemente usado como um guarda-chuva para uma variedade de problemas específicos. Podemos organizar as tarefas de categorização com base em diferentes critérios, como a natureza das categorias, a relação entre elas e o objetivo final da aplicação.

Uma primeira grande divisão se dá entre a **classificação de tópicos** (ou assuntos) e a **classificação por metadados**. A **classificação de tópicos** (*topic classification*) é a forma mais tradicional de categorização, cujo objetivo é identificar o assunto principal de um texto. Por exemplo, um artigo de jornal pode ser classificado como “Ciência” e uma publicação em rede social como “Viagem”. As categorias são definidas pelo conteúdo semântico do documento. Na **classificação por metadados**, as categorias não se referem ao tópico do texto, mas a um atributo ou metadado associado a esse tópico. Exemplos incluem:

- **Análise de sentimentos:** classificar a polaridade de um texto (e.g., positivo, negativo, neutro), uma tarefa explorada em aplicações de redes sociais, como discutido no Capítulo 16.
- **Detecção de autoria:** identificar o autor de um texto com base em seu estilo de escrita (Varela et al., 2011).
- **Detecção de spam:** diferenciar e-mails legítimos de spam (Almeida; Yamakami, 2012).
- **Detecção de notícias falsas:** classificar uma notícia como legítima ou falsa, uma aplicação de grande impacto social abordada no Capítulo 8.

Neste capítulo não trataremos deste tipo de classificação, porém essa distinção é crucial, pois os tipos de características (*features*) relevantes para cada tarefa podem variar drasticamente. Enquanto a classificação de tópicos depende muito do vocabulário específico do domínio, a análise de sentimentos pode depender de palavras com carga afetiva e a detecção de autoria, de padrões estilísticos sutis.

A estrutura e a relação entre as categorias também definem diferentes tipos de problemas, como veremos em detalhe na próxima seção, abordando desde a simples classificação binária até cenários mais complexos, como a classificação hierárquica e multirrótulo.



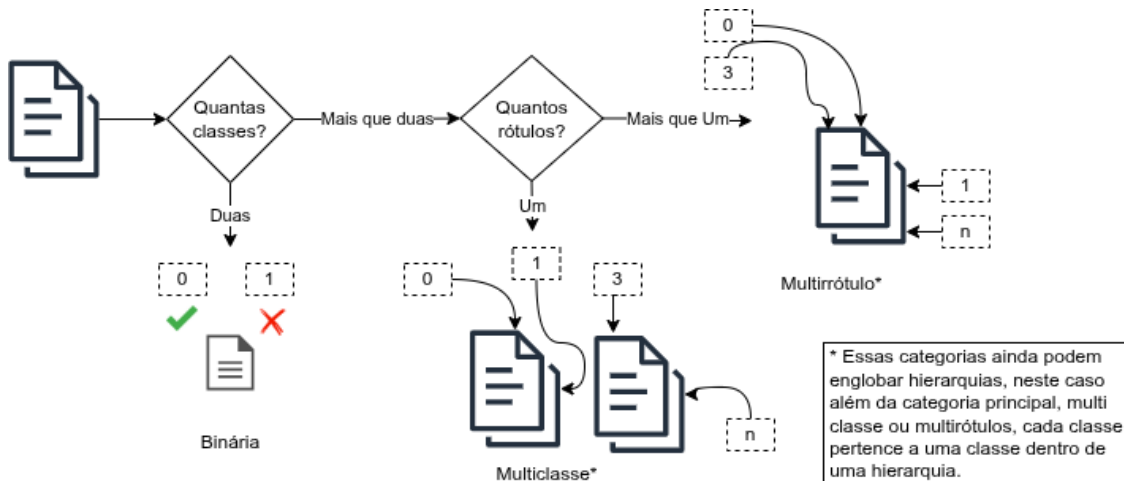
1.2 Tipos de problemas

A taxonomia apresentada na seção anterior descreve o propósito da categorização. Agora, vamos formalizar os tipos de problemas com base na estrutura do conjunto de categorias C (também chamado de espaço de rótulos ou *label space*). A definição do problema tem impacto direto na forma como o modelo é projetado, treinado e avaliado (como veremos nas Seções 1.4 e 1.5).

Para treinar e avaliar os modelos que executam essas tarefas, dependemos fundamentalmente de conjuntos de dados anotados, os chamados *corpora*, um conceito central detalhado no Capítulo [Conjunto de dados, dataset e corpus](#). A estrutura e a relação entre as **classes (categorias ou rótulos)** também definem diferentes tipos de problemas. A identificação do tipo de classificação de texto é importante para levantar os principais desafios e as melhores estratégias para solução. Para isso, é necessário primeiro identificar o número de classes na tarefa e, em seguida, o número de rótulos por instância.

A Figura 1.1 apresenta um fluxograma que guia a identificação do tipo de classificação a que uma categorização textual pertence. Seja D um conjunto de documentos e $C = c_1, c_2, \dots, c_K$ um conjunto de K categorias (rótulos) predefinidas. O objetivo de um classificador f é aprender uma função de mapeamento $f : D \rightarrow C'$ que associa um documento $d \in D$ a um subconjunto de rótulos $C' \subseteq C$. A natureza desse mapeamento define os problemas descritos a seguir (Aggarwal, 2018; Jurafsky; Martin, 2023).

Figura 1.1: Fluxograma dos tipos de problemas de classificação de texto.



1.2.1 Classificação binária

É o cenário mais simples e fundamental de categorização. O conjunto de categorias contém exatamente duas classes, $K = 2$ (ou seja, $C = \{c_1, c_2\}$), que são mutuamente exclusivas. Neste caso, o classificador aprende uma função $f : D \rightarrow \{c_1, c_2\}$. Frequentemente, as classes são representadas como $\{+1, -1\}$ (positivo/negativo) ou como $\{0, 1\}$ (ausente/presente).

Exemplos clássicos: incluem detecção de *spam*, na qual um e-mail é classificado como “spam” ou “legítimo” (*ham*); análise de sentimentos (polaridade), na qual uma avaliação de produto ou serviço é “positiva” ou “negativa”; e detecção de notícias falsas, na qual uma notícia pode ser classificada como “legítima” ou “falsa”.

1.2.2 Classificação multiclasse

O conjunto de categorias contém mais de duas classes ($K > 2$), mutuamente exclusivas. Cada documento d deve estar associado a exatamente uma categoria do conjunto C . Neste caso, o classificador aprende uma função $f : D \rightarrow C$, com $|C| = K > 2$.

Exemplos tradicionais incluem: classificação de tópicos de notícias, no qual um artigo de notícia é atribuído a um único assunto, como “Esporte”, “Política” ou “Tecnologia”; análise de sentimentos (granular), na qual uma avaliação de produto ou serviço é classificada como “muito positiva”, “positiva”, “neutra”, “negativa” ou “muito negativa”; e identificação de idioma, que determina se um texto está escrito em “Português”, “Inglês”, “Espanhol” etc (Silva et al., 2017b).

1.2.3 Classificação multirrótulo

A classificação multirrótulo rompe a restrição da exclusividade mútua. As categorias não são mutuamente exclusivas, e um documento d pode ser associado a zero, a um ou a múltiplos rótulos simultaneamente (Tsoumakas; Katakis, 2007).

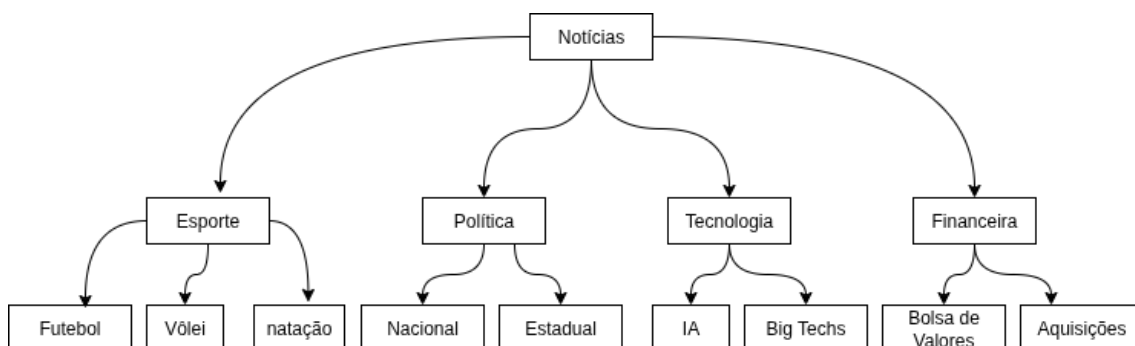
Exemplos incluem: categorização de artigos científicos, no qual um estudo sobre aprendizado de máquina aplicado à medicina pode receber os rótulos “Inteligência Artificial”, “PLN” e “Saúde”; classificação de conteúdo jurídico, no qual uma peça processual pode ser relevante para múltiplos ramos do direito, como “Direito Civil” e “Direito do Consumidor”; e marcação de publicações em redes sociais, no qual um *post* sobre uma viagem gastronômica pode ser marcado com *tags* como “#Viagem”, “#Gastronomia” e “#Férias”.

Este problema é significativamente mais complexo, pois o espaço de saída cresce exponencialmente (2^K combinações possíveis) e exige métodos específicos, como a Relevância Binária (*Binary Relevance*) ou algoritmos adaptados que modelam a correlação entre os rótulos (Bittencourt et al., 2020).

1.2.4 Classificação hierárquica

Este é um caso especial de classificação (que pode ser multiclasse ou multirrótulo), em que as próprias categorias estão organizadas em uma estrutura hierárquica, como uma árvore ou um grafo acíclico dirigido. As categorias em C mantêm relações de super-classe/subclasse (pai/filho). A decisão de classificar um documento em uma categoria mais específica (e.g., “Futebol”) pode depender da classificação na categoria mais geral (e.g., “Esporte”) (Figura 1.2).

Figura 1.2: Exemplo de hierarquia de classes em uma tarefa de classificação de notícias.



Exemplos incluem classificação de produtos, na qual um item em um *e-commerce* pode ser classificado em “Eletrônicos” → “Áudio e vídeo” → “Fones de ouvido”; classificação de documentos médicos, na qual registros de pacientes ou artigos científicos são classificados de acordo com o uso de taxonomias como a CID (Classificação Internacional de Doenças¹) ou a MeSH (*Medical Subject Headings*²); e taxonomias biológicas, nas quais uma espécie de ser vivo pode ser classificada em Reino, Filo, Classe, Ordem etc.

Abordagens para este problema podem ser planas (*flat*), ignorando a hierarquia (i.e., tratando-a como um problema multiclasse/multirrotulo padrão), ou hierárquicas, que utilizam a estrutura da taxonomia para guiar o processo de classificação (Silla; Freitas, 2011; Tavares et al., 2018). Neste caso, a classificação hierárquica pode ser resolvida com abordagens *top-down* (vários classificadores em cascata) ou *big-bang* (um único classificador que conhece a estrutura).

Compreender qual desses problemas estamos resolvendo é o primeiro passo para selecionar a representação de texto (Seção 1.3), o método de classificação (Seção 1.4) e as métricas de avaliação (Seção 1.5) mais adequados.

1.3 Representações computacionais de texto

Modelos de aprendizado de máquina, como os que veremos na Seção 1.4, não operam sobre texto puro; exigem entradas numéricas. O processo de conversão de um texto (uma sequência de caracteres) em um vetor numérico — seja ele esparsos ou denso — é chamado de representação de texto ou *feature extraction* (extração de características).

A qualidade dessa representação é fundamental, pois ela deve capturar a informação semântica e discriminativa do documento, filtrando ruídos de modo que o classificador possa encontrar padrões. A evolução das representações de texto é um pilar central da própria evolução do PLN, passando de modelos baseados em contagem a modelos que aprendem o significado a partir do contexto.

1.3.1 Representações distributivas (esparsas)

A primeira abordagem bem-sucedida, que serviu de base para a Recuperação de Informação (ver Capítulo 2) por décadas, é baseada na Hipótese Distributiva, que, em sua forma mais simples, pode ser resumida como: “*você conhecerá uma palavra pela companhia que ela mantém*” (Firth, 1957). Embora essa hipótese seja a base de todas as representações modernas, suas primeiras implementações eram baseadas em contagem.

1.3.1.1 Saco de palavras

A representação em saco de palavras ou *bag-of-words* (BoW) é a forma mais simples de representar um documento de texto. Nela, um texto é dividido em palavras e cada palavra é associada a sua ocorrência ou ao valor de sua frequência nesse texto.

Considere o exemplo a seguir³.

¹Sistema de classificação mantido pela OMS para codificação de doenças e problemas de saúde. Disponível em: <http://www2.datasus.gov.br/cid10/V2008/cid10.htm>

²Vocabulário controlado desenvolvido pela National Library of Medicine dos EUA para indexação de artigos biomédicos. Disponível em: <https://www.nlm.nih.gov/mesh/meshhome.html>

³Frase extraída do livro Memórias Póstumas de Brás Cubas e livremente adaptada para o português moderno, disponível gratuitamente no Projeto Gutenberg (<https://www.gutenberg.org/>).



Exemplo 1.1:

talvez espante ao leitor a franqueza com que lhe exponho e realço a minha mediocridade

Em um modelo de saco de palavras, o texto é primeiramente dividido em palavras unitárias, separadas por um espaço em branco (processo conhecido como tokenização explicado no Capítulo [Sequência de caracteres e palavras](#)), resultando no mapeamento a seguir onde cada palavra é associada a sua frequência de ocorrência.

```
{ "a": 2, "ao": 1, "com": 1, "e": 1, "espante": 1, "exponho": 1,
  "franqueza": 1, "leitor": 1, "lhe": 1, "mediocridade": 1, "minha": 1,
  "que": 1, "realço": 1, "talvez": 1 }
```

Essa representação ignora completamente a gramática, a sintaxe e a ordem das palavras, tratando o documento apenas como um multiconjunto (um “saco”) de seus termos.

Formalizando, dado um vocabulário V (o conjunto de todas as palavras únicas do *corpus*), cada documento d é representado por um vetor v de tamanho $|V|$. Cada posição i do vetor v corresponde a um termo $t_i \in V$, e o valor v_i é a frequência (contagem) do termo t_i no documento d .

Essa representação tem a vantagem de ser simples e de apresentar fácil interpretabilidade (é fácil ver quais palavras contribuíram para a classificação) e eficácia surpreendente na classificação de tópicos (onde a presença de palavras-chave é um forte indicador). Por outro lado, os vetores resultantes podem ter elevada dimensionalidade (dezenas de milhares de posições), ser esparsos (a maioria das posições é zero), apresentar perda total da ordem (por exemplo, “o cachorro mordeu o homem” *vs* “o homem mordeu o cachorro”) e ser incapazes de capturar sinonímia (os vetores para “carro” e “automóvel” são ortogonais, ou seja, totalmente independentes).

1.3.1.2 Term Frequency-Inverse Document Frequency (TF-IDF)

Uma melhoria crucial em relação à contagem bruta é o TF-IDF, que combina *Term Frequency* – *TF* (frequência de termo) e *Inverse Document Frequency* – *IDF* (frequência inversa de documentos). Essa junção visa dar mais peso a palavras frequentes em um documento específico, mas raras em todos os demais documentos. Palavras muito comuns em toda a coleção (como *stopwords* “de”, “que”, “o”) recebem um peso baixo, pois não têm poder discriminativo.

O peso de um termo t no documento d é dado por:

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D), \quad (1.1)$$

onde $tf(t, d)$ é a frequência do termo em d , computada por:

$$tf(t, d) = \frac{freq_{t,d}}{\sum_{t' \in d} freq_{t',d}}, \quad (1.2)$$

e $idf(t, D)$ é a frequência inversa do documento no corpus D :

$$idf(t, D) = \log \frac{N}{n_t} \quad (1.3)$$



Na Equação 1.2, $freq_{t,d}$ é a frequência de um termo t no documento d . O denominador $\sum_{t' \in d} freq_{t',d}$ representa a soma das frequências de todos os termos t' presentes em d . O objetivo desta soma é normalizar a frequência de um termo em relação ao tamanho do documento. Um documento longo deve apresentar termos com frequência elevada, enquanto um documento pequeno apresenta termos com frequência baixa. Para evitar distorções, a quantidade de termos de um documento é considerada.

Após o cálculo da frequência de um termo, é necessário considerar o quão raro ou frequente ele é em um conjunto de textos D (Equação 1.3). Para seu cálculo, utilizam-se a quantidade de documentos no corpus (N) e o número de documentos em D em que t ocorre (n_t).

O resultado dessa representação ainda é um vetor esparsos e de alta dimensionalidade, mas os valores são ponderados pela sua relevância informacional, tornando-o uma *baseline* muito mais robusta para classificadores tradicionais.

Para um arcabouço teórico mais profundo, indicamos ao leitor o conteúdo sobre vetores esparsos (Seção **Vetores esparsos**). A representação TF-IDF também pode ser utilizada no contexto da Recuperação de Informação (RI). O Capítulo 2 fornece mais detalhes sobre o TF-IDF neste contexto.

1.3.2 Representações distribuídas (densas)

A grande revolução na representação de texto veio com a aplicação de redes neurais para aprender representações, em vez de apenas contar as palavras. Este é o foco da Seção **Vetores densos estáticos**.

As representações vetoriais são formalmente chamadas de Modelos Semânticos Distribucionais (MSD). *Word embeddings* (vetores de palavras) gerados por modelos como Word2Vec (Mikolov et al., 2013a) e GloVe (Pennington et al., 2014) aprendem vetores densos (normalmente de 50 a 300 dimensões) para cada palavra, em que cada dimensão idealmente representa uma característica semântica da palavra. A principal vantagem dessa representação em relação àquela apresentada anteriormente é que palavras com significados semelhantes (como “carro” e “automóvel”) terão vetores próximos no espaço vetorial.

Contudo, a tarefa de categorização opera sobre documentos, não sobre palavras. Dessa forma, para obter um vetor que represente um documento, as abordagens mais comuns são:

1. **Concatenação:** o vetor do documento é obtido a partir da concatenação dos vetores das palavras;
2. **Média de vetores:** o vetor do documento é obtido calculando-se a média dos vetores de todas as palavras no documento, elemento a elemento (Jurafsky; Martin, 2026);
3. **Média ponderada:** o vetor do documento é obtido como resultado da média ponderada na qual as palavras de maior peso (por exemplo, o TF-IDF) contribuem mais para o vetor final.

As representações distribuídas resolvem o problema da sinonímia, mas ainda sofrem com a perda de ordem e, crucialmente, com a polissemia: a palavra “banco” terá o mesmo vetor em “sentei no banco da praça” e em “fui ao banco sacar dinheiro”. Existem, entretanto, representações distribuídas densas cujo objetivo é representar por completo um



documento textual ou frases (Le; Mikolov, 2014). Este tipo de representação, assim como métodos específicos, como Word2Vec, GloVe e FastText, é detalhado na Seção **Vetores densos estáticos**.

1.3.3 Representações contextuais (modelos de linguagem profundos)

A fronteira atual, que resolve o problema da polissemia, é o uso de modelos de linguagem profundos baseados na arquitetura Transformer, como o BERT (Devlin et al., 2019) e suas variantes (Liu et al., 2019b) (RoBERTa, ALBERT e outras).

Esses modelos são pré-treinados em volumes massivos de texto (como discutido no Capítulo **Modelos de linguagem**) e não geram um vetor fixo por palavra. Em vez disso, eles geram um vetor para cada *token* (palavra ou subpalavra) que é sensível ao contexto no qual esse *token* aparece. Neste caso, o vetor de representação da palavra “banco” será diferente nas duas sentenças de exemplo da seção anterior.

Para a tarefa de categorização de textos, temos duas estratégias principais:

1. **Extração de características (*feature extraction*):** o documento é passado pelo modelo pré-treinado (sem modificá-lo), e os vetores de saída são usados como representação. Frequentemente, utiliza-se a representação do *token* especial [CLS] (no caso do BERT), projetado para agregar o significado de toda a sequência. Alternativamente, pode-se aplicar uma operação de *pooling* (por exemplo, a média) aos vetores de todos os *tokens* da sequência. Esse vetor denso e contextual (por exemplo, de 768 dimensões) é então usado como entrada de um classificador clássico⁴.
2. **Ajuste fino (*fine-tuning*):** esta é a abordagem dominante. O modelo de linguagem pré-treinado é “descongelado” e treinado (com uma taxa de aprendizado baixa) juntamente com uma camada de classificação simples (por exemplo, softmax) adicionada ao topo. O modelo inteiro é ajustado para a tarefa específica de categorização. Isso permite que as representações, que eram de propósito geral, se especializem nos padrões discriminativos da base de dados de aplicação.

A ascensão dos modelos de linguagem de larga escala (do inglês, *Large Language Models* – LLMs), como o GPT e modelos de código aberto (discutidos no Capítulo **ChatGPT, MariTalk e outros agentes de conversação**), também permite abordagens de *zero-shot* ou *few-shot learning*, nas quais o modelo pode categorizar textos com base apenas na descrição das categorias (ou seja, um *prompt*), sem necessitar de um grande conjunto de treinamento rotulado.

A escolha da representação (Quadro 1.1) depende do problema (Seção 1.2) e dos recursos disponíveis: o TF-IDF ainda é uma *baseline* rápida e robusta para tópicos claros; *fine-tuning* de modelos de linguagem é o estado da arte para a maioria das tarefas que exigem compreensão semântica profunda.

⁴Em modelos mais recentes (como o RoBERTa), o uso da média dos vetores (conhecida como *mean pooling*) tem se mostrado mais estável do que o [CLS] isolado para certas tarefas de categorização

Quadro 1.1: Comparativo das abordagens de representação para classificação de textos

Abordagem	Unidade Básica	Dimension.	Captura Semântica	Sensível ao Contexto
TF-IDF (Distributiva)	Palavra (Termo)	Alta e Esparsa (e.g., 50.000)	Nenhuma (Contagem)	Não
Word Embeddings (Distribuída)	Palavra	Baixa e Densa (e.g., 300)	Alta (Estática)	Não
LLMs (Contextual)	Subpalavra (Token)	Média e Densa (e.g., 768+)	Altíssima (Dinâmica)	Sim

A identificação e a análise dos documentos que se deseja classificar são importantes para determinar o tipo de classificação. Após isso, é necessária a representação destes documentos para a aprendizagem de um modelo.

1.4 Métodos de classificação

Uma vez que o texto foi transformado em uma representação numérica (Seção 1.3), entramos na fase de tomada de decisão. O objetivo desta seção é apresentar os métodos que atuam como o “cérebro” do sistema de categorização, responsáveis por aprender padrões nesses números e atribuir a classe correta a um documento inédito.

Formalmente, o problema de classificação pode ser visualizado como a busca por uma função de mapeamento $f : D \rightarrow C'$ que associa cada documento $d \in D$ a $C' \subseteq C$. Aqui, d é um documento (um ponto ou um vetor) e C' é o subconjunto de categorias discretas. Classificar, portanto, nada mais é do que encontrar uma regra — matemática ou lógica — que divida esse espaço de forma que documentos de categorias diferentes fiquem em regiões distintas (Faceli et al., 2025).

Para o leitor que está iniciando na área, é útil pensar em um classificador através de duas fases fundamentais:

1. **Fase de treinamento (aprendizado):** o método analisa um conjunto de dados previamente rotulado (o *corpus* de treinamento, conforme discutido no Capítulo [Conjunto de dados, dataset e corpus](#)). Durante esse processo, o modelo tenta ajustar seus parâmetros internos para minimizar os erros de classificação, identificando quais características (palavras, contextos ou padrões) são os melhores indicadores de cada categoria.
2. **Fase de inferência (predição):** o modelo, agora “treinado”, recebe um novo documento que ele nunca observou. Ele aplica o que aprendeu para projetar esse documento no espaço de categorias e decide qual rótulo é o mais adequado.

Ao longo das últimas décadas, a área de PLN testemunhou uma evolução fascinante nos métodos de classificação. Partimos de modelos baseados em intuições probabilísticas simples (Gonçalves; Quaresma, 2003; Pinto; Melgar, 2016), passamos por algoritmos que buscam fronteiras geométricas ideais (Gonçalves; Quaresma, 2003; Teixeira et al., 2018) e estruturas de árvores lógicas (Fonseca et al., 2014), até chegarmos às redes neurais profundas (Cortes et al., 2020; Matos et al., 2024) e aos modelos de linguagem de larga

escala (LLMs), que operam quase intuitivamente por meio de instruções (Gillin, 2024; Oliveira et al., 2024a).

Nesta seção, exploraremos esses métodos, dividindo-os em seus paradigmas fundamentais. Veremos que a escolha do método não depende apenas da precisão desejada, mas também da quantidade de dados disponíveis, da capacidade computacional e do nível de explicabilidade exigido pela aplicação — afinal, em domínios sensíveis como a Saúde (Capítulo 9) ou o Direito (Capítulo 12), entender “*por que*” um modelo tomou uma decisão pode ser tão importante quanto a decisão em si.

1.4.1 Métodos probabilísticos: Naïve Bayes

O método Naïve Bayes (NB) pertence à classe de modelos generativos e baseia-se na aplicação direta do Teorema de Bayes. Apesar de sua simplicidade matemática, ele é eficiente e serve como uma *baseline* robusta para tarefas como a filtragem de *spam* (Alberto et al., 2015; Almeida et al., 2011; Almeida et al., 2016), a identificação de *fake news* (Silva et al., 2020a) e *fake reviews* (Cardoso et al., 2018) e a classificação de tópicos (Gonçalves; Quaresma, 2003).

Para entender o NB, imagine que você está lendo um e-mail e encontra as palavras “promoção”, “desconto” e “clique”. Instantaneamente, sua mente aumenta a probabilidade de que este e-mail seja *spam*. Se, por outro lado, você encontrar palavras como “aula”, “professor” e “prova”, a probabilidade de que seja um e-mail legítimo cresce. O NB faz exatamente isso: ele conta a frequência com que cada palavra aparece em cada categoria durante o treino e utiliza essas contagens como “evidências” para decidir a classe de um novo texto.

Formalizando, dado um documento d (que visualizamos como um conjunto de palavras $w_1, w_2, \dots, w_n$), queremos encontrar a classe c que maximiza a probabilidade $P(c|d)$. Pelo Teorema de Bayes, temos:

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)} \quad (1.4)$$

Como o denominador $P(d)$ (a probabilidade do documento ocorrer) é constante para todas as classes, o classificador foca apenas no numerador:

1. $P(c)$ (**probabilidade a priori**): é a probabilidade de uma classe ocorrer antes mesmo de lermos o texto. Se 80% dos e-mails no seu treinamento são *spam*, $P(\text{spam}) = 0,8$.
2. $P(d|c)$ (**verossimilhança ou likelihood**): é a probabilidade de observarmos aquele conjunto de palavras dentro de uma classe específica.

A dificuldade computacional reside em calcular $P(d|c)$, ou seja, a probabilidade de uma sequência exata de palavras ocorrer. Para viabilizar o cálculo, o modelo assume a *premissa de independência*: a presença de uma palavra em um documento é totalmente independente da presença de outras palavras, dada a classe.

Sob essa suposição, a probabilidade conjunta torna-se o produto das probabilidades individuais:

$$P(d|c) \approx P(w_1|c) \times P(w_2|c) \times \dots \times P(w_n|c) \quad (1.5)$$

Sabemos que esta premissa é falsa na linguagem natural — a palavra “Inteligência” aumenta muito a chance da palavra seguinte ser “Artificial”. No entanto, mesmo sendo



“ingênuo”, o modelo preserva a força relativa das evidências, o que o torna preciso na prática.

Ao implementar o NB, há dois desafios clássicos que precisam de atenção (Manning et al., 2008a):

- **O problema da divisão por zero (suavização de Laplace):** se uma palavra w nunca apareceu na classe “Esportes” durante o treino, sua probabilidade $P(w|Esportes)$ será 0. Como estamos multiplicando probabilidades, um único zero anularia todo o cálculo. Para resolver isso, adicionamos 1 ao contador de todas as palavras (técnica de suavização de Laplace), garantindo que nenhuma probabilidade seja nula.
- **Espaço de logaritmos:** multiplicar muitas probabilidades pequenas (ex.: $0,0001 \times 0,0002$) pode levar a erros de precisão numérica. Por isso, na prática, os logaritmos das probabilidades são somados: $\log(P(c|d)) \propto \log(P(c)) + \sum \log(P(w_i|c))$.

O NB é, portanto, um exemplo de como uma modelagem estatística simplificada pode oferecer excelente desempenho a baixo custo computacional, capaz de processar milhões de documentos em segundos (Freitas et al., 2019).

1.4.2 Métodos baseados em distância: K-vizinhos mais próximos

O método k -vizinhos mais próximos (do inglês, *K-Nearest Neighbors* – KNN) adota uma abordagem intuitiva e geométrica para a classificação (Cover; Hart, 1967). Diferente de outros métodos que tentam criar um modelo matemático abstrato sobre os dados, o KNN baseia-se na premissa de que “documentos semelhantes tendem a pertencer à mesma categoria”.

Intuitivamente, imagine que você está em uma biblioteca imensa e encontra um livro sem capa nem título. Para descobrir o assunto dele, você olha para os livros guardados em posições imediatamente ao redor. Se a maioria dos vizinhos próximos trata de “Cálculo”, é muito provável que o livro misterioso também pertença a essa categoria.

O KNN é classificado como um algoritmo de *aprendizado preguiçoso (lazy)*, pois não possui uma fase de treinamento explícita. Ele simplesmente armazena todos os exemplos de treinamento na memória. O verdadeiro trabalho ocorre apenas no momento da predição, quando o algoritmo compara o novo documento com todos os documentos armazenados.

O processo de classificação de um novo documento d segue três passos:

1. **Cálculo de distância:** o método calcula a distância (ou similaridade) entre o vetor de d e todos os vetores dos documentos do conjunto de treinamento.
2. **Seleção dos vizinhos:** o método identifica os k documentos mais próximos (vizinhos).
3. **Votação majoritária:** o método atribui, ao documento d , a classe que aparece com maior frequência entre esses k vizinhos.

No primeiro passo, em problemas de computação tradicionais (que envolvem dados tabulares), a distância euclidiana⁵ é a mais utilizada. No entanto, em textos, ela apresenta uma

⁵A distância euclidiana entre dois vetores é computada como $d(u, v) = \sqrt{\sum_{i=1}^n (u_i - v_i)^2}$. Note que o comprimento do vetor não é levado em consideração nesta equação, assim, textos muito grandes, mas parecidos podem produzir similaridade baixa.



falha crítica: a sensibilidade ao tamanho do documento. Dois textos sobre o mesmo tema, mas com tamanhos muito diferentes, teriam uma distância euclidiana elevada, mesmo sendo semanticamente idênticos. Por isso, na categorização de texto, preferimos a *similaridade de cosseno*, discutida no Capítulo [Representação vetorial e semântica distribucional](#). Ela mede o ângulo entre dois vetores, considerando a orientação e não a magnitude. Isso garante que a comparação seja feita com base na proporção de palavras, tornando o modelo invariante ao tamanho do texto.

A escolha do valor do hiperparâmetro k (para o segundo passo) é o principal desafio deste método. Geralmente, utiliza-se um valor de k ímpar para evitar empates em classificações binárias:

- Se $k = 1$, o modelo fica muito sensível a ruídos ou erros de rotulação no treino (um vizinho “errado” define a classe).
- Se k for muito grande, o modelo pode perder a capacidade de capturar nuances locais e pode ser dominado pelas classes majoritárias (aquelas que possuem mais documentos no total).

O KNN é valorizado por ser não-paramétrico, o que significa que ele não assume nenhuma distribuição específica para os dados — ele se adapta perfeitamente a fronteiras de decisão complexas e irregulares. Contudo, sua principal limitação no contexto de PLN é a eficiência. Como ele precisa comparar cada novo documento a toda a base de dados, o tempo de resposta cresce linearmente com o tamanho do *corpus*. Para aplicações que exigem baixa latência ou lidam com milhões de documentos, o custo computacional e de memória pode ser proibitivo, o que exige o uso de estruturas de dados avançadas para busca aproximada de vizinhos (como *Ball Trees* (Omohundro, 1989) ou *Locality Sensitive Hashing* (Indyk; Motwani, 1998)).

1.4.3 Métodos de otimização: Regressão Logística e SVM

Enquanto o Naïve Bayes computa probabilidades e o KNN calcula vizinhanças, os métodos de otimização tentam encontrar uma fronteira de decisão — uma linha, plano ou hiperplano — que separe as categorias de forma ótima no espaço de características. Nesta seção, exploraremos dois dos métodos desse paradigma mais utilizados em PLN: a Regressão Logística e as Máquinas de Vetores de Suporte (SVM) (Faceli et al., 2025).

Apesar do nome sugerir uma tarefa de regressão, a **Regressão Logística** é um dos métodos de classificação mais robustos e interpretáveis da literatura. No contexto de PLN, ele é frequentemente utilizado para entender quais palavras ou características têm maior peso na decisão de uma categoria. Diferente da Regressão Linear, que pode retornar qualquer valor, a Regressão Logística utiliza a função logística (ou sigmoide) para ajustar o resultado da soma ponderada ao intervalo entre 0 e 1, representando a probabilidade de pertencer à classe (Cox, 1958).

Como destacado por Faceli et al. (2025), a vantagem deste método é a sua interpretabilidade: ao analisarmos os coeficientes (pesos) aprendidos, conseguimos explicar exatamente por que um documento foi classificado em determinada categoria, o que é essencial em domínios como o Direito (Capítulo [12](#)).

As **Máquinas de Vetores de Suporte** (do inglês, *Support Vector Machines* – SVM) foram, durante anos, o estado da arte na categorização de texto antes da ascensão das

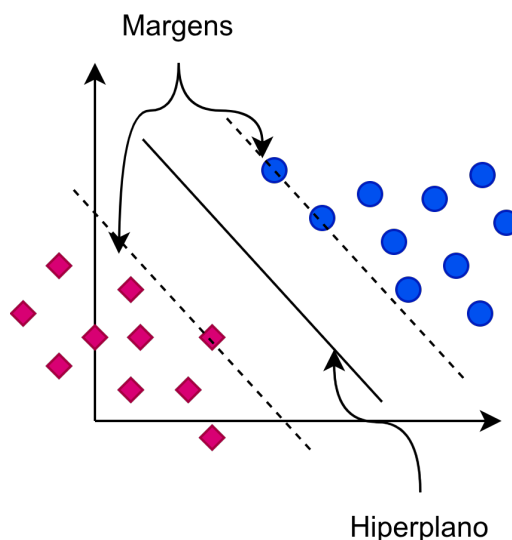


redes neurais profundas. O SVM é particularmente eficaz quando lidamos com espaços de alta dimensionalidade, como no caso das representações TF-IDF discutidas na Seção 1.3).

Intuitivamente, imagine que as categorias sejam duas cidades separadas por uma floresta. O SVM não quer apenas desenhar uma linha divisória entre elas; quer construir a estrada mais larga possível que as separe. A linha central dessa estrada é chamada *hiperplano de decisão*, e as bordas da estrada são chamadas de *margens* (Figura 1.3).

O objetivo do SVM é maximizar essa margem. Quanto maior o espaço entre a fronteira e os documentos mais próximos de cada classe, maior será a capacidade do modelo de classificar novos documentos corretamente (generalização) (Cortes; Vapnik, 1995).

Figura 1.3: Máquina de vetores suporte com o hiperplano de separação entre os dados



Um detalhe didático do SVM é que a decisão final não depende de todos os documentos de treino. Ela depende apenas dos documentos que estão “na beira da estrada”, ou seja, nos limites das margens. Esses documentos são os *vetores de suporte*. Se removermos qualquer outro documento do treino, a fronteira permanece igual; se movermos um vetor de suporte, a fronteira muda.

Muitas vezes, os textos não são linearmente separáveis (não é possível traçar uma linha reta que os divida perfeitamente). O SVM resolve isso ao projetar os dados em um espaço de maior dimensão. Essa transformação matemática é realizada de forma eficiente por meio de funções de *kernel*, permitindo que o modelo capture relações não lineares complexas entre as palavras.

No uso prático em PLN:

- A Regressão Logística é preferida quando a velocidade de treinamento e a interpretabilidade dos pesos das palavras são cruciais.
- O SVM é a escolha ideal quando o volume de dados é menor, mas a dimensionalidade é muito alta, oferecendo uma robustez matemática superior contra o sobreajuste (*overfitting*).

1.4.4 Métodos baseados em árvores e *ensembles*

Enquanto os modelos anteriores buscam equações matemáticas ou vizinhanças espaciais, os métodos baseados em árvores adotam uma abordagem de divisão e conquista. Eles particionam o espaço de características por meio de uma sequência de perguntas lógicas, criando uma estrutura que se assemelha a um fluxograma de decisão (Faceli et al., 2025).

Uma **Árvore de Decisão** (Quinlan, 1993) é uma estrutura onde cada nó interno representa um teste sobre uma característica do texto (ex.: “A palavra lucro está presente?”). Cada ramo da árvore representa o resultado do teste (Sim ou Não), e cada nó-folha representa o rótulo da categoria final.

Pense na classificação como um jogo de adivinhação. O modelo tenta fazer a pergunta mais informativa primeiro para eliminar o maior número possível de incertezas. Em um *corpus* de notícias, a pergunta “O texto contém o termo campeonato?” pode separar quase instantaneamente a categoria “Esportes” das demais.

Para decidir qual palavra usar em cada nó, a árvore utiliza medidas estatísticas de desordem⁶. A mais comum é a *entropia*. Se um conjunto de documentos contém textos de várias categorias misturadas, a entropia é alta. O objetivo do algoritmo é escolher uma divisão que resulte em subconjuntos mais “puros” (de menor entropia), aumentando o *Ganho de Informação*.

A maior vantagem desse método é a explicabilidade: qualquer pessoa consegue seguir o caminho do topo da árvore até a folha para entender a decisão. O risco, porém, é o sobreajustamento: a árvore pode se tornar tão profunda e específica que acaba “decorando” o ruído do treino em vez de aprender padrões gerais. Para mitigar as falhas de uma única árvore, normalmente utilizam-se métodos de *ensemble* (conjuntos), que combinam as previsões de várias árvores para obter um resultado mais robusto e estável. Duas estratégias neste sentido são a Floresta Aleatória e os métodos de *boosting*.

Uma **Floresta Aleatória** cria centenas de árvores de decisão distintas. O segredo de sua eficácia é a diversidade:

- Cada árvore é treinada em um subconjunto aleatório dos dados (*bagging*);
- Em cada nó, apenas um subconjunto aleatório de palavras é considerado para a divisão.

A decisão final é tomada por votação majoritária. Como destacado por Breiman (2001), essa colaboração entre árvores independentes reduz drasticamente o erro e torna o modelo um dos mais confiáveis para características extraídas manualmente, como as medidas de complexidade textual discutidas no Capítulo 6.

Diferentemente das florestas aleatórias, nas quais as árvores crescem em paralelo, no *boosting* as árvores são construídas sequencialmente. A lógica é: a primeira árvore faz uma previsão inicial (que pode ser fraca); a segunda é treinada especificamente para corrigir os erros da primeira, e assim por diante. O **XGBoost** (Chen; Guestrin, 2016) é uma implementação altamente otimizada desse processo. Ele lida nativamente com dados esparsos (característica comum em BoW/TF-IDF), o que justifica sua alta performance no

⁶Aqui desordem refere-se a um conceito estabelecido na Teoria da Informação chamado de Entropia de Shannon, que visa quantificar a “impureza” ou a incerteza dos dados (Faceli et al., 2025). Em uma tarefa de classificação de texto, um conjunto de dados com alta “impureza” significa que os elementos de diversas categorias estão misturados e um modelo busca organizar esta “impureza”, tornando os conjuntos mais homogêneos.

PLN tradicional, frequentemente vencendo competições de ciência de dados. Em PLN, ele se destaca quando há vetores de características bem definidos ou metadados estruturados.

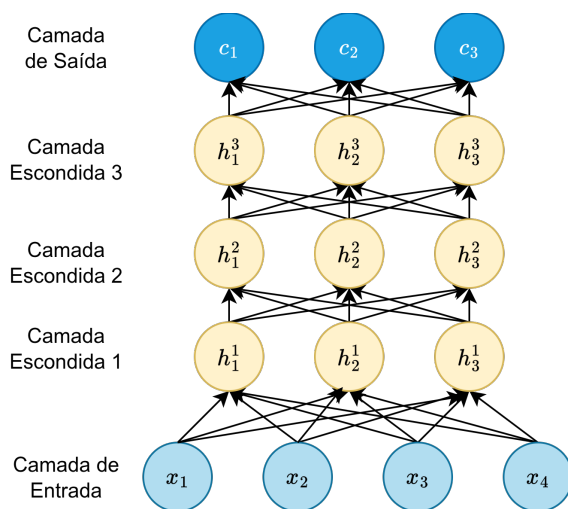
Embora as árvores tenham dificuldade em lidar diretamente com vetores TF-IDF de altíssima dimensão (em que a maioria dos valores é zero), elas são altamente eficientes quando o problema envolve a combinação de texto com dados estruturados (e.g. classificar um anúncio com base no texto da descrição e no preço do produto) ou quando a interpretabilidade das regras de negócio é um requisito legal.

1.4.5 Métodos connexionistas: Redes neurais e *Deep learning*

Os métodos connexionistas, inspirados na estrutura biológica do cérebro, representam uma mudança de paradigma na categorização de texto. Enquanto os métodos anteriores dependem de uma representação de entrada fixa, como o TF-IDF, as redes neurais são capazes de aprender características (*feature learning*) e extrair padrões complexos e semânticos diretamente dos dados brutos. A unidade básica desta abordagem é o neurônio artificial, que realiza uma soma ponderada de suas entradas e aplica uma função de ativação não linear (Faceli et al., 2025). Para a categorização de texto, o modelo mais simples é o *Multilayer Perceptron* (MLP) (Figura 1.4).

Intuitivamente, imagine que a primeira camada da rede observa palavras individuais, a segunda camada combina essas palavras em sentenças curtas e a terceira camada identifica conceitos abstratos (como “ironia” ou “urgência”). Ao final, a rede decide a categoria com base nessas abstrações aprendidas durante o treinamento.

Figura 1.4: Exemplo de uma rede neural simples



Em vez de contagens esparsas, as redes neurais frequentemente utilizam camadas de *embeddings* (como detalhado no Capítulo [Representação vetorial e semântica distribucional](#)). Essas camadas transformam palavras em vetores densos que capturam afinidade semântica. A rede ajusta seus pesos por meio do algoritmo de retropropagação (*back-propagation*), minimizando o erro entre a predição e o rótulo real. A seguir uma breve explicação das principais arquiteturas existentes.

Redes neurais convolucionais (CNN). Originalmente propostas para o processamento de imagens, as CNNs foram adaptadas com sucesso para o processamento de texto (Kim, 2014). Elas funcionam como “detectores de n -gramas” automáticos.

Uma CNN desliza janelas (filtros) sobre o texto. Um filtro pode se especializar em detectar a expressão “muito caro”, enquanto outro pode focar em “excelente atendimento”. A rede não se importa de onde essas expressões aparecem no documento; ela apenas registra que ocorreram e as utiliza para classificar o sentimento ou o tópico.

Redes recorrentes (RNN) e LSTMs. A linguagem humana é inerentemente sequencial e a ordem das palavras altera o sentido (e.g. “O gato caçou o rato” *vs* “O rato caçou o gato”). As Redes Neurais Recorrentes (RNNs) foram projetadas para lidar com essa característica, mantendo um “estado oculto” que serve como uma memória do que foi lido anteriormente.

Como as RNNs tradicionais sofrem com o esquecimento de informações distantes, a arquitetura *Long Short-Term Memory* (LSTM) (Hochreiter; Schmidhuber, 1997) foi proposta para resolver esse problema introduzindo “portões” (*gates*) que controlam o que deve ser lembrado e o que deve ser descartado. A principal vantagem é que elas conseguem entender que uma negação no início de um parágrafo longo afeta o sentido de uma palavra no final. Devido ao alto desempenho em diversas tarefas de PLN, elas foram o padrão-ouro para categorização de documentos longos até a consolidação dos *Transformers*.

1.4.6 Modelos de linguagem

A grande vantagem desses métodos é a flexibilidade. Podemos combinar CNNs e LSTMs para criar modelos híbridos que capturam tanto padrões locais quanto dependências globais. No entanto, o treinamento dessas redes do zero exige grandes volumes de dados rotulados. Essa limitação preparou o terreno para o uso de modelos pré-treinados, onde redes neurais profundas (os *Transformers*) aprendem a língua em *corpora* massivos antes de serem ajustadas para a tarefa de categorização, como veremos na próxima seção e no Capítulo [Modelos de linguagem](#).

1.4.7 Redes com atenção: Transformers

Se as CNNs focam em padrões locais e as RNNs em sequências, a arquitetura *Transformers* (ou rede de transformadores) (Vaswani et al., 2017a) introduziu uma abordagem que mudou permanentemente o PLN: o mecanismo de auto-atenção (do inglês, *self-attention*). Esta arquitetura abandonou a recorrência e a convolução em favor de uma estrutura que processa todas as palavras de um documento simultaneamente.

Imagine que, ao ler uma frase, você pudesse projetar um “holofote” sobre as palavras mais importantes para entender o sentido de uma palavra específica. Na frase “O banco da praça estava quebrado”, ao ler a palavra “banco”, seu cérebro se concentra em “praça” para entender que se trata de um assento, e não de uma instituição financeira. Uma rede de transformadores faz exatamente isso através da autoatenção: ela calcula o peso (a importância) de cada palavra do texto em relação a todas as outras, criando uma representação contextualizada.

Para a tarefa de categorização, o modelo mais emblemático é o BERT (*Bidirectional Encoder Representations from Transformers*) (Devlin et al., 2019). Diferentemente de modelos que leem o texto da esquerda para a direita, o BERT lê o texto em ambas as direções simultaneamente, o que lhe permite capturar o contexto completo de cada termo.

Para classificar um texto com o BERT, utilizamos um procedimento específico:

1. **O *token* [CLS]:** adicionamos um *token* especial chamado [CLS] (de *classification*) no início de cada documento;



2. **Agregação de sentido:** após passar por todas as camadas de atenção da rede, o vetor resultante deste *token* [CLS] funciona como um “resumo” matemático de todo o documento;
3. **Camada de saída:** adicionamos uma camada densa simples (como uma Regressão Logística) que processa este vetor para determinar a categoria final.

Diferente das LSTMs, que processam o texto palavra por palavra e podem “esquecer” o início de documentos longos, as redes de transformadores mantêm o acesso direto a todas as palavras ao mesmo tempo. Além disso, por não ser uma arquitetura sequencial, no treinamento é possível realizar o processamento paralelo em massa, o que viabiliza o treinamento em *corpora* de escala global (como discutido no Capítulo [Modelos de linguagem](#)).

Este método é hoje o ponto de partida para qualquer sistema de categorização de alto desempenho, servindo de base para os modelos que exploraremos na próxima seção.

1.4.8 Classificação via *prompt* e modelos de linguagem de larga escala (LLMs)

A chegada dos modelos de linguagem de larga escala (LLMs), como as famílias GPT, Llama e a brasileira MariTalk (detalhadas no Capítulo [ChatGPT, MariTalk e outros agentes de conversação](#)), introduziu uma nova forma de classificar textos. Nos métodos anteriores, o foco era treinar um modelo específico para uma tarefa; aqui, o foco passa a ser a instrução de um modelo de propósito geral.

Imagine que, em vez de construir uma máquina complexa para separar e-mails, você pudesse simplesmente entregar um documento a um assistente com amplo conhecimento e solicitar:

Quadro 1.2: Exemplo de instrução com *zero-shot learning* para a categorização de um texto

“Você é um jornalista encarregado de classificar uma dada notícia em uma das seguintes categorias: política, economia, esporte e policial. Agora, considere a notícia a seguir e responda apenas em qual categoria ela melhor se encaixa.”
{texto da notícia}

Neste caso, o assistente usa todo o seu conhecimento prévio sobre o mundo e a língua para tomar a decisão, sem precisar que você mostre exemplos antes.

No paradigma de *prompting*, não alteramos os pesos do modelo (como fazemos no ajuste fino do BERT). Em vez disso, fornecemos um texto de entrada (o *prompt*) que contém a descrição da tarefa e o documento a ser classificado.

O principal mecanismo por trás desta abordagem é a capacidade do modelo de realizar o que chamamos de aprendizado em contexto. Existem duas formas principais de aplicar isso à categorização:

- ***Zero-shot learning* (Aprendizado sem exemplos):** o modelo recebe apenas a instrução e o texto. Exemplo: “Classifique o sentimento deste tweet como Positivo ou Negativo: ‘O suporte técnico resolveu meu problema em minutos!’”.
- ***Few-shot learning* (Aprendizado com poucos exemplos):** o modelo recebe dois ou três exemplos de textos já classificados antes de solicitar a classificação do



novo documento. Isso ajuda o modelo a entender o formato esperado e as sutilezas das categorias (Brown et al., 2020b).

Quando possível, a inclusão de bons exemplos (*few-shot learning*) fornece mais contexto para que o modelo realize a inferência e, portanto, costuma oferecer resultados superiores aos da abordagem *zero-shot learning*. Veja o exemplo a seguir:

Quadro 1.3: Exemplo de instrução com *few-shot learning* para a categorização de um texto

“Você está analisando textos curtos de redes sociais para determinar se expressam sentimento positivo, negativo ou neutro. Os textos são relacionados ao esporte e podem conter linguagem informal e ironia. Veja, a seguir, alguns casos e a classificação correta de cada um.

Caso 1) Ninguém segura o Mengão \o/
Resposta 1) Positivo.

Caso 2) Mais uma vez o VAR!!! Ninguém aguenta mais juiz!!!!
Resposta 2) Negativo.

Caso 3) Mais uma partida e estamos aqui aguardando...
Resposta 3) Neutro.”

{texto a ser classificado}

Veja que, neste exemplo, foram fornecidos três casos diferentes, um para cada classe. Dessa forma, o modelo pode compreender melhor a tarefa e oferecer um poder preditivo maior. Contudo, é importante destacar que a ordem dos exemplos no *prompt* pode influenciar a resposta do modelo (problema conhecido como viés de recência), o que constitui um desafio prático real da engenharia de *prompt*.

Para que os modelos se tornassem bons classificadores por meio de *prompt*, passaram por um processo chamado *instruction tuning* (Ouyang et al., 2022). Como discutido no Capítulo Modelos de linguagem, os modelos de linguagem básicos (da primeira geração) são treinados apenas para prever a próxima palavra. Modelos modernos (de larga escala) foram ajustados para seguir ordens humanas, o que os torna capazes de entender restrições complexas na categorização, como: “Classifique este contrato jurídico, mas se houver ambiguidade, atribua o rótulo ‘Revisão Necessária’”.

Esta abordagem resolve um dos maiores gargalos do PLN: a *necessidade de dados rotulados por humanos*. É possível criar um classificador funcional em minutos. No entanto, é preciso ficar atento a dois pontos críticos:

1. **Custo e latência:** rodar um LLM para classificar milhões de documentos é muito mais caro e lento do que usar um Naïve Bayes ou um SVM, por exemplo.
2. **Instabilidade do *prompt*:** pequenas mudanças na forma como escrevemos a instrução podem gerar resultados distintos, um fenômeno que exige técnicas de engenharia de *prompt* (Kojima et al., 2022).

A categorização via LLMs é ideal para prototipagem rápida, tarefas com categorias altamente subjetivas ou quando não temos dados de treino suficientes para os métodos conexionistas clássicos (Brown et al., 2020b; Wei et al., 2022).



1.4.9 Síntese e Recomendação

Para facilitar a escolha do método mais adequado, de acordo com a natureza do problema e os recursos disponíveis, o Quadro 1.4 apresenta uma síntese comparativa dos paradigmas discutidos nesta seção.

Quadro 1.4: Quadro comparativo dos métodos de classificação de texto

Método	Características principais	Aplicações recomendadas
Naïve Bayes	Baixo custo computacional; assume independência entre os termos; robusto com poucos dados.	Filtragem de <i>spam</i> ; classificação de tópicos em tempo real; <i>baselines</i> rápidas.
KNN	Não-paramétrico; sem fase de treino (<i>lazy</i>); sensível à métrica de distância e ao ruído.	Busca por documentos similares; sistemas de recomendação; bases de dados pequenas e dinâmicas.
SVM & Reg. Logística	Fronteiras geométricas claras; excelente em alta dimensionalidade; interpretável (pesos dos termos).	Análise de sentimentos; classificação jurídica e médica; cenários com alta dimensionalidade (TF-IDF).
Árvores & Ensembles	Baseado em regras lógicas; lida bem com dados heterogêneos; alta precisão (XGBoost).	Avaliação de legibilidade; classificação que combina texto com metadados estruturados.
Deep Learning (BERT)	Captura contexto e ordem; exige alto poder computacional (GPU); estado da arte em precisão.	Categorização de alta complexidade semântica; tarefas que exigem compreensão profunda do contexto.
LLMs (Prompt)	<i>Zero/Few-shot</i> ; baseados em instruções; dispensam grandes bases de dados rotuladas.	Prototipagem rápida; categorias subjetivas ou mutáveis; cenários sem dados de treino disponíveis.

1.5 Avaliação de modelos

A avaliação de um classificador de texto é o estágio em que verificamos se o aprendizado na fase de treinamento (Seção 1.4) de fato resultou em um modelo capaz de generalizar para dados nunca vistos. Como discutido no Capítulo [Avaliação de tecnologias de linguagem](#), a escolha de uma métrica isolada pode ser enganosa. Portanto, é necessário um protocolo de validação robusto para garantir que o desempenho reportado seja estatisticamente confiável e livre de vícios.

1.5.1 Metodologias de validação

O princípio fundamental da avaliação é a separação entre os dados usados para o aprendizado e os usados para a medição do desempenho. Caso o modelo seja testado com os mesmos dados de treinamento, estaríamos medindo apenas sua capacidade de memorização (*overfitting*) e não de categorização.



1.5.1.1 Divisão por repartição (*Hold-out*)

A forma mais comum de validação, especialmente em grandes *corpora* ou ao utilizar modelos de *Deep Learning* (Seção 1.4.5), é a divisão simples em subconjuntos. A literatura recomenda, usualmente, uma proporção de 80% para treino e 20% para teste (Murphy, 2012), embora variações como 70/30 também sejam frequentes.

Para um ajuste fino de modelos, como a escolha de hiperparâmetros (o valor de k no KNN ou o *learning rate* em LLMs), subdividimos o treinamento para criar um conjunto de Validação. Assim, temos três divisões do conjunto de dados:

- **Treino:** partição dos dados utilizada para ajustar os pesos do classificador.
- **Validação:** partição dos dados usada para escolher a melhor configuração do modelo.
- **Teste:** partição dos dados utilizada para a avaliação final e definitiva do desempenho.

1.5.1.2 Validação cruzada (*Cross-Validation*)

Quando o conjunto de dados é pequeno, a divisão simples pode gerar resultados instáveis dependendo de quais documentos caíram no teste. A Validação Cruzada de k partições (*k-fold cross-validation*) resolve isso ao dividir os dados em k subconjuntos (*folds*). O experimento é repetido k vezes: em cada rodada, uma partição diferente é usada para teste e as outras $k - 1$ para treinamento.

Por exemplo, em um *corpus* de 500 notícias dividido em 5 partições ($k = 5$), cada partição teria 100 documentos. Na primeira rodada, a Partição 1 é o teste e as outras 4 partições são usadas para treinamento; na segunda, a Partição 2 é o teste, e assim sucessivamente. Ao final, reportamos a média e o desvio padrão dos resultados, o que nos dá uma visão da variância e da estabilidade do modelo frente a diferentes recortes dos dados.

1.5.1.3 Estratificação e Significância

Um aspecto crítico no PLN é o desequilíbrio de classes. Em um portal de notícias, pode haver muito mais textos sobre “Política” do que sobre “Astronomia”. Se a divisão dos dados for feita de forma puramente aleatória, corremos o risco de a partição de teste não conter nenhum exemplo de uma classe rara. Para evitar isso, utilizamos o *Particionamento Estratificado*, que garante que a proporção original de cada categoria seja mantida em todos os subconjuntos de treino e de teste.

Para comparar se um modelo é genuinamente superior a outro, além da média, utilizamos técnicas como o *bootstrapping*. Nesta técnica, geramos múltiplas amostras aleatórias com reposição a partir do conjunto original (Efron; Tibshirani, 1993). Isso nos permite calcular intervalos de confiança e afirmar, com suporte estatístico, se a diferença de desempenho entre, por exemplo, um SVM e um Naïve Bayes, não ocorreu meramente ao acaso.

1.5.2 Métricas de desempenho

Para quantificar a eficácia, utilizamos métricas derivadas da *Matriz de Confusão*. Ela é uma tabela que cruza as predições do modelo com os rótulos reais, conforme ilustrado na Tabela 1.3.



Tabela 1.3: Exemplo de matriz de confusão para três classes (A, B e C).

Classe Real \ Predita	Classe A	Classe B	Classe C
Classe A	50 (VP)	2 (FN)	3 (FN)
Classe B	4 (FP)	45	6
Classe C	1 (FP)	5	49

A partir da matriz, definimos três conceitos fundamentais para cada classe:

- **Verdadeiro Positivo (VP)**: documentos da classe A corretamente classificados como A.
- **Falso Positivo (FP)**: documentos de outras classes classificados erroneamente como A (“falso alarme”).
- **Falso Negativo (FN)**: documentos que eram da classe A, mas o modelo atribuiu-os a outra categoria (“omissão”).

Idealmente, a matriz de confusão deve conter a maior quantidade de instâncias na diagonal principal, indicando que o modelo classificou a maior parte das instâncias na classe correta. A partir da matriz de confusão é possível definir a **acurácia**, uma das métricas mais usadas para avaliar o desempenho de um modelo em classificação. O objetivo da acurácia é medir a taxa global de acertos a partir da seguinte equação $Acurácia = \frac{VP+VN}{VP+VN+FP+FN}$.

1.5.2.1 Precisão, Revocação e Medida-F

Diferente da *Acurácia*, que mede apenas a taxa global de acertos, as seguintes métricas são mais informativas:

1. **Precisão (P)**: proporção de acertos nas predições de uma classe. Indica a confiabilidade do modelo ao atribuir um rótulo.

$$P = \frac{VP}{VP + FP}$$

2. **Revocação (R)**: proporção de documentos de uma classe que o modelo conseguiu recuperar. Indica a abrangência (ou cobertura) do modelo.

$$R = \frac{VP}{VP + FN}$$

3. **Medida-F1 (F_1)**: média harmônica entre Precisão e Revocação, equilibrando os dois aspectos. A média harmônica (e não a aritmética) é utilizada para punir valores extremos. Se um modelo tem Precisão 1.0 e Revocação 0.0, a média aritmética seria 0.5 (enganosa), enquanto a média harmônica seria 0 (realista).

$$F1 = 2 \times \frac{P \times R}{P + R}$$



1.5.2.2 Agregação: Macro e Micro

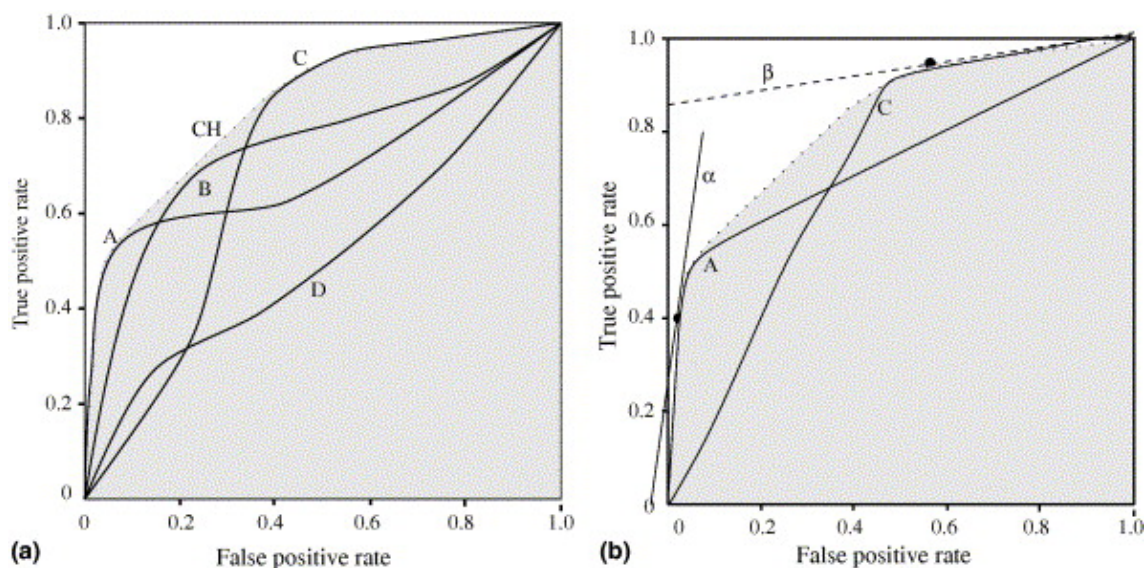
Em problemas multiclasse, existem duas formas de resumir o desempenho global:

- **Macro-Média:** calcula a métrica para cada classe individualmente e calcula a média simples. Trata todas as classes como igualmente importantes, sendo a melhor escolha para avaliar o desempenho em classes raras.
- **Micro-Média:** agrega as contagens globais de VP, FP e FN de todas as classes antes do cálculo. É dominada pelas classes mais frequentes, o que reflete o desempenho médio por documento.

1.5.3 Análise via Curva ROC e AUC

A Curva ROC (*Receiver Operating Characteristics*) é uma ferramenta visual para analisar o desempenho de classificadores probabilísticos (como a Regressão Logística ou o Naïve Bayes), variando o limiar de decisão t (Fawcett, 2006). A curva coloca em gráfico a Taxa de Verdadeiros Positivos ($TPR = \frac{VP}{VP+FN}$) no eixo Y e a Taxa de Falsos Positivos ($FPR = \frac{FP}{FP+VN}$) no eixo X. Um classificador perfeito alcançaria o canto superior esquerdo (100% de acerto com 0% de alarme falso). A Figura 1.5 ilustra dois exemplos de curvas ROC de diferentes modelos extraídos de (Fawcett, 2006). Visualmente, quanto mais próxima a curva estiver do canto superior esquerdo do gráfico, melhor o desempenho do modelo, pois isso indica uma alta taxa de verdadeiros positivos e uma baixa taxa de falsos positivos. Na Figura 1.5 (a), por exemplo, nota-se que as curvas de diferentes classificadores se cruzam; isso implica que um modelo pode ser superior a outro dependendo do limiar de decisão escolhido ou das restrições de custo do problema. Já a Figura 1.5 (b) demonstra o conceito de *Convex Hull*, que representa o limite superior de desempenho possível ao combinar diferentes modelos.

Figura 1.5: Exemplos de Curva ROC



Fonte: (Fawcett, 2006)



Para resumir essa curva em um único valor, utilizamos a AUC (*Area Under the Curve*). Um valor de 1, 0 indica separação perfeita, enquanto 0, 5 indica um desempenho equivalente ao acaso. A AUC é especialmente útil em PLN para comparar modelos em tarefas de detecção binária, como a identificação de notícias falsas Capítulo 8.

1.6 Recursos em português

A aplicação prática dos métodos de categorização para a língua portuguesa exige recursos capazes de captar as nuances morfosintáticas e semânticas da nossa língua. Como discutido no Capítulo [Conjunto de dados, dataset e corpus](#), a qualidade do modelo é diretamente proporcional à qualidade dos dados de treino. Nesta seção, detalhamos os principais conjuntos de dados, modelos pré-treinados e ferramentas disponíveis para o processamento do português brasileiro.

1.6.1 Conjuntos de dados para categorização

A comunidade brasileira de PLN tem sido prolífica na criação de recursos que abrangem desde a classificação de tópicos tradicionais até tarefas complexas de inferência e detecção de viés. A Tabela 1.4 sintetiza alguns desses recursos, detalhando a tarefa, o número de classes e, quando disponível, o desempenho de referência.

Tabela 1.4: Lista de recursos em português para classificação.

Nome	Tarefa	Classes	Resultados
CHAVE (Santos; Rocha, 2004a)	Classificação multirrótulo de notícias	9	–
Fake.Br (Monteiro et al., 2018)	Classificação de notícias (falsas/reais)	2	Acurácia/F1: 0,950 (Pires et al., 2024)
B2W-Reviews01 (Real et al., 2019)	Críticas de consumidores (escalas de 1 a 5)	5	–
UTLCorpus (Sousa et al., 2019)	Utilidade de críticas (apps e filmes)	2	ROC-AUC: 97,1 (apps) / 92,4 (filmes) (Souza; Filho, 2022)
Decisões Corte (Lage-Freitas et al., 2019)	Sumários judiciais com decisões	3	F1: 0,80
ASSIN 2 (Real et al., 2020)	Inferência textual (premissa/hipótese)	2	F1: 0,89 (Silva et al., 2023b)
HateBR (Vargas et al., 2022a)	Discurso de ódio no Instagram	2	F1: 0,85
FactNews (Vargas et al., 2023)	Viés vs. Factualidade e tipo de viés	2 / 12	F1: 0,88 / 0,67
LegiSubject-Br (Nunes, 2024)	Projetos de lei (multirrótulo)	59	–

No domínio “Jornalístico e Combate à Desinformação”, o CHAVE (Santos; Rocha, 2004a) é um marco histórico no português, oferecendo um desafio de classificação multirrótulo em notícias. No cenário contemporâneo de combate à desinformação (Capítulo 8), o Fake.Br (Monteiro et al., 2018) consolidou-se como um recurso essencial, apresentando resultados de alto desempenho com modelos modernos, como indicado por Pires et al. (2024).



Em “Análise de Sentimento e Consumo”, para o estudo de opiniões, o B2W-Reviews01 (Real et al., 2019) oferece uma granularidade de 5 classes (estrelas), permitindo avaliar a satisfação do cliente. Complementarmente, o UTLCorpus (Sousa et al., 2019) foca não no conteúdo da crítica em si, mas em sua utilidade (*helpfulness*), um problema de classificação binária que ajuda a filtrar informações relevantes em plataformas de aplicativos e de filmes.

Na área de “Inferência e Fenômenos Sociais”, o ASSIN 2 (Real et al., 2020) desafia os modelos a identificar relações de inferência textual, essenciais para tarefas de semântica profunda. No campo das redes sociais, o HateBR (Vargas et al., 2022a) fornece dados para a detecção de discurso de ódio no Instagram, enquanto o FactNews (Vargas et al., 2023) permite classificar o viés em frases, indo além da polaridade para identificar tipos específicos de viés jornalístico.

No “Domínio Jurídico e Legislativo”, os recursos “Decisões da Corte Brasileira” (Lage-Freitas et al., 2019) e LegiSubject-Br-Keywords (Nunes, 2024) são fundamentais para a automação e organização do sistema jurídico. O último destaca-se pela alta complexidade, envolvendo 59 categorias multirrótulo para indexação de projetos de lei.

1.6.2 Modelos de linguagem pré-treinados

A revolução dos Transformers (Seção 1.4.7) trouxe modelos treinados em volumes massivos de texto em português, fundamentais para os desempenhos reportados na Tabela 1.4.

- **BERTimbau** (Souza et al., 2020a): o modelo BERT (Devlin et al., 2019) padrão para o português brasileiro. Treinado no *corpus* brWaC (Wagner Filho et al., 2018), ele é a base para a maioria dos classificadores de alto desempenho da atualidade.
- **PTT5** (Carmo et al., 2020a): uma versão do T5 (Raffel et al., 2020) que permite abordar a classificação como um problema de geração de texto para texto.
- **MariTalk**⁷: representa a nova geração de LLMs nacionais, facilitando a classificação via *prompt* (Seção 1.4.8).

1.6.3 Ferramentas e Bibliotecas

Para implementar classificadores com esses recursos, são comumente empregadas as bibliotecas *Hugging Face Transformers* (para modelos neurais)⁸, *spaCy* (para modelos industriais prontos)⁹ e *Scikit-learn* (para métodos clássicos e métricas de avaliação)¹⁰.

1.7 Desafios e Tendências

Embora a categorização de texto seja uma das áreas mais maduras do PLN, o cenário contemporâneo – marcado pela onipresença de modelos de linguagem de larga escala (LLMs) e pela aplicação de IA em domínios críticos – impõe novos desafios. Nesta seção, discutimos as fronteiras da pesquisa e as tendências que moldarão a próxima geração de sistemas de categorização.

⁷<https://chat.maritaca.ai/>

⁸`pip install transformers`

⁹<https://spacy.io/>

¹⁰<https://scikit-learn.org/stable/>



1.7.1 Dinamismo e Desvio de Conceito (*Concept Drift*)

Modelos de classificação não são estáticos; operam em um mundo em que a linguagem e os conceitos evoluem. O *Desvio de Conceito* ocorre quando a relação entre as características do texto e as categorias muda ao longo do tempo. Imagine um classificador de notícias políticas treinado em 2010. Seus padrões de vocabulário e de nomes de entidades estarão obsoletos para classificar o cenário de 2026. Em redes sociais, gírias e neologismos surgem semanalmente, exigindo que o sistema de categorização possua estratégias de *Aprendizado Contínuo* para se adaptar, sem necessidade de ser retreinado do zero constantemente (Silva et al., 2022b; Silva; Almeida, 2021).

1.7.2 Escalabilidade e Eficiência energética

A escalabilidade permanece como um dos desafios centrais na categorização de textos, especialmente diante da necessidade de processar documentos cada vez mais extensos e volumes massivos de dados (Minaee et al., 2021). Embora a evolução dos modelos de linguagem de larga escala (LLMs) tenha permitido a expansão da janela de contexto, observa-se que a eficácia preditiva e a performance do modelo tendem a degradar à medida que o tamanho da entrada aumenta (Dong et al., 2024).

Paralelamente ao desafio técnico, o debate sobre a **IA Verde** (*Green AI*) tornou-se fundamental para a sustentabilidade da área (Schwartz et al., 2020). Sistemas que utilizam modelos com bilhões de parâmetros exigem infraestrutura computacional de alto desempenho e um consumo energético que pode ser proibitivo para muitas instituições. Mesmo modelos recentes e otimizados para tarefas de raciocínio, como os da família Qwen (Bai et al., 2023), ainda demandam um consumo de energia considerável, evidenciando que a busca por um equilíbrio entre alto desempenho e eficiência operacional é a nova fronteira no desenvolvimento de tecnologias de linguagem.

1.7.3 Explicabilidade e Interpretabilidade (XAI)

À medida que migramos de modelos lineares para redes neurais profundas e LLMs, ganhamos em precisão, mas perdemos em transparência. Em domínios sensíveis como a saúde ou o direito, uma classificação sem justificativa é insuficiente. A área de *IA Explicável* (XAI) busca desenvolver métodos que permitam entender *por que* um modelo atribuiu um rótulo específico. Técnicas como mapas de atenção (que mostram quais palavras o Transformer priorizou) ou métodos de atribuição de características (como LIME (Ribeiro et al., 2016) ou SHAP (Lundberg; Lee, 2017)) são essenciais para que o especialista humano possa confiar nas decisões automáticas e auditar as decisões automáticas. Embora para texto, estudos indicam que estes métodos se mostram pouco confiáveis (Létoffé et al., 2025). Modelos transparentes são um objetivo que muitos pesquisadores desejam e, portanto, ainda um caminho aberto de pesquisa em PLN.

1.7.4 Ética, Equidade (*Fairness*) e Mitigação de vieses

Como discutido no Capítulo *Questões éticas em IA e PLN*, modelos de PLN aprendem com dados gerados por humanos, o que pode significar que herdamos e amplificamos preconceitos históricos de gênero, raça ou classe. Classificadores de currículos ou de sentimentos podem apresentar disparidades de desempenho ou rotulações injustas devido a vieses nos *corpora* de treino. A tendência atual é o desenvolvimento de *benchmarks* de equidade específicos



para o português brasileiro e de técnicas de *desenviesamento* (*debiasing*) que garantam que a categorização seja justa para todos os grupos demográficos.

1.7.5 Soberania digital

A dependência de modelos globais de larga escala suscita questões sobre a soberania dos dados brasileiros. Há um movimento crescente em direção ao desenvolvimento de modelos nacionais robustos, como a MariTalk (Sales Almeida et al., 2024), e de técnicas de *destilação de conhecimento* (Gou et al., 2021). O objetivo é criar classificadores “pequenos e inteligentes” que consigam manter o desempenho de um LLM, mas que possam ser executados localmente em servidores brasileiros ou dispositivos móveis, com menor latência e custo.

1.7.6 Categorização em baixos recursos e dialetos

O Brasil é um país de diversidade linguística imensa. A maioria dos recursos da Seção 1.6 foca no português padrão. Um desafio aberto é a categorização eficaz de textos em dialetos regionais, variedades populares e línguas indígenas brasileiras, casos em que a escassez de dados rotulados impede o uso de métodos conexionistas tradicionais (Finger, 2021), exigindo avanços em aprendizado *zero-shot* e na transferência de conhecimento.

1.8 Referências para aprofundamento

Dada a amplitude e a rápida evolução do campo da classificação de texto, este capítulo oferece uma base sólida, mas não esgota as diversas nuances e os desenvolvimentos recentes da área. Para os leitores que desejam expandir seus conhecimentos teóricos e práticos, recomendamos as obras mencionadas a seguir.

Para uma compreensão aprofundada das bases teóricas e dos métodos clássicos de classificação e recuperação de informação, a obra de Manning; Schütze (1999) permanece como uma referência fundamental, detalhando algoritmos que formaram a base do PLN moderno. Para uma visão panorâmica e atualizada sobre o impacto dos modelos *Transformers* e das LLMs na tarefa de categorização, o levantamento realizado por Fields et al. (2024) constitui um importante guia para mapear o estado da arte. Por fim, os aspectos técnicos e comparativos entre as estratégias contemporâneas de adaptação de modelos, como o *fine-tuning* e o *instruction-tuning*, podem ser aprofundados no trabalho de Chae; Davidson (2025).



Capítulo 2

Recuperação de Informação

Viviane P. Moreira

Publicado em: 26/09/2023

2.1 Introdução

A necessidade de organizar a informação é inerente à espécie humana – bibliotecas existem desde pelo menos o ano 2600 AC. Dado o grande volume de informação que as bibliotecas armazenam, a partir dos anos 1960, iniciaram-se esforços a fim de automatizar o armazenamento e a busca de materiais bibliográficos através da computação. Esses esforços marcaram o início da área de recuperação de informação (RI). A RI trata de encontrar, a partir de grandes coleções, material (geralmente documentos) de natureza não estruturada (geralmente texto) que satisfaça uma necessidade de informação (Manning et al., 2008b). Em outras palavras, o objetivo central da RI é a busca, ou seja, a tarefa de encontrar material **relevante** a partir de uma consulta de um usuário. Esta tarefa é comumente conhecida por **recuperação ad hoc**. Apesar de a RI poder ser aplicada a diferentes tipos de dados não estruturados como imagem, áudio, vídeo etc. o foco deste capítulo é **informações textuais**.

2.1.1 Relação com o PLN

Há muita interseção entre RI e PLN, pois ambas lidam com a linguagem humana. Contudo, há algumas diferenças. A primeira diferença é quanto à origem – enquanto que o PLN teve origem na inteligência artificial e na linguística computacional, a RI teve origem na biblioteconomia e na ciência da informação¹. Outra diferença é quanto ao escopo – podemos dizer que o escopo do PLN é mais abrangente (i.e., compreensão e geração de linguagem) enquanto que o da RI é mais restrito às tarefas relacionadas à busca por informação.

2.1.2 O Foco da Recuperação de Informação

A **tarefa central** da RI é casar a consulta de um usuário com os documentos que são potencialmente relevantes a ela. A principal dificuldade é que os termos utilizados pelo usuário podem não ter sido usados nos documentos relevantes. Este problema é conhecido como **incompatibilidade de vocabulário** (*vocabulary mismatch*) e é decorrente de dois fenômenos comuns na linguagem: a **sinonímia** e a **polissemia**. A sinonímia refere-se ao fato de usarmos palavras diferentes para nos referirmos ao mesmo conceito, e.g., “bergamota”, “tangerina” e “mexerica”. O problema para RI é que uma consulta pelo termo

¹Para um relato mais detalhado sobre as origens da RI, há o artigo de Lesk (1995).



“bergamota” não consegue recuperar documentos relevantes que contenham “tangerina”. A polissemia refere-se ao fato de que uma mesma palavra pode apresentar sentidos distintos, e.g., “manga” que tanto pode ser a fruta como a parte da vestimenta que cobre o braço da pessoa. O efeito negativo da polissemia para a RI é a recuperação de documentos que contêm a palavra pesquisada, mas não no sentido pretendido. Ao longo dos anos, houve uma vasta gama de propostas de solução para este problema. Uma visão geral dessas propostas é fornecida na Seção 2.5.

Até meados dos anos 1990, o interesse em RI estava restrito a bibliotecários, jornalistas e profissionais do direito (i.e., profissões que tinham bastante necessidade de buscar informações). Com a popularização da Internet e dos motores de busca para a web, a RI ganhou muita importância. Sistemas de RI fazem parte da vida diária de uma grande parte da população mundial. Há estimativas que o Google, o motor de busca mais utilizado, receba cerca de 100 mil consultas por segundo e tenha 4,3 bilhões de usuários². Os desafios de se lidar com coleções contendo bilhões de documentos (i.e., páginas web) motivaram o desenvolvimento de novos algoritmos e técnicas, objetivando tanto eficiência (baixo custo computacional) quanto eficácia (qualidade do resultado).

Um sistema de RI também pode ser utilizado como um componente em tarefas de PLN como sistemas de perguntas e respostas e de detecção de plágio. A vantagem é que a RI consegue recuperar documentos candidatos com um custo computacional baixo. Assim, as fases subsequentes que requerem comparações exaustivas usando modelos mais custosos podem trabalhar com um conjunto menor de documentos.

2.1.3 O Conceito de Relevância

Vimos na Seção 2.1.2 que a tarefa central de RI é recuperar itens que sejam **relevantes** a uma necessidade de informação expressa por meio de palavras-chave. A relevância é um julgamento feito **pelo usuário** que indica o quão bem um documento satisfaz a consulta. Em sua forma mais simples, ela pode ser tratada como binária – cada documento é considerado como **relevante** ou como **não relevante**. Também é possível utilizar diversos níveis de relevância como por exemplo: “muito relevante”, “moderadamente relevante”, “marginalmente relevante” e “não relevante”.

É comum ouvirmos que um motor de busca ou sistema de RI retorna os resultados em **ordem de relevância**. Esta afirmação não está tecnicamente correta pois a **relevância é subjetiva** e só pode ser atribuída pelo usuário. Dois usuários podem fazer a mesma consulta q e julgar um mesmo documento d de forma diferente – enquanto um usuário considera que d é relevante, o outro usuário pode considerar que d não é relevante para q . Os sistemas de RI ordenam os documentos utilizando uma função de ranqueamento baseada em evidências que o sistema consegue calcular (e.g., estatísticas de ocorrências das palavras nos documentos, tamanho dos documentos etc.). Idealmente, espera-se que esta função de ranqueamento esteja positivamente correlacionada com a ideia subjetiva de relevância dos usuários.

2.1.4 Principais Livros

A RI é uma área de pesquisa por si só e este capítulo não pretende ser uma revisão exaustiva. vamos focar nos aspectos centrais da área e na sua interseção com PLN. Há

²<https://wpdevshed.com/how-many-people-use-google/>



diversos livros em **língua inglesa** que são referência em RI. Para o leitor que deseja saber mais sobre a área, indicamos os seguintes livros:

- Introduction to Information Retrieval de Manning, Schütze e Raghavan (Manning et al., 2008b), publicado em 2008 pela Cambridge University Press e disponível gratuitamente online³.
- Search engines: Information retrieval in practice de Croft, Metzler, e Strohman (Croft et al., 2010), publicado em 2010 pela Addison-Wesley e disponível gratuitamente online⁴.
- Modern Information Retrieval - the concepts and technology behind search de Baeza-Yates e Ribeiro-Neto (Baeza-Yates; Ribeiro-Neto, 2011), publicado em 2011 pela Addison-Wesley.

Em português, há uma versão resumida do livro Modern Information Retrieval publicada em 2013 (Baeza-Yates; Ribeiro-Neto, 2013). Esta versão contém os principais capítulos do livro original.

2.1.5 Organização deste Capítulo

Nesta primeira versão do livro **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**, o capítulo sobre Recuperação da Informação tem como foco as operações de pré-processamento e indexação (Seção 2.2), nos modelos clássicos (Seção 2.3), no paradigma de avaliação (Seção 2.4) e em técnicas de melhoria de qualidade (Seção 2.5). Também indicamos algumas bibliotecas e ferramentas amplamente utilizadas (Seção 2.6). As próximas versões do livro irão incluir tópicos avançados como o ranqueamento neural.

2.2 Visão Geral de um Sistema Típico de Recuperação de Informação

A Figura 2.1 mostra uma visão geral do processo de RI. Em uma ponta, há a usuária que tem uma **necessidade de informação** e, na outra ponta, temos uma coleção de documentos textuais. A consulta da usuária é submetida a um sistema de RI (representado pelo retângulo laranja). As Etapas 1 e 2 ocorrem offline, pois o índice precisa estar pronto antes que o sistema possa receber consultas. Na Etapa 1, os documentos da coleção são pré-processados para então, na Etapa 2, serem indexados. Na Etapa 3, o sistema executa sobre a consulta as mesmas operações de pré-processamento que foram aplicadas na Etapa 1. O texto pré-processado da consulta é utilizado, na Etapa 4, para buscar no índice os documentos que mais bem atendam a consulta. Os resultados da consulta são então retornados à usuária sob a forma de uma lista ordenada.

2.2.1 Pré-Processamento

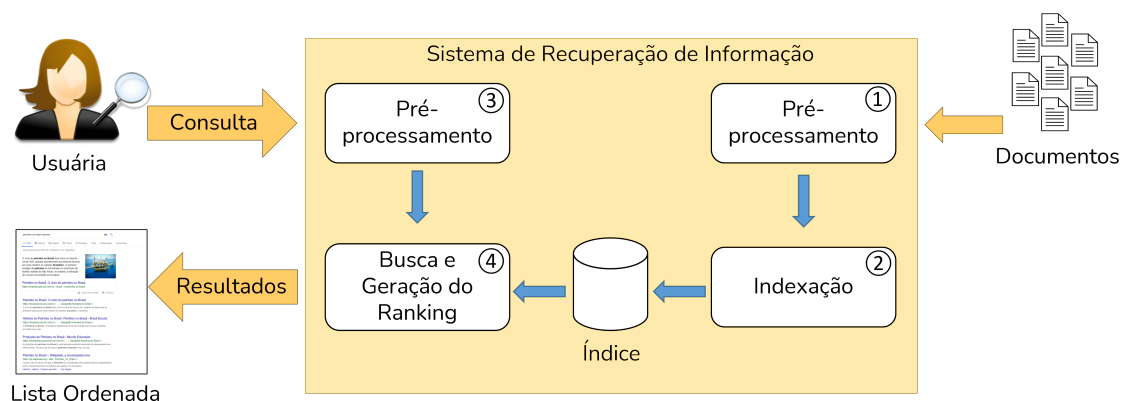
As operações de pré-processamento são muito semelhantes às estudadas para as tarefas de PLN relatadas no Capítulo **Sequência de caracteres e palavras**. É preciso definir a nossa **unidade de indexação**, ou seja, o que é um **documento** no contexto tratado.

³<https://nlp.stanford.edu/IR-book/>

⁴<https://ciir.cs.umass.edu/irbook/>



Figura 2.1: Visão Geral do Processo de RI.



Por exemplo, uma notícia, um comentário, um e-mail, um artigo científico, um livro, um capítulo do livro etc.

2.2.1.1 Tokenização

Dada uma sequência de caracteres e uma unidade de indexação, a **tokenização** irá separar a sequência em *tokens* (i.e., unigramas de palavras). Nesse processo, é comum descartar caracteres de pontuação e mantermos apenas as palavras (i.e., também chamadas de termos). Embora à primeira vista este processo pareça bastante trivial (i.e., podemos apenas considerar que os caracteres não alfabéticos como símbolos e sinais de pontuação sejam separadores), há vários casos especiais que podemos precisar tratar. Por exemplo, endereços de email, emojis, e termos que misturam letras e símbolos como C++.

O processo de tokenização costuma ser implementado utilizando-se expressões regulares e há vários tokenizadores prontos em bibliotecas como NLTK⁵ e spaCy⁶. Após a tokenização, é comum converter todos os caracteres para letra minúscula. O objetivo é fazer com que a busca não seja *case-sensitive*, isto é, impactada por texto em letras maiúscula e minúscula. Por outro lado, essa operação dificulta a identificação de entidades nomeadas, uma vez que entidades são geralmente nomes próprios com grafia em letra maiúscula.

A remoção de acentos, cuja denominação mais apropriada é **remoção de sinais diacríticos** também pode ser realizada, principalmente quando se trata de textos informais nos quais esses sinais são menos utilizados. Um sinal diacrítico é uma marca que colocamos sobre ou sob o caractere como acento agudo, grave, circunflexo, til ou cedilha.

2.2.1.2 Remoção de *Stop Words*

As *stop words* são palavras que têm pouca utilidade para a RI pois ocorrem em muitos documentos e assim não servem para distinguir o conteúdo semântico dos documentos. Preposições, conjunções, artigos e verbos de ligação são normalmente consideradas como *stop words*. Essas palavras são muitas vezes completamente descartadas em um processo conhecido como **remoção de *stop words***.

O processo de remoção de *stop words* é bastante simples e consiste em verificar a presença de cada palavra do texto em uma lista de *stop words* previamente construída. Existem

⁵<https://www.nltk.org/>

⁶<https://spacy.io/models/pt>

listas prontas que podemos usar, como a do NLTK, por exemplo, que contém 204 palavras. Contudo, é recomendável revisar se as listas contêm palavras que podem ser úteis para o contexto da aplicação sendo desenvolvida. Por exemplo, na lista elaborada pela Linguateca⁷ com as palavras mais frequentes dos textos da Folha de São Paulo, encontramos a palavra “brasil”. Esta palavra, muito provavelmente, não deve ser removida dos textos da coleção de documentos.

O principal argumento a favor da remoção de *stop words* é a redução do tamanho do vocabulário pois elas representam cerca de 40% das ocorrências de palavras em um *corpus*. Por outro lado, essa remoção pode trazer uma perda de informação relevante. Se pensarmos na famosa expressão “ser ou não ser, eis a questão”, com a remoção de *stop words*, sobraria apenas “questão”, o que descaracteriza completamente a expressão. O impacto negativo na busca seria que o sistema não mais conseguiria distinguir entre documentos que contenham a expressão completa daqueles que contêm apenas a palavra “questão”.

2.2.1.3 Stemming

O objetivo do *stemming* é gerar uma mesma representação para formas variantes de uma mesma palavra através da remoção dos sufixos. Por exemplo, removendo-se os sufixos de “apresentação”, “apresentando”, e “apresentar”, obteríamos o radical “apresent”. Com isso, uma usuária buscando por “apresentar artigos científico” consegue recuperar um documento com o trecho “apresentando artigos científicos”. O benefício é aumentar o número de documentos relevantes recuperados em resposta à consulta.

O processo de *stemming* e seu impacto sobre RI vêm sendo estudado há diversos anos. Para a língua inglesa, o primeiro algoritmo de *stemming*, ou *stemmer*, data de 1968 e foi proposto por Julie Beth Lovins (Lovins, 1968). Em 1980, Martin Porter propôs o Porter Stemmer (Porter, 1980) que mostrou obter bons resultados (também para a língua inglesa) e posteriormente foi traduzido para outros idiomas incluindo o português⁸.

O primeiro *stemmer* desenvolvido especialmente para a língua portuguesa foi o Removedor de Sufixos da Língua Portuguesa (RSLP)⁹ (Orengo; Huyck, 2001). Posteriormente veio o STEMBR (Alvares et al., 2005). Há resultados experimentais que mostram a validade da aplicação de *stemmers* para RI em português (Flores et al., 2010; Orengo et al., 2006). Muitas vezes, na média de um conjunto de consultas, o impacto do *stemming* pode ser pequeno. Contudo, ao analisarmos consultas individuais, percebemos melhorias muito grandes em alguns casos. Em outros, podem aparecer mais documentos não relevantes recuperados pois a remoção do sufixo pode gerar ambiguidade. Os resultados mostram que *stemmers* menos agressivos, ou seja, com menos regras de remoção tendem a gerar resultados melhores pois têm menos impacto negativo na precisão.

O Quadro 2.1 mostra um trecho do livro O Tempo e o Vento – O Continente, de Érico Veríssimo, e suas versões após cada etapa do pré-processamento. Observemos a redução de 51 para 27 *tokens* com a remoção de *stop words*. O *stemmer* utilizado no exemplo foi o Snowball¹⁰. Também vale ressaltar a potencial introdução de ambiguidade com a redução de algumas palavras, por exemplo de “legalistas” para “legal”.

⁷<https://www.linguateca.pt/chave/stopwords/>

⁸<http://snowball.tartarus.org/algorithms/portuguese/stemmer.html>

⁹<https://www.vivianemoreira.org/rslp>

¹⁰<https://snowballstem.org/>



Quadro 2.1: Excerto do livro O Tempo e o Vento – O Continente, de Érico Veríssimo, e suas versões após cada etapa do pré-processamento.

Original	Quando pela primeira vez aparecera em Santa Fé, no ano em que fora assinada a paz entre farroupilhas e legalistas, causara a pior das impressões. Chegara escoteiro, montado num cavalo magro e manco, e fazendo questão de mostrar a toda a gente que tinha as guaiacas atestadas de moedas de ouro.
Tokenizado	Quando pela primeira vez aparecera em Santa Fé no ano em que fora assinada a paz entre farroupilhas e legalistas causara a pior das impressões Chegara escoteiro montado num cavalo magro e manco e fazendo questão de mostrar a toda a gente que tinha as guaiacas atestadas de moedas de ouro
Convertido para minúsculo	quando pela primeira vez aparecera em santa fé no ano em que fora assinada a paz entre farroupilhas e legalistas causara a pior das impressões chegara escoteiro montado num cavalo magro e manco e fazendo questão de mostrar a toda a gente que tinha as guaiacas atestadas de moedas de ouro
Após remoção de <i>stop words</i>	primeira vez aparecera santa fé ano assinada paz farroupilhas legalistas causara pior impressões chegara escoteiro montado cavalo magro manco fazendo questão mostrar gente guaiacas atestadas moedas ouro
Após <i>stemming</i>	primeir vez aparec sant fé ano assin paz farroupilh legal caus pior impressõ cheg escoteir mont caval magr manc faz questã mostr gent guaiac atest moed our

2.2.2 Indexação

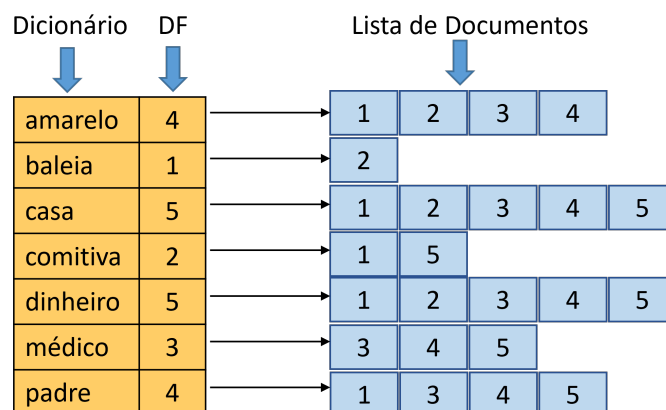
A forma que temos para evitar ter de “varrer” os textos da coleção de documentos em busca dos termos da consulta (o que seria um procedimento extremamente lento) é contar-mos com um **índice** previamente construído. A indexação recebe como entrada o texto pré-processado e cria um índice chamado de **arquivo invertido**. Um arquivo invertido assemelha-se ao índice remissivo que comumente encontramos ao final de um livro e contém as palavras (únicas) do texto e a lista de documentos em que elas aparecem.

A Figura 2.2 mostra um exemplo de índice que pode ser usado por um sistema de RI. O índice tem dois componentes: (i) o dicionário, também é chamado de vocabulário, que possui a lista de termos únicos da coleção de documentos e (ii) a lista de identificadores de documentos em que o termo aparece (conhecida por lista de *postings*). Junto ao dicionário, também armazena-se o número de documentos em que o termo aparece, conhecido como *document frequency* (*DF*). Observemos que o dicionário fica em ordem alfabética e a lista de *postings* fica em ordem de identificador de documento.

Nos casos em que desejamos fazer uma busca sobre um campo específico do documento (e.g., título ou autor), vamos precisar de **índices paramétricos**, um para cada campo. Além disso, se quisermos permitir buscas por múltiplos termos adjacentes (e.g., “memória prodigiosa”), precisaremos de um **índice posicional** que guarde também as posições em que os termos aparecem nos documentos. Certamente, tanto os índices paramétricos como os posicionais possuem um custo maior em termos de espaço de armazenamento e de processamento.



Figura 2.2: Exemplo de um trecho de índice armazenado como um arquivo invertido.



2.3 Modelos Clássicos de recuperação de informação

Um modelo de RI especifica como representar os documentos, as consultas e como compará-los. Ao longo dos anos, diversos modelos de RI foram propostos. Vamos explorar os modelos clássicos. Todos os modelos clássicos pressupõem que a distribuição dos termos nos documentos é independente e utilizam a abordagem *bag of words* (BoW). Em um BoW, a ordem dos termos nos documentos é desprezada. A vantagem é a simplificação dos modelos, os quais conseguem processar consultas de maneira mais rápida. A desvantagem é a perda de semântica em alguns casos pois as sentenças “João é mais velho do que José” e “José é mais velho do que João” têm representações idênticas apesar de significado oposto.

2.3.1 Modelo Booleano

O primeiro modelo que foi usado para RI foi o modelo Booleano. Trata-se de um modelo simples, baseado na teoria dos conjuntos e na álgebra booleana. As consultas são compostas por palavras-chave combinadas com os operadores AND, OR e NOT e o seu processamento utiliza o índice.

Para processar uma consulta do tipo p_1 AND p_2 , precisamos selecionar as listas de *postings* de p_1 e p_2 e computar a sua interseção. Por exemplo, com o trecho de índice da Figura 2.2, podemos ver que a consulta “comitiva” AND “médico” retornaria apenas o documento 5. Já a consulta “baleia” AND “padre” não retornaria nenhum documento.

Consultas com OR implicam na união de conjuntos. Por exemplo, considerando o índice da Figura 2.2, a consulta “comitiva” OR “médico” retornaria os documentos 1, 3, 4 e 5.

A principal limitação do modelo Booleano é que ele não é capaz de ordenar o resultado da consulta. Ou um documento satisfaz ou não satisfaz a expressão Booleana da consulta; não há a opção “satisfaz parcialmente”. Apesar dessa limitação, este foi o modelo comercial mais usado por três décadas, mesmo após a proposta de modelos superiores. Até hoje ainda vemos o modelo Booleano sendo usado, por exemplo, em alguns sistemas de bibliotecas.

2.3.2 Modelo Vetorial

Com o objetivo de poder ordenar os documentos em resposta às consultas, há duas premissas simples que podemos utilizar: (i) documentos que contêm mais vezes os termos da consulta têm mais chance de estarem relacionados a ela (e de serem relevantes) e (ii) os

termos mais raros na coleção são mais úteis para diferenciar o conteúdo dos documentos. A primeira premissa é atribuída ao pesquisador alemão Hans Peter Luhn e data de 1957. E a segunda premissa foi desenvolvida pela pesquisadora inglesa Karen Spärck Jones em 1972.

Com base nessas premissas, Gerard Salton propôs, na década de 1960, o modelo vetorial (em inglês, *vector space model*) (Salton; McGill, 1983). Nesse modelo, os documentos e as consultas são representados como vetores em um espaço de t dimensões, onde t é o número de termos distintos (i.e., o tamanho do dicionário). No espaço vetorial, os termos são os eixos. As consultas e os documentos são representados no espaço de acordo com a força da associação que eles têm com o termo. A força da associação do documento d_j com o termo i (w_{i,d_j}) é computada pelo esquema TF-IDF (*Term Frequency times Inverse Document Frequency*), segundo a Equação 2.1:

$$w_{i,d_j} = TF_{i,d_j} \times IDF_t = freq_{i,d_j} \times \log_{10} \frac{N}{DF_i} \quad (2.1)$$

onde $freq_{i,d_j}$ é o número de ocorrências do termo i no documento d_j , N é o número de documentos da coleção e DF_i é o número de documentos que contêm o termo i . É comum normalizar o componente TF para evitar que documentos longos sejam beneficiados, pois um documento longo tem mais chances de possuir os termos da consulta. A normalização do componente TF pode ser feita de diferentes formas: (i) dividindo-se o número de ocorrências do termo no documento pelo número de palavras no documento ou (ii) dividindo-se o número de ocorrências do termo no documento pelo número de ocorrências do termo mais comum naquele documento.

O componente IDF tem por objetivo dar um peso maior para os termos raros pois eles são mais úteis para discriminar o conteúdo dos documentos. Pensando em uma situação extrema, um termo que ocorra em todos os documentos não tem nenhuma utilidade para diferenciar um documento de outro. Sendo assim, valores de IDF mais altos serão atribuídos para termos que ocorram em poucos documentos. Note-se que o IDF não tem efeito caso a consulta tenha apenas um termo. Em consultas com dois ou mais termos, o efeito do IDF é dar mais importância para os termos que ocorrem em menos documentos. Por exemplo, imaginemos uma coleção que trata de instrumentos musicais. Um usuário faz a consulta “flauta bansuri”. Considerando que o termo “flauta” é muito mais comum do que o termo “bansuri” (que se refere a um tipo específico de flauta transversal), o efeito do IDF será atribuir bem mais importância ao termo “bansuri”. Os vetores gerados pelo TF-IDF são longos e esparsos. **Longos** pois o número de dimensões é igual ao número de palavras do vocabulário (o que pode chegar a centenas de milhares para algumas coleções). **Esparsos** pois a grande maioria dos elementos têm valor zero já que os documentos possuem uma pequena fração dos termos do vocabulário.

A similaridade entre um documento e uma consulta é dada pelo **cosseno** entre os seus vetores. O cosseno entre uma consulta q e um documento d_j é o produto escalar normalizado dos vetores $\vec{q}$ e $\vec{d}_j$, dado pela Equação 2.2:

$$cosseno(q, d_j) = \frac{\vec{q} \cdot \vec{d}_j}{|\vec{q}| \times |\vec{d}_j|} = \frac{\sum_{i=1}^t w_{i,q} \times w_{i,d_j}}{\sqrt{\sum_{i=1}^t w_{i,q}^2} \times \sqrt{\sum_{i=1}^t w_{i,d_j}^2}} \quad (2.2)$$

onde $w_{i,q}$ e w_{i,d_j} representam o peso do i -ésimo termo na consulta q e no documento d_j , respectivamente.



O cosseno é máximo (i.e., igual a 1) se os vetores possuem um ângulo de 0 graus e é mínimo (i.e., igual a 0) se os vetores formarem um ângulo de 90 graus¹¹. Em outras palavras, o cosseno será mínimo se o documento e a consulta não compartilharem nenhum termo. Vale ressaltar que não é preciso computar o cosseno entre a consulta e todos os documentos. Utilizando um índice (como o da Figura 2.2), podemos selecionar apenas os documentos que têm pelo menos uma das palavras da consulta, e calcular o cosseno somente entre a consulta e esses documentos. Também é possível estender essa noção e selecionar apenas documentos com valores de TF-IDF que ultrapassem um limiar para os termos da consulta ou restringir ainda mais e calcular o cosseno apenas para documentos que possuam duas ou mais das palavras da consulta.

Exemplo 2.1: Para ilustrar o conceito de TF-IDF e o ranqueamento dos documentos no modelo vetorial através do cosseno, vamos trabalhar com um exemplo. Nossa coleção de documentos é composta por cinco importantes livros da literatura brasileira. São eles:

- d_1 – O Alienista de Machado de Assis;
- d_2 – Vidas Secas de Graciliano Ramos;
- d_3 – O Continente (de O Tempo e o Vento) de Érico Veríssimo;
- d_4 – Capitães de Areia de Jorge Amado; e
- d_5 – Os Sertões de Euclides da Cunha.

Essa pequena coleção possui 350.260 *tokens*, e um vocabulário de 32.594 termos (já desconsiderando-se as *stop words*) – isso quer dizer que os documentos e as consultas são representados por vetores de 32.594 dimensões. Como seria inviável mostrarmos todos esses termos, iremos focar em sete termos selecionados: amarelo, baleia, casa, comitiva, dinheiro, médico e padre. A Tabela 2.1 mostra os sete termos selecionados juntamente com o número de documentos em que cada um deles aparece (DF). A partir do DF, podemos computar o IDF (segundo elemento da Equação 2.1). Por exemplo, o termo “comitiva” ocorre em dois documentos, então $IDF_{comitiva} = \log_{10} \frac{5}{2} = 0,3979$. Já para o termo “casa” que ocorre em todos os cinco documentos, o $IDF_{casa} = \log_{10} \frac{5}{5} = 0$.

Tabela 2.1: Exemplo de DF e IDF para algumas palavras do texto

Termo	DF	IDF
amarelo	4	0,0969
baleia	1	0,6990
casa	5	0,0000
comitiva	2	0,3979
dinheiro	5	0,0000
médico	4	0,0969
padre	4	0,0969

¹¹Note que o valor do cosseno irá ficar no intervalo [0,1] e não [-1,1] pois os escores gerados pelo TF-IDF são positivos.

Tabela 2.2: Exemplo de matriz de incidência de termos em documentos com valores de TF (i.e., o número de ocorrências dos termos nos documentos).

	d_1	d_2	d_3	d_4	d_5
Termo	O Alienista	Vidas Secas	O Tempo e O Vento	Capitães de areia	Os Sertões
amarelo	1	42	6	3	0
baleia	0	86	0	0	0
casa	109	37	247	120	30
comitiva	4	0	0	0	4
dinheiro	7	9	33	43	3
médico	18	0	157	7	8
padre	22	0	120	252	9

Tabela 2.3: Exemplo de matriz de incidência de termos em documentos com valores de TF-IDF

	d_1	d_2	d_3	d_4	d_5
Termo	O Alienista	Vidas Secas	O Tempo e O Vento	Capitães de areia	Os Sertões
amarelo	0,0009	0,0473	0,0024	0,0012	0,0000
baleia	0,0000	0,6990	0,0000	0,0000	0,0000
casa	0,0000	0,0000	0,0000	0,0000	0,0000
comitiva	0,0146	0,0000	0,0000	0,0000	0,0549
dinheiro	0,0000	0,0000	0,0000	0,0000	0,0000
médico	0,0160	0,0000	0,0616	0,0031	0,0258
padre	0,0196	0,0000	0,0471	0,0969	0,0291

Para computar os vetores de pesos TF-IDF para os cinco documentos da coleção, vamos iniciar com uma matriz de incidência dos termos nos documentos. Temos sete termos e cinco documentos; então, nossa matriz terá sete linhas e cinco colunas. Nessa matriz, ilustrada na Tabela 2.2, temos o número de ocorrências dos termos nos documentos. A seguir, vamos normalizar o componente TF dividindo-o pelo número de ocorrências do termo mais comum (i.e., o termo que mais ocorre no documento), que para d_1 é “casa”. Com isso, valor de TF-IDF para o termo “padre” no documento d_1 é $(22/109) \times 0,0969 = 0,0196$. Esses valores estão na Tabela 2.3. Note-se que os escores para os termos “casa” e “dinheiro” são iguais a zero pois eles aparecem em todos os cinco documentos, resultando em um IDF = 0. Também podemos observar que o maior escore é o do termo “baleia” no documento d_2 . Isso ocorre porque esse termo ocorre muitas vezes em d_2 e nenhuma vez nos outros quatro documentos – ele é, portanto, um termo que tem bastante poder de diferenciar esse documento dos demais.

Para exemplificar o cálculo do cosseno, vamos imaginar uma consulta $q = \{\text{comitiva}, \text{médico}\}$. Assim como os documentos, o vetor da consulta terá sete dimensões e será composto por pesos TF-IDF. Os termos que não aparecem na consulta têm peso zero. Assim, o vetor da consulta será $\vec{q} = [0; 0; 0; 0,3979; 0; 0,0961; 0]$. Para montar o *ranking* dos documentos em relação à consulta, calculamos o cosseno entre $\vec{q}$ e os vetores dos documentos que estão na Tabela 2.3 utilizando a Equação 2.2. Por exemplo, o cosseno entre $\vec{q}$ e $\vec{d}_1$ é:



$$\frac{0 \times 0,009 + 0 \times 0 + 0 \times 0 + 0,3979 \times 0,0146 + 0 \times 0 + 0,0969 \times 0,0160 + 0 \times 0,0196}{\sqrt{0^2 + 0^2 + 0^2 + 0,3979^2 + 0^2 + 0,0969^2 + 0^2} \times \sqrt{0,0009^2 + 0^2 + 0^2 + 0,0146^2 + 0^2 + 0,0160^2 + 0,0196^2}} =$$

$$\frac{0 + 0 + 0 + 0,0058 + 0 + 0,0016 + 0}{\sqrt{0 + 0 + 0 + 0,1583 + 0 + 0,0009 + 0} \times \sqrt{0 + 0 + 0 + 0,0002 + 0 + 0,0003 + 0,0004}} =$$

$$\frac{0,0074}{\sqrt{0,1677} \times \sqrt{0,0009}} = \frac{0,0074}{0,4096 \times 0,0292} = \frac{0,0074}{0,012} = 0,6156$$

Os escores de cosseno entre $\vec{q}$ e os cinco documentos estão na Tabela 2.4. Note que como q e d_2 não têm nenhum termo em comum, o escore desse documento é zero. Para montar o *ranking*, ordenamos os documentos em ordem decrescente de cosseno. Dessa forma, o primeiro do *ranking* seria d_5 , seguido por d_1 , d_3 e d_4 .

Tabela 2.4: Cosseno entre o vetor da consulta q e os vetores dos documentos da coleção.

d_1	d_2	d_3	d_4	d_5
0.6156	0.0000	0.1879	0.0075	0.8765

2.3.3 Modelos Probabilísticos

Há diversos modelos que utilizam abordagens para RI baseadas em um arcabouço probabilístico. A principal motivação é que a teoria das probabilidades fornece uma base sólida para representar e manipular a incerteza que é inerente ao processo de casamento entre consultas e documentos.

A ideia original foi proposta por Rijsbergen (1979) e intitulada *Probability Ranking Principle*. A ideia é ranquear os documentos da coleção em ordem decrescente de probabilidade de relevância para a consulta. Se tivéssemos uma amostra de documentos sabidamente relevantes ou não relevantes, seria possível estimar a probabilidade de um termo aparecer em um documento relevante. Essa informação poderia então ser usada para ranquear os documentos. As probabilidades são estimadas com a maior precisão possível com base em estatísticas que conseguimos calcular. O primeiro modelo probabilístico foi proposto por Robertson; Spärck Jones (1976) – o *Binary Independence Model* (BIM). O BIM é um modelo binário, ou seja, ele considera apenas se o termo está ou não presente no documento. Outra simplificação é a utilização do princípio da independência, que supõe que os termos ocorrem nos documentos de maneira independente. Esse é um modelo simples que não incorpora algumas características, como a frequência dos termos nos documentos e a importância dos termos na coleção (vistas na Seção 2.3.2), que contribuem para a geração de resultados melhores para as consultas (i.e., uma melhor ordenação dos documentos).

O modelo probabilístico mais influente até os dias de hoje é o **Okapi BM25** que foi projetado entre os anos 1980 e 1990 por Spärck Jones et al. (2000). O Okapi era um sistema de busca para o catálogo da biblioteca da City University de Londres. O modelo BM25 (BM é a abreviatura de *best match*) foi resultado de uma série de experimentos com variações de fórmulas de modelos probabilísticos. Os pesquisadores usaram identificadores para as variações da fórmulas e a que obteve melhores resultados foi a BM25, que é dada conforme a Equação 2.3 a seguir:

$$BM25(q, d_j) = \sum_{i \in q} \log \frac{(r_i + 0,5)/(R - r_i + 0,5)}{(df_i - r_i + 0,5)/(N - df_i + r_i + 0,5)} \times \frac{(k_1 + 1)tf_i}{K + tf_i} \times \frac{(k_2 + 1)qf_i}{k_2 + qf_i} \quad (2.3)$$



onde R é o número de documentos sabidamente relevantes para a consulta q ; r_i é o número de documentos sabidamente relevantes que contêm o termo i ; N é o número de documentos da coleção; df_i é o número de documentos que contêm o termo i ; tf_i é o número de ocorrências do termo i no documento j ; qf_i é o número de ocorrências do termo i na consulta q ; K , k_1 e k_2 são parâmetros que precisam ser definidos empiricamente. A adição de 0,5 em cada termo da equação é para evitar a divisão por zero. Note que o somatório é efetuado para todos os termos da consulta.

O parâmetro k_1 controla a escala da frequência dos termos (tf) nos documentos. Usar $k_1 = 0$ faz com que o modelo se comporte como um modelo binário. Por outro lado, altos valores de k_1 correspondem ao uso do TF sem normalização. Na prática, é comum usarmos $k_1 = 1,2$. De forma similar, o parâmetro k_2 controla a escala da frequência dos termos na consulta (qf). Os valores recomendados para k_2 são entre 0 e 1000.

Por fim, o parâmetro K é calculado por:

$$K = k_1((1 - b) + b \times \frac{dl}{avdl})$$

onde dl é o tamanho do documento, $avdl$ é o tamanho médio dos documentos da coleção e b controla a normalização em função do tamanho do documento. Se $b = 0$, então a normalização não é realizada. Já $b = 1$ normaliza os pesos em função do tamanho dos documentos. O valor recomendado para b é 0,75.

É importante ressaltar que os escores gerados pelos modelos probabilísticos não são probabilidades, i.e., não estão no intervalo $[0,1]$. Isso não representa um problema pois a utilidade do escore é apenas ordenar os documentos, sem a necessidade de permitir uma interpretação probabilística.

Como na imensa maioria dos casos práticos nós não temos conhecimento de uma amostra de documentos relevantes, o BM25 também pode ser computado **sem a informação de relevância**. Em outras palavras, R e r_i são zero. Com isso, a Equação 2.3 pode ser simplificada para:

$$BM25(q, d_j) = \sum_{i \in q} \log \frac{N - df_i + 0,5}{df_i + 0,5} \times \frac{(k_1 + 1)tf_i}{K + tf_i} \times \frac{(k_2 + 1)qf_i}{k_2 + qf_i} \quad (2.4)$$

Exemplo 2.2: Para ilustrar o cálculo do BM25, vamos utilizar a mesma consulta da Seção 2.3.2 $q = \{\text{comitiva, médico}\}$ e faremos o cálculo do escore passo a passo para o d_1 . Não temos informação prévia de relevância, então vamos usar a Equação 2.4. Cada termo ocorre apenas uma vez na consulta, então $qf_1 = 1$ e $qf_2 = 1$. O escore final é formado pela soma dos escores das palavras da consulta que aparecem no documento d_1 . Para facilitar o entendimento, os cálculos referentes ao termo “comitiva” estão em laranja e os os cálculos referentes ao termo “médico” estão em azul. As estatísticas da coleção estão nas Tabelas 2.1, 2.2, 2.3. Os parâmetros utilizados foram $k_1 = 1,2$, $k_2 = 100$ e $b = 0,75$. O tamanho médio dos documentos da coleção ($avdl$) é de 276 e o tamanho de d_1 é 161. Com isso, $K = 1,2((1 - 0,75) + 0,75 \times 161/279) = 0,825$. Observe que o escore final para o documento é negativo, mas o valor absoluto não é importante pois estamos interessados apenas em ranquear os documentos. Calculando o escore para os outros quatro documentos, verificamos que o primeiro documento do *ranking* seria d_5 , seguido por d_1 , d_4 e d_3 . Note-se que o ordenamento fica quase igual ao gerado pelo modelo vetorial – apenas os documentos d_4 e d_3 invertem as posições.

$$\begin{aligned}
 BM25(q, d_1) &= \log \frac{5 - 2 + 0,5}{2 + 0,5} \times \frac{(1, 2 + 1)4}{0,825 + 4} \times \frac{(100 + 1)1}{100 + 1} + \\
 &\quad \log \frac{5 - 4 + 0,5}{4 + 0,5} \times \frac{(1, 2 + 1)18}{0,825 + 18} \times \frac{(100 + 1)1}{100 + 1} = \\
 &\quad \log \frac{3,5}{2,5} \times \frac{8,8}{4,825} \times \frac{101}{101} + \log \frac{1,5}{4,5} \times \frac{39,6}{18,825} \times \frac{101}{101} = \\
 &\quad \log 1,4 \times 1,824 \times 1 + \log 0,33 \times 2,1 \times 1 = \\
 &\quad 0,33 \times 1,824 + -1,10866 \times 2,1 = \\
 &\quad 0,6137 - 2,311 = -1,6973
 \end{aligned}$$

2.4 Avaliação da Qualidade de Sistemas de recuperação de informação

A RI é uma disciplina altamente empírica e a história da avaliação em sistemas de RI nasceu junto com a área. Os primeiros trabalhos sobre avaliação em RI foram coordenados por Cyril W. Cleverdon na escola de aeronáutica de Cranfield (Inglaterra) nos anos 1960. Por esta razão, o modelo de avaliação até hoje é conhecido como paradigma Cranfield. A avaliação consiste no cálculo de uma série de métricas que são calculadas com base em uma coleção de teste. Nesta seção abordaremos as principais métricas (Seção 2.4.1) e as coleções de teste para português (Seção 2.4.2). O leitor que deseja se aprofundar no processo de avaliação de RI pode referir-se a Sanderson et al. (2010).

2.4.1 Métricas

As métricas são baseadas na noção de relevância, i.e., uma avaliação que diz se um documento d_j é relevante a uma consulta q . Inicialmente, vamos tratar a relevância como **binária**, ou seja, o *ground truth* julga o documento como relevante (1) ou não relevante (0). Todas as métricas vistas aqui variam no intervalo $[0,1]$, sendo que 1 representa a recuperação ideal.

2.4.1.1 Métricas Baseadas em Conjuntos

A Figura 2.3 mostra um exemplo do que ocorre tipicamente quando fazemos uma consulta. A coleção de documentos está representada pelo oval cinza. O círculo amarelo representa os documentos que foram recuperados em resposta à consulta, enquanto que os documentos que de fato são relevantes para a consulta estão representados pelo círculo azul. Podemos ver que há documentos relevantes que foram recuperados (i.e., os documentos que estão na intersecção entre os conjuntos, que aparece representada na cor verde), mas há também documentos relevantes que deixaram de ser recuperados e documentos não relevantes que foram recuperados. Com base nesses conjuntos, podemos definir duas métricas básicas para avaliar a qualidade da recuperação: **precisão** e **revocação** (em inglês, *recall*).

- Precisão (P) mede a proporção dos documentos recuperados que de fato é relevante.

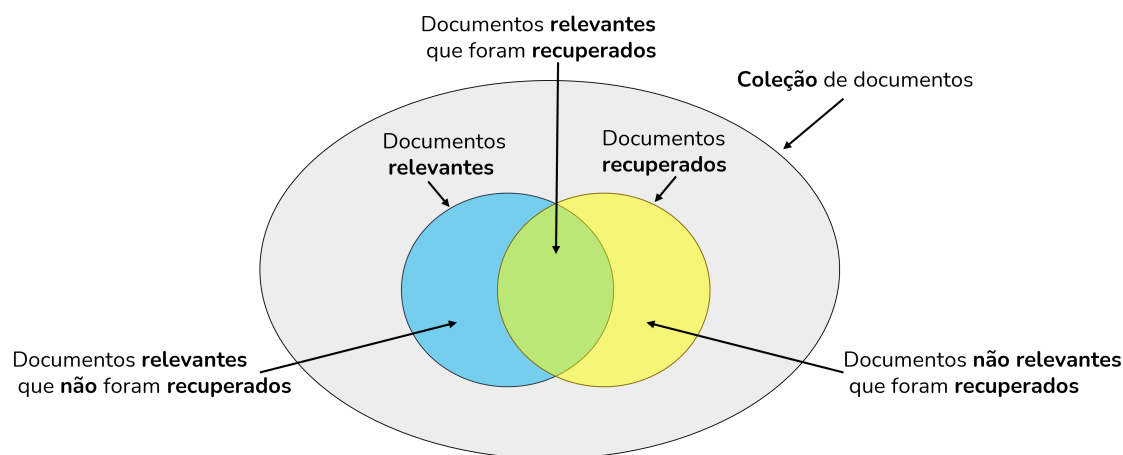
$$P = \frac{\# \text{ relevantes recuperados}}{\# \text{ recuperados}}$$



- Revocação (R) mede a proporção dos documentos relevantes que de fato foram recuperados.

$$R = \frac{\# \text{ relevantes recuperados}}{\# \text{ relevantes}}$$

Figura 2.3: Exemplo de recuperação de documentos em resposta a uma consulta.



Por exemplo, vamos imaginar uma situação em que um sistema recupera 100 documentos em resposta a uma consulta para a qual existam 30 documentos relevantes. Sabendo-se que entre os recuperados há 15 documentos relevantes, os valores das métricas de avaliação seriam $P = \frac{15}{100} = 15\%$ e $R = \frac{15}{30} = 50\%$. Imaginemos, então, uma segunda situação em que, na tentativa de aumentar a precisão, reduzimos o número de documentos recuperados para apenas 5 e todos eles sejam relevantes. Como resultado, nossa precisão aumentaria para $P = \frac{5}{5} = 100\%$. O problema é que, com isso, a nossa revocação cairia para $\frac{5}{30} = 16,67\%$. Indo na direção oposta, poderíamos tentar aumentar a revocação e recuperar mil documentos em vez de 100. Com isso, haveria uma chance bem maior de recuperarmos documentos relevantes, mas por outro lado, a precisão seria reduzida. Em resumo: altos níveis de revocação costumam ser acompanhados por baixos níveis de precisão e vice-versa. Por esta razão, é comum utilizarmos uma medida que agregue P e R , a conhecida **medida F**, definida a seguir.

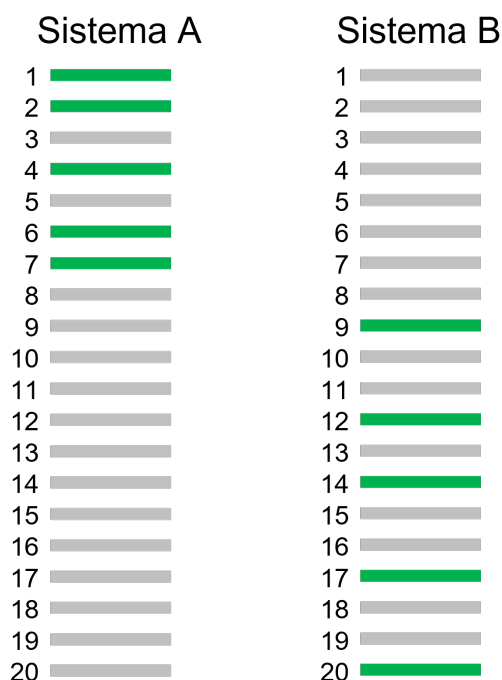
$$F = \frac{(\beta^2 + 1) \times P \times R}{(\beta^2 \times P) + R}$$

onde β é um parâmetro que permite definirmos a ênfase dada às métricas. Valores de β maiores do que 1 enfatizam a revocação enquanto que valores de β menores do que 1 enfatizam a precisão. Se quisermos atribuir a mesma ênfase às duas métricas, usamos $\beta = 1$. Nesse caso, a métrica é comumente chamada de $F1$ e sua fórmula é simplificada para $F1 = \frac{2 \times P \times R}{P + R}$. Considerando o nosso primeiro exemplo, em que $P = 0,15$ e $R = 0,50$, a $F1$ seria $F1 = \frac{2 \times 0,15 \times 0,50}{0,15 + 0,50} = 0,23$.

2.4.1.2 Métricas para *Rankings*

Os sistemas de RI retornam os documentos em um *ranking*. A Figura 2.4 mostra exemplos de *rankings* de documentos recuperados em resposta a uma mesma consulta por dois sistemas diferentes (Sistemas A e B). São 20 documentos recuperados e, dentre eles, há 5

Figura 2.4: Exemplos de *rankings* de documentos recuperados em resposta a uma consulta por dois sistemas diferentes. Os documentos relevantes à consulta estão representados em verde.



relevantes (representados em verde). Vamos supor que existam 7 documentos relevantes para essa consulta. As métricas vistas na Seção 2.4.1.1 resultariam em $P = \frac{5}{20} = 25\%$, $R = \frac{5}{7} = 71\%$ e $F1 = \frac{2 \times 0,25 \times 0,71}{0,25 + 0,71} = 0,37$ tanto para o Sistema A quanto para o Sistema B. Esse resultado não é o ideal pois podemos ver que o Sistema A recuperou os documentos relevantes mais perto do topo do *ranking* e por isso deveria receber um escore mais alto do que o Sistema B. Esse problema ocorre porque precisão, revocação e medida-F são baseadas em conjuntos – e conjuntos, por definição, não têm ordenação. Então concluímos que precisão, revocação e F1 **não são adequadas para avaliar *rankings***. Uma boa métrica para avaliar resultados ranqueados deve conseguir calcular quantos documentos relevantes foram recuperados e o **quão próximos estão do topo do ranking**. Com isso em mente, a métrica **precisão média** (do inglês, *average precision*, AP) foi proposta.

$$AP = \frac{\sum_{k=1}^n P(k) \times rel(k)}{\# \text{ relevantes}}$$

onde $rel(k)$ é uma função de relevância binária que resulta em 1 caso o documento na k -ésima posição do *ranking* seja relevante. Os documentos relevantes que são foram recuperados serão penalizados com $P = 0$. Em outras palavras, percorreremos o *ranking* calculando, para cada documento relevante recuperado, a precisão naquele ponto. Para o Sistema A da Figura 2.4, temos a seguinte $AP = \frac{1/1 + 2/2 + 3/4 + 4/5 + 5/7}{7} = 0,61$. E para o Sistema B temos $AP = \frac{1/9 + 2/12 + 3/14 + 4/17 + 5/20}{7} = 0,14$. Concluimos, então, que a AP consegue capturar o fato do Sistema A ter sido superior ao B.

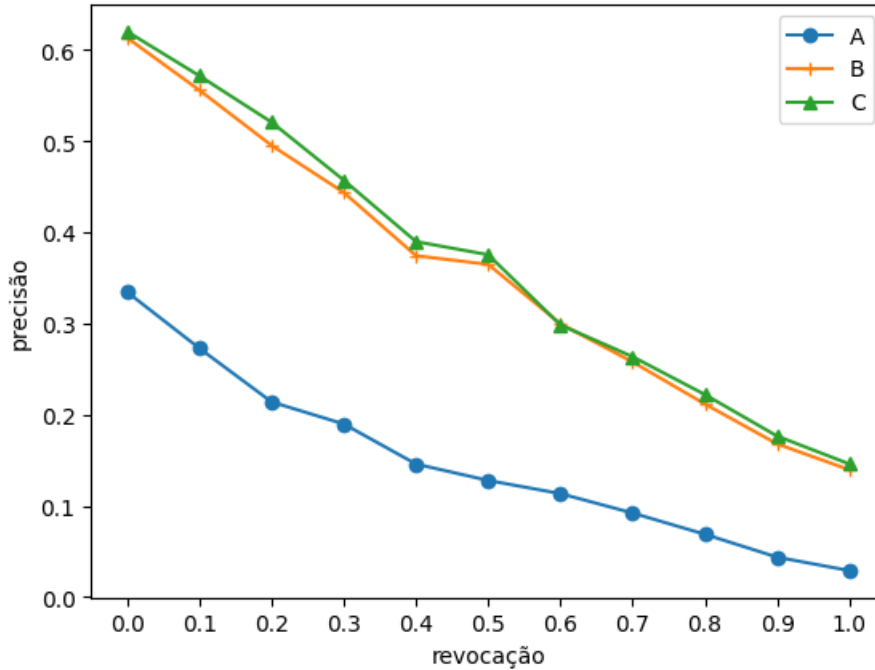
Quando temos um conjunto de consultas, calculamos a média das precisões médias (do inglês, *mean average precision*, MAP), que é simplesmente a média das APs para $|Q|$

consultas.

$$MAP = \frac{\sum_{q=1}^{|Q|} AP_q}{|Q|}$$

Muitas vezes, é útil fazermos uma comparação visual do desempenho de dois sistemas. Neste caso, podemos utilizar as **curvas de revocação e precisão**. Para desenhar as curvas, precisamos ter os valores de precisão calculados para 11 pontos de revocação padrão (de 0 a 1 com incrementos de 0,1). Para calcular a precisão nesses 11 pontos, é necessário utilizar uma regra de interpolação que diz que “a precisão interpolada para um nível de revocação j é o maior valor de precisão para qualquer nível de revocação maior ou igual a j ”. Aplicando-se a regra da interpolação, teremos curvas monotonicamente decrescentes, como as que vemos na Figura 2.5. Quanto maior a área sob a curva, melhor o resultado do sistema. Podemos ver claramente que a função A é pior do que as outras duas e que as funções B e C obtiveram resultados muito similares, com uma leve superioridade da função C . Para sabermos se as diferenças entre os sistemas são significativas, é comum fazermos testes estatísticos. O teste-T pareado é bastante usado e podemos fazê-lo comparando as precisões médias (AP) para um mesmo conjunto de consultas executadas em sistemas diferentes. Para os resultados das consultas que deram origem às curvas da Figura 2.5, quando comparamos os resultados da função C com a função A , o teste-T resulta em um p -valor = $4.6e-08$. Utilizando um nível de significância $\alpha = 0.01$, a interpretação é que a função C é significativamente melhor do que a função A pois o p -valor é $\ll 0.01$. Já quando comparamos a função C com a função B , o p -valor é 0.12. Então concluímos que não há diferença significativa entre B e C pois o p -valor é > 0.01 .

Figura 2.5: Exemplos de curvas de revocação e precisão para a comparação de três funções de *ranking*.



Além da MAP, é comum avaliar-se *rankings* com a métrica $P@k$, onde k é uma posição

específica do *ranking* (ex: 1, 5, 10, 100 etc.). Basta calcular o número de documentos relevantes encontrados até a k -ésima posição do *ranking*. Por exemplo, a $P@10$ conta quantos documentos relevantes foram encontrados até a décima posição do *ranking*. Essas métricas são especialmente úteis quando não sabemos o número de documentos relevantes para cada consulta, como é o caso das buscas na Web.

Outra métrica por vezes usada é o *mean reciprocal rank* (MRR) que calcula o inverso da posição do primeiro documento relevante retornado. Por exemplo, caso o primeiro documento relevante esteja na segunda posição do *ranking*, $MRR = 1/2$. Para agregar os resultados para um conjunto de consultas, fazemos a média:

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i}$$

Até agora, abordamos apenas métricas que utilizam a relevância como binária. Contudo, em alguns casos, é importante diferenciar um documento muito relevante de um documento marginalmente relevante. Nesses casos, podemos calcular o **ganho acumulado descontado normalizado** (do inglês, *normalized discounted cumulative gain*, NDCG) (Järvelin; Kekäläinen, 2002). Para entender o NDCG, primeiro temos que entender o DCG e o IDCG.

$$DCG = \sum_{i=1}^n \frac{2^{rel_i} - 1}{\log_2(i + 1)}$$

onde n é o número de documentos recuperados, i é a posição do documento no *ranking* e rel_i é o nível de relevância do documento (i.e., podemos usar quatro níveis onde o score 3 indica um documento altamente relevante, 2 indica um documento moderadamente relevante e 1 indica um documento marginalmente relevante e 0 é um documento não relevante). O IDCG é o DGC ideal, ou seja, a pontuação de DCG obtida se a ordenação fosse perfeita. Em outras palavras, primeiramente teríamos todos os documentos de score 3, seguidos pelos de score 2, depois os de score 1 e por fim os de score zero. O NDCG é a fração $\frac{DCG}{IDCG}$. A normalização é importante para fazer com que a métrica fique no intervalo $[0,1]$.

Além dessas medidas, muitas outras continuam sendo propostas e apresentadas nos fóruns de RI. Contudo, ainda não são amplamente adotadas. Há também muitos estudos sobre a validade e estabilidade das métricas, como por exemplo o trabalho de Buckley; Voorhees (2017).

2.4.2 Coleções de Teste

O cálculo das métricas apresentadas na Seção 2.4.1 só é possível caso exista uma coleção de teste com três componentes: (i) um conjunto de documentos, (ii) um conjunto de tópicos de consulta e (iii) os julgamentos de relevância. O primeiro componente é o mais fácil de ser obtido, pois é trivial fazer uma coleta a partir da Web e obter um grande número de documentos. O segundo componente, i.e., as consultas, necessita da intervenção humana na sua elaboração pois é preciso imaginar o tipo de consulta que um usuário que estivesse pesquisando a coleção de documentos poderia fazer. É comum a avaliação de consultas ad hoc utilizar por volta de 50 tópicos de consulta. O maior custo na elaboração de uma coleção de testes está no terceiro componente, i.e., os julgamentos de relevância. Os julgamentos de relevância dizem quais são os documentos relevantes para cada tópico de consulta e necessitam de intervenção humana na sua elaboração.



A primeira coleção de testes elaborada, a Cranfield Collection, tem 1398 resumos de artigos sobre aerodinâmica e 225 tópicos de consulta. A elaboração dos julgamentos de relevância foi exaustiva, ou seja, a relevância de todos os documentos foi avaliada em relação a cada consulta. Além disso, cada par {documento, consulta} foi julgado por dois avaliadores – ou seja $1398 \text{ documentos} \times 225 \text{ consultas} \times 2 \text{ avaliadores} = 629.100$, i.e., mais de meio milhão de avaliações para uma coleção muito pequena. Aplicar essa mesma abordagem em uma coleção com 100 mil documentos e 50 consultas, demandaria 10 milhões de julgamentos de relevância, o que seria impraticável.

A solução amplamente adotada foi a criação de um método em que nem todos os documentos são avaliados em relação às consultas – apenas os documentos que têm alguma chance de serem relevantes. Essa foi a ideia do método *pooling* (Spärck Jones, 1975) que tem esse nome pois são criados repositórios (*pools*) com o subconjunto de documentos que serão avaliados. O processo envolve a criação de um repositório para cada consulta. Os repositórios são populados com documentos recuperados utilizando diferentes funções de *ranking*, variando-se parâmetros e sistemas. Os documentos do repositório são então ordenados de acordo com alguma heurística – por exemplo documentos que foram retornados mais vezes ficam no topo do *ranking*. Então, apenas os n primeiros são analisados pelos avaliadores humanos. Os documentos que não foram avaliados são considerados como não relevantes. É possível que essa estratégia deixe de identificar alguns documentos relevantes, mas há evidências de que ainda assim ela é robusta. Um estudo realizado por Zobel (1998) identificou que a ordenação entre funções de *ranking* (ou sistemas) permanecia a mesma ainda que o conjunto de julgamentos de relevância mudasse. Em outras palavras, a superioridade de um método em relação a outro é mantida mesmo com conjuntos diferentes de julgamentos de relevância.

Mesmo com a grande redução no esforço manual de avaliação na criação de coleções trazida pelo método *pooling*, este processo ainda é bastante trabalhoso. As coleções de teste foram criadas, na imensa maioria das vezes, dentro do escopo de campanhas de avaliação. Essas campanhas tiveram um grande impacto no desenvolvimento de técnicas para RI. Para a língua inglesa, as campanhas TREC (*Text REtrieval Conference*)¹² mantiveram tarefas de recuperação ad hoc entre os anos de 1992 e 1999. Posteriormente, uma iniciativa europeia deu origem às campanhas CLEF (*Cross Language Evaluation Forum*) que entre 2000 e 2009 organizaram tarefas de recuperação ad hoc para idiomas europeus. Semelhante a essas campanhas, há o NTCIR¹³ dedicado aos idiomas do leste da Ásia.

Foi no contexto das campanhas CLEF e por iniciativa da Linguateca¹⁴ que a primeira coleção de teste para a língua portuguesa foi elaborada – a coleção CHAVE¹⁵ (Santos; Rocha, 2004b). A coleção CHAVE é composta por cerca de 209 mil notícias dos jornais Folha de São Paulo e Público (de Portugal), publicadas nos anos de 1994 e 1995. Há 100 tópicos de consulta com os seus respectivos julgamentos de relevância. Por muitos anos, esta foi a única coleção de teste disponível para português e sem dúvida ainda representa o recurso mais valioso para a recuperação ad hoc neste idioma.

A segunda coleção de teste para a recuperação ad hoc na língua portuguesa é a coleção REGIS¹⁶ (*Retrieval Evaluation for Geoscientific Information Systems*). Ela é composta por teses, dissertações e artigos científicos no domínio geocientífico. São mais de 20 mil

¹²<https://trec.nist.gov/>

¹³<https://research.nii.ac.jp/ntcir/>

¹⁴<https://www.linguateca.pt/>

¹⁵<https://www.linguateca.pt/CLEF/>

¹⁶<https://github.com/Petroles/regist-collection>



documentos e 34 tópicos de consulta. Uma característica distintiva é que os documentos são bem mais longos do que os encontrados em coleções de notícias de jornal e tratam de um domínio mais restrito. Além disso, os julgamentos de relevância foram feitos em quatro níveis (muito relevante, moderadamente relevante, marginalmente relevante e não relevante).

Recentemente, uma versão traduzida da MS MARCO (*Microsoft Machine Reading Comprehension*)¹⁷ foi disponibilizada para português e mais 12 idiomas (Bonifacio et al., 2021). A MS MARCO contém um conjunto de *datasets* que são amplamente utilizados no treinamento de algoritmos de aprendizado profundo. Para RI, há uma coleção com 8,8 milhões de documentos e 6980 consultas. As consultas foram obtidas a partir de buscas reais submetidas ao Bing e os documentos são passagens curtas extraídas de páginas web.

O Quadro 2.2 mostra exemplos de tópicos de consulta das três coleções de teste existentes para a língua portuguesa. O formato adotado pelas coleções CHAVE e REGIS é o formato tradicional utilizado pelas campanhas TREC e CLEF. Cada tópico de consulta é composto por um identificador, um **título** que descreve sucintamente o tópico, uma **descrição** mais detalhada e uma **narrativa** que visa auxiliar o avaliador humano a distinguir os documentos relevantes dos não relevantes. Já os tópicos da MS MARCO são diferentes, há apenas um identificador e o texto da consulta submetida ao motor de busca.

Quadro 2.2: Exemplos de tópicos de consulta das coleções de teste de RI em português

CHAVE

```
<top>
<num> C267 </num>
<PT-title> Melhor Filme Estrangeiro </PT-title>
<PT-desc> Quais foram os filmes candidatos ao Oscar de Melhor Filme Estrangeiro? </PT-desc>
<PT-narr> Documentos relevantes devem indicar o título e nacionalidade dos filmes indicados para o Óscar na categoria de Melhor Filme em Língua Estrangeira. </PT-narr>
</top>
```

REGIS

```
<top>
<num>Q5</num>
<title>Permeabilidade em Marlim</title>
<desc>Informações sobre o campo de Marlim, mas não dos campos de “Marlim Sul” ou de “Marlim Leste”.</desc>
<narr>Interessam documentos que contenham informação sobre a permeabilidade das rochas do campo de Marlim. Principalmente, interessam informações quantitativas (como dados de permeabilidade expressos em Darcies ou milidarcies).</narr>
</top>
```

MS MARCO

350,732	como limpar copos de melamina manchados
1,067,511	por que meu ligamento colateral lateral dói
906,071	o que adicionar à sua água para ser mais saudável

¹⁷<https://microsoft.github.io/msmarco/>



2.5 Modificação Automática de Consultas

Conforme visto na Introdução, o problema central da RI é casar a consulta com os documentos que a satisfazem. Contudo, o casamento puramente léxico aplicado pelos Modelos Clássicos (vistos na Seção 2.3) implica que uma consulta somente recupera documentos que possuam alguma de suas palavras-chave. A consulta não é capaz de recuperar um documento que contenha apenas sinônimos das suas palavras-chave e isso tem um impacto negativo na revocação. Por exemplo, se um usuário consulta “tributos pagos por donos de carros” um documento que não contenha as palavras da consulta mas mencione “imposto sobre a propriedade de veículos automotores” não seria recuperado. Desde os anos 1960, os pesquisadores vêm propondo formas de solucionar esses problemas. Essa sub-área de pesquisa é conhecida como **modificação automática de consultas**¹⁸ (*automatic query modification*) AQM e conta com dezenas de milhares de artigos publicados. Esta seção apresenta um breve resumo e fornece referências para que o leitor possa encontrar mais informações. A AQM é um ponto forte de interseção entre RI e PLN pois os diversos métodos adotados em PLN para solucionar o problema da sinonímia encontram aqui uma ótima área de aplicação.

Um bom ponto de partida para quem quer se inteirar-se sobre esse tema é o *survey* de Carpineto; Romano (2012). vamos adotar a taxonomia introduzida nesse trabalho para classificar as abordagens para AQM.

- **Análise Linguística.** Uma das técnicas mais simples que pode ser enquadrada nessa categoria é o uso de *stemming* (ver Seção 2.2.1.3). O *stemming* reduz as formas variantes de uma palavra ao mesmo radical e assim conseguiria, por exemplo, que uma consulta pelo termo “imposto” recuperasse documentos com a forma plural “impostos”. Contudo, o *stemming* não é capaz de recuperar um documento que contenha apenas o sinônimo “tributo”. Para isso, poderíamos nos valer de dicionários de sinônimos, tesouros ou ontologias. A modificação da consulta consiste em adicionar termos sinônimos à consulta original. A adição de sinônimos irá recuperar mais documentos. O problema ocorre quando os sinônimos adicionados mudam o significado da consulta, por exemplo “homenagem” é sinônimo de “tributo”, mas sua adição à consulta iria trazer documentos irrelevantes, baixando a precisão. Alguns domínios como a medicina dispõe de tesouros que podem ser diretamente usados. Por outro lado, essa é uma exceção – a maioria dos domínios não dispõem desses recursos e a sua construção manual é cara. Uma solução seria descobrir automaticamente essas relações a partir de *corpora*.
- **Técnicas baseadas em *corpora*.** Esse grupo de técnicas analisa o conteúdo da coleção de documentos para extrair padrões de coocorrência de termos. A ideia é que termos que coocorrem (i.e., que frequentemente aparecem juntos nos mesmos documentos) estão correlacionados e a sua adição à consulta pode fornecer contexto adicional. A informação de coocorrência pode ser obtida a partir de métodos diferentes como a Indexação Semântica Latente (Deerwester et al., 1990) ou *word embeddings* (Mikolov et al., 2013b) discutidas no Capítulo **Representação vetorial e semântica distribucional**. Há experimentos que obtiveram resultados positivos (Kuzi et al., 2016), contudo se as *embeddings* forem obtidas a partir de textos genéricos e a coleção for de um domínio específico, os termos adicionados podem piorar a

¹⁸O termo **expansão automática de consultas** é bastante empregado. Contudo, a consulta nem sempre é expandida e, por isso, preferimos adotar o termo **modificação automática de consultas**.



qualidade do *ranking*.

- **Técnicas baseadas no resultado da consulta** analisam os documentos recuperados pela consulta inicial para então modificá-la. A primeira técnica nessa categoria é a **realimentação de relevância** (em inglês, *relevance feedback*) (Rocchio-Jr, 1971). Ela consiste em solicitar que o usuário marque alguns documentos que ele considera relevantes e não relevantes. A reformulação da consulta consiste em modificar seu vetor para que ele se aproxime mais dos vetores dos documentos marcados como relevantes e se afaste dos vetores dos documentos não relevantes. Essa técnica costuma obter bons resultados, mas o ponto negativo é que os usuários não gostam de investir seu tempo fornecendo feedback. A alternativa então é supor que os primeiros k documentos recuperados pela consulta original sejam relevantes e modificar o vetor da consulta com base neles. Este processo é conhecido como **pseudo realimentação de relevância** e pode ter o efeito negativo de piorar ainda mais os resultados caso o primeiro conjunto de documentos recuperados não tenha sido bom. Outras propostas que se enquadram nessa categoria incluem o uso de *word embeddings* geradas a partir do resultado da consulta (Diaz et al., 2016).
- **Análise de logs de consulta** de motores de busca pode fornecer termos que os usuários frequentemente adicionam às suas consultas. Esses termos podem então ser usados para expandir a consulta original. Essa abordagem mostrou resultados positivos em (Cui et al., 2002).

Mais recentemente, uma nova gama de trabalhos têm focado no caminho inverso, i.e., expandir os documentos. Nessa linha de investigação, destaca-se o trabalho de Nogueira et al. (2019) em que os documentos são enriquecidos com consultas que poderiam ser feitas com o intuito de recuperar o documento.

2.6 Ferramentas e Bibliotecas

Há várias ferramentas (ou sistemas) de RI disponíveis tanto para fins comerciais como para fins de pesquisa. Esta seção aborda algumas das mais utilizadas.

2.6.1 Sistemas Comerciais

Os sistemas comerciais mais populares são da Apache e baseiam-se no **Lucene**¹⁹, uma biblioteca Java de código aberto. A biblioteca não é um sistema de RI completo. Com base nela, foram desenvolvidos o **Solr**²⁰ e o **Elasticsearch**²¹ que fornecem interfaces de consulta mais amigável além de indexação distribuída para fins de escalabilidade. Tanto o Solr como o Elasticsearch são ferramentas de código aberto, possuem vasta documentação e são amplamente adotados na indústria. O Elasticsearch fornece melhor suporte a sistemas distribuídos e pode ser integrado com ferramentas de análise e visualização. Por outro lado, o Solr possui mais flexibilidade para customização.

2.6.2 Ferramentas para Pesquisa

Há várias ferramentas de código aberto construídas por pesquisadores da área de RI que estão disponíveis e são usadas em trabalhos acadêmicos. Dentre elas, as mais utilizadas

¹⁹<https://lucene.apache.org/>

²⁰<https://solr.apache.org/>

²¹<https://www.elastic.co/>



atualmente são:

- O **Anserini**²² e sua interface em Python (Pyserini²³) foram elaborados por pesquisadores das universidades de Delaware e Waterloo (Yang et al., 2017) e tem o objetivo de facilitar o processo experimental em RI com ênfase na reprodutibilidade.
- O **Terrier**²⁴ e sua interface em Python, o Pyterrier²⁵ (Macdonald; Tonellotto, 2020), foram criados por pesquisadores da universidade de Glasgow que tem longa tradição na pesquisa em RI. O Terrier também dispõe de uma interface gráfica que permite indexar e consultar uma coleção sem a necessidade de escrever código.
- O projeto **Lemur**²⁶ reúne uma série de componentes. O Indri é o motor de busca em si. O projeto é uma colaboração entre as Universidades de Massachusetts e Carnegie Mellon. O sistema Galago também é parte desse projeto e foi disponibilizado em conjunto com o livro de Croft et al. (2010).

2.7 Conclusão

Este capítulo forneceu uma visão geral sobre a área de RI. O processo típico de RI foi introduzido, bem como os modelos clássicos. Damos ênfase à metodologia de avaliação dos resultados das consultas, dada a importância que esse tema tem na área. Também foram apontadas coleções de teste em português e ferramentas que podem ser usadas tanto no meio acadêmico como na indústria.

Nessa primeira versão, o capítulo não abordou um tópico relevante e atual: a RI utilizando vetores densos. Em especial, o emprego da arquitetura Transformers (Vaswani et al., 2017b) tem sido bastante difundido em RI (assim como em diversas tarefas de PLN). Por ora, o leitor interessado pode referir-se o artigo de (Lin et al., 2020b) que apresenta um levantamento abrangente.

Agradecimentos

Agradeço a Helena Caseli e a Graça Nunes pelo convite a contribuir com este livro. Também sou grata a Adriana Pagano, Felipe Paula, Lucas Pessutto, Luciana Bencke e João Comba pela revisão e comentários.

²²<https://github.com/castorini/anserini>

²³<http://pyserini.io/>

²⁴<http://terrier.org/>

²⁵<https://github.com/terrier-org/pyterrier>

²⁶<https://lemurproject.org/>



Capítulo 3

Extração de Informação

Daniela Barreiro Claro

Joaquim Santos

Marlo Souza

Renata Vieira

Vládia Pinheiro

Publicado em: 26/09/2023

Atualizado em: 16/04/2026

 <https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/>

3.1 Introdução

A Extração de Informação (EI) é desenvolvida com o objetivo de se obter informação estruturada de dados não-estruturados (Jurafsky; Martin, 2023; Konstantinova, 2014).

Os primeiros trabalhos a debruçarem-se sobre o problema remontam à década de 1970, com a aplicação de gramáticas formais e *parsers* sintáticos para a estruturação de informação em domínios como prontuários médicos (Sager, 1978; Sager et al., 1987) e textos jornalísticos (DeJong, 1979). A comunidade científica demonstrou grande interesse pela área nas décadas posteriores devido à sua utilidade prática, seu foco no processamento de dados reais, suas tarefas bem-definidas e a facilidade de mensurar a qualidade dos resultados em comparação com o desempenho humano na mesma tarefa (Cowie; Lehnert, 1996).

Para autores como Eisenstein (2019) e Jurafsky; Martin (2023), a EI é normalmente dividida em diversas tarefas de interesse, com foco no tipo de informação a ser extraída do texto. Entre as mais comumente citadas na literatura estão o Reconhecimento de Entidades Nomeadas (REN), a Extração de Relações (ER) e a Extração de Eventos (EE).

O Reconhecimento de Entidades Nomeadas (REN) consiste em identificar e classificar entidades mencionadas em textos através de designadores rígidos como nomes próprios, expressões temporais e espécies biológicas (Nadeau, 2007). Esse é considerado por alguns como um primeiro passo na análise semântica de um texto (Santos; Cardoso, 2007a), pois permite identificar as entidades às quais se faz referência nele.

A Extração de Relações (ER), também chamada de extração de informação tradicional ou somente extração de informação, por sua vez, diz respeito à identificação de relacionamentos semânticos entre duas ou mais entidades, ou seja, identificar “quem fez o que para quem e quando”. Ananiadou; Mcnaught (2005) a definem como o processo de extrair fatos (em nossa terminologia, relacionamentos) a partir de uma fonte textual e representá-los a partir de um gabarito (em inglês, *template*). As *relações* são elementos essenciais para o entendimento da informação relatada no texto e sua identificação é passo essencial para a estruturação da mesma. Assim, identificar relações entre entidades é tarefa essencial



para construção de bases de conhecimento e de grande utilidade na construção de soluções para a resposta automática a perguntas (em inglês, *query answering*), sumarização, recuperação de informação e mais (Nasar et al., 2021).

A extração de eventos consiste na tarefa de identificação de uma menção a um evento em uma sentença e, se existirem, extração de outras informações sobre o evento. Um evento pode, por sua vez, ser entendido como uma ocorrência específica envolvendo participantes (Consortium, 2005), i.e., algo que acontece e que pode ser descrito como uma mudança de estado da qual participam entidades como agentes. Devido a intrínseca natureza temporal dos eventos, tal problema possui uma natureza mais complexa e costuma possuir tratamento específico.

Assim, nesse capítulo, iniciaremos com um pouco de história da Extração de Informação (EI) e sua evolução para Extração de Informação Aberta, e destacaremos as tarefas de Reconhecimento de Entidades Nomeadas (REN) e Extração de Relação (ER).

3.2 Um pouco de história

Os primeiros trabalhos que abordaram o problema de EI dos quais temos conhecimento surgiram no final da década de 1970. Esses primeiros trabalhos da década de 1970 e 1980 tinham como modelo geral a aplicação de regras para a identificação de informações especificadas em um gabarito. Tais sistemas empregavam analisadores sintáticos (*parsers*) e regras definidas especificamente para o domínio e gênero textual estudado.

Entre esses primeiros trabalhos, estão aqueles de Sager (1978), Sager et al. (1987), de DeJong (1979) e de Cowie (1983). Sager et al. exploraram como identificar informações do estado de saúde de pacientes através dos textos de prontuários médicos. DeJong (1979), por sua vez, descrevem o sistema FRUMP que, a partir de um *parser* e regras de análise conceitual baseadas em uma arquitetura cognitiva proposta pelos autores e no conceito de dependência conceitual de Schank et al. (1973), processavam textos de notícias e realizavam tarefas como sumarização e identificação de papéis semânticos associados aos constituintes da sentença. Cowie (1983), por fim, descreve um sistema que emprega regras simples de segmentação e análise sintática rasa para identificar propriedades de plantas a partir de textos descritivos no campo da botânica. Diferente dos métodos anteriores, o trabalho dos autores se baseia em grande parte no estudo de padrões de descrição das informações a serem identificadas, em detrimento do emprego de *parsers* robustos da língua.

A década de 1990 traz um grande interesse na área de EI com a implementação das conferências MUC (do inglês, *Message Understanding Conference*, ou Conferência de Compreensão de Mensagem), promovidas pela Agência de Projetos de Pesquisa Avançada de Defesa (DARPA, do inglês *Defense Advanced Research Projects Agency*). As conferências MUC, realizadas e financiadas pelo exército americano, representaram um esforço em avançar a tecnologia de EI e consistiam de tarefas de avaliação conjunta de métodos desenvolvidos por pesquisadores para problemas propostos pelos organizadores. As sete conferências realizadas de 1987 a 1997, foram cruciais para definir aspectos centrais da área, como estruturar a tarefa de ER, definindo suas métricas de avaliação, e propor a tarefa de REN (Grishman; Sundheim, 1996).

A partir da MUC-3, em 1991, a conferência passa a ter foco no processamento de textos jornalísticos em detrimento dos relatórios militares utilizados anteriormente (DARPA, 1991). Com a disponibilidade de dados e o incentivo no desenvolvimento de soluções para



a tarefa, vemos na década de 1990 o surgimento das primeiras aplicações comerciais de EI, como o JASPER (Andersen et al., 1992), construído para a agência de notícias Reuters.

A MUC-6, ocorrida em 1995, introduz a tarefa de REN com o intuito de ser uma tarefa de uso prático, independente de domínio e que poderia ser realizada automaticamente em um futuro próximo (Grishman; Sundheim, 1996). Enquanto os trabalhos em REN se avolumaram a partir de sua proposição na MUC-6, trabalhos anteriores como Rau (1991) e Wolinski et al. (1995) já se debruçavam sobre o problema de identificação e classificação de nomes próprios. Desde então, o interesse na tarefa cresceu significativamente e outras conferências de avaliação conjunta têm sido dedicadas a essa tarefa, como a *Automatic Content Extraction* (ACE) e a conferência Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas (HAREM), dedicada exclusivamente à língua portuguesa, com sua primeira edição em 2005 (Santos; Cardoso, 2007a).

Por outro lado, houve um crescimento de abordagens baseadas em dados nesta década, a partir da análise de *corpora*. Tais esforços são impulsionados pelos resultados positivos na área, como o trabalho de Hearst (1992). Logo, métodos baseados em dados passaram também a explorar o emprego de análise estatística e aprendizado de máquina na construção de padrões para a extração de relações (Riloff et al., 1993, 1999; Roark; Charniak, 2000; Soderland et al., 1995).

Não foi somente na extração de padrões que métodos de aprendizado de máquina, em particular aprendizado supervisionado, foram aplicados. A década de 2000 viu a proliferação de métodos supervisionados aplicados à ER (Culotta et al., 2006; Kambhatla, 2004; Zelenko et al., 2003; Zhao; Grishman, 2005) e ao REN (Asahara; Matsumoto, 2003; McCallum; Li, 2003; Sekine, 1998).

Devido à dificuldade de construção de dados para treinamento e padrões para extração, além da pouca adaptabilidade dos sistemas construídos para outros escopos e domínios, nos anos 2000, sistemas baseados em métodos de aprendizado semi-supervisionado, como o DIPRE (Brin, 1998) e Snowball (Agichtein; Gravano, 2000) começaram a aparecer, juntamente com os estudos sobre expansão automatizada de anotações (*bootstrapping*) (Riloff et al., 1999). Também para entidades nomeadas, estudos investigaram como utilizar recursos da Web (Etzioni et al., 2005; Nadeau, 2007) ou *corpora* (Cucchiarelli; Velardi, 2001) para aprender entidades com pouco ou nenhum esforço de anotação.

Buscando superar as dificuldades da limitação de escopo, i.e. das relações-alvo a serem extraídas e categorias de entidades a serem identificadas, ainda restritas à definição de padrões desde a criação dessas tarefas, Banko et al. (2007) propõe a tarefa de extração de informação aberta (EIA), também conhecida como **Open Information Extraction**, OpenIE ou OIE, a qual busca extrair todas as relações possíveis expressas em um texto, sem necessidade de pré-definição de relações e entidades.

Devido ao recente sucesso da aplicação de métodos baseados em redes neurais, em particular *deep learning* e grandes modelos de linguagem, às tarefas de Processamento de Linguagem Natural, uma tendência atual da área se delineou como o estudo de arquiteturas neurais para os problemas de EI e a geração de grandes conjuntos de dados por supervisão fraca. *Surveys* recentes, como (Cui et al., 2018; Konstantinova, 2014; Nasar et al., 2021), nos mostram a evolução da área em direção à aplicação de métodos neurais. Na vertente de geração de dados, vemos o emprego da Wikipédia e Freebase como fontes mais usadas para obter anotações de entidades e relações em textos (Nguyen et al., 2016; Smirnova; Cudré-Mauroux, 2018; Takamatsu et al., 2012).

Porém, toda a tarefa de EI necessita de uma concordância entre as definições de Entidade e Relação. Neste sentido, a próxima seção discute a conceituação de relação adotada neste

capítulo, assim como o conceito de entidade.

3.3 Conceituação formal: Relação e Entidade

A natureza das relações estudadas na área de Extração de Informação e os critérios para reconhecer sua ocorrência em um texto têm recebido pouca atenção na literatura. Este é um passo importante para estabelecer metodologias adequadas para avaliar os sistemas, bem como para criar conjuntos de dados que possam apoiar a criação de sistemas futuros.

Enquanto as noções de Relação e Entidade são de grande importância e já bem estudadas nas áreas de Computação, Linguística, Ciência da Informação e Filosofia da Linguagem, esses conceitos não são empregados de forma consistente entre as áreas, ou mesmo entre suas subáreas.

3.3.1 Entidade

Para Chen (1976), uma entidade é um objeto que pode ser concreto, tal como pessoa, livro, casa ou ainda abstrato, tal como um emprego, um sentimento, uma disciplina. As entidades podem estabelecer relações entre si. Duas ou mais entidades são vinculadas, ou seja conectadas por uma relação¹.

Tradicionalmente em reconhecimento de entidades nomeadas, as entidades consideradas são aquelas referenciadas por um nome próprio, acrescidas das referências temporais e valores que são expressões numéricas. Essas expressões, portanto, geralmente não constituem uma entrada em uma base lexical. Porém a tarefa se expandiu para domínios especializados, onde as entidades de interesse são mais conceituais. No domínio bio-médico por exemplo, podemos ter como exemplo de entidades de interesse, sintomas e tratamento que não são referenciadas por nomes próprios.

3.3.2 Relação

Os conceitos de *relação* e *relacionamento* são noções fundamentais que vêm sendo estudadas em áreas como Ciência da Computação, Linguística e Filosofia.

No campo de bancos de dados e modelagem conceitual, Chen (1976) define um relacionamento, no contexto da modelagem de Entidade-Relacionamento, como uma associação entre entidades. Guarino; Guizzardi (2015), por sua vez, estudando a natureza ontológica dos relacionamentos com base na semântica de veridadores (*truthmaker semantics*) (Fine, 2017), postulam relacionamentos como entidades que atuam como veridadores (*truthmakers*) de alguma proposição relacionando duas ou mais entidades, ou seja, *uma relação mantida entre essas entidades*. Um verificador é um elemento cuja existência torna verdadeira uma proposição particular. Por exemplo, considerando a sentença (1) “*a* é uma maçã”, a existência de um objeto denotado pelo nome *a* que por acaso é uma maçã é uma condição suficiente para a verdade da frase (1). Como tal, dizemos que esse objeto é o verificador de (1). Tal definição nos permite adotar critérios ontológicos para validar a existência de relacionamentos a partir da informação relatada em um texto e, por isso, adotaremos tal definição de relacionamento neste capítulo.

O conceito de relações é muito menos consistente na literatura. Ainda na área de modelagem conceitual, Guarino; Guizzardi (2015) definem as relações como proposições para as quais os relacionamentos são veridadores e, portanto, possuem conteúdo proposicional.

¹Em nossa terminologia, por um relacionamento.



Assim, podemos entender uma relação como um tipo para entidades como relacionamentos. Ou seja, relações são universais ontológicos que descrevem a natureza dos relacionamentos.

Xavier et al. (2015), no entanto, argumentam que a noção de relacionamento adotada na área de Extração de Informação é mais geral do que isso, não se limitando àquelas entre objetos e propriedades, mas também àquelas que descrevem ou implicam propriedades de classes gerais como descrito pela sentença (2) “Filósofos são autores de Livros”. Assim, para o contexto de EI consideramos relações como tipos de relacionamentos de primeira ou segunda ordem. Isso significa que uma relação é um tipo de relacionamento que existe entre objetos, suas propriedades e classes de objetos ou suas propriedades.

Enquanto os métodos tradicionais de Extração de Informação dependem de um conjunto pré-existente de relações semânticas bem definidas que são relevantes para um domínio específico, a noção de “relação” e “entidade” na literatura da área mais recente, tais como a Extração de Informação Aberta, requer mais aprofundamento por demandar um significado diferente, principalmente com diferentes visões de autores. Esta indeterminação terminológica pode trazer problemas para comparar os resultados dos métodos propostos ou para reutilizar os conjuntos de dados criados na área.

As seções seguintes exploram essas duas áreas: Extração de Informação e Extração de Informação Aberta.

3.4 Extração de Informação (EI)

A Extração de Informação é caracterizada por obter informação estruturada a partir de textos, sendo entidades ou fatos, i.e. relacionamentos entre entidades, de tipos previamente definidos, conforme exemplo no Quadro 3.1. Métodos com limitação de escopo possuem como uma de suas principais desvantagens a necessidade de intervenção humana para especificar novos fatos a serem extraídos. Esta limitação impede que sistemas de Extração de Informação, doravante denominados de EI tradicional extraiam fatos fora do escopo pré-definido.

Quadro 3.1: Exemplos de relações específicas na EI tradicional

Relação Específica	Exemplo de Sentença	Extração
location-of(algo/alguém, local)	Um aluno pode ser encontrado na escola	location-of(aluno, escola)
is-a(subclasse, superclasse)	Salvador é uma cidade	is-a(Salvador, cidade)
part-of(todo, parte)	Roda é um componente de um carro	part-of(roda, carro)

Fonte: (Souza; Claro, 2014)

3.4.1 Reconhecimento de Entidades Nomeadas

O Reconhecimento de Entidades Nomeadas (REN) consiste na tarefa de identificar e classificar expressões linguísticas, denominadas entidades nomeadas (EN), que referenciam entidades específicas num domínio de discurso, como nomes próprios, expressões temporais e espécies biológicas (Mota et al., 2007; Nadeau, 2007). De uma forma geral, o REN pode ser dividido em duas etapas: a identificação (ou delimitação) da expressão, na qual



as palavras que formam a EN são selecionadas; a classificação, em que é atribuída uma categoria semântica à EN.

A classificação das ENs determina os tipos de entidades a serem consideradas e são especificadas a partir do escopo definido previamente para a tarefa. Algumas das categorias mais comumente utilizadas incluem as entidades que referenciam Pessoas Singulares (antropônimos); Coletivas (empresas e organizações) e Lugares (topônimos) (Mota et al., 2007). Para exemplificar tomemos a sentença: “Renata Silva e Maria Costa palestraram na Universidade Federal da Bahia”. No exemplo temos três ENs: “Renata Silva”, “Maria Costa”, “Universidade Federal da Bahia”, sendo as duas primeiras correspondentes à categoria semântica Pessoa e a última, à categoria semântica Organização. Entretanto, existem outras categorias de ENs, como as menções a Obras (por exemplo, “Código Da Vinci”); Acontecimentos (por exemplo, “Festa de Santo Antônio”), Tempo (por exemplo, “meio-dia”); Coisa (por exemplo, “barco”), entre outras.

O REN é uma tarefa com grande importância para o Processamento de Linguagem Natural, pois consiste numa primeira tarefa de análise semântica de um texto, com potencial aplicações a diversas tarefas. Por exemplo, em sistemas de perguntas e respostas, as perguntas frequentemente se referem a informações sobre entidades. Também, métodos de identificação de estruturas mais complexas, como eventos ou relações, dependem do bom desempenho do REN como uma etapa de pré-processamento (Socher et al., 2012; Zelenko et al., 2003).

3.4.2 Extração de Relações

A tarefa de extração de relações (ou de relacionamentos) (ER) refere-se a identificar relacionamentos entre entidades de um determinado escopo mencionadas em um texto (Jurafsky; Martin, 2023). O escopo, no contexto da ER, refere-se a um conjunto de relações-alvo de um determinado domínio de conhecimento ou aplicação a ser investigado. Por exemplo, o Quadro 3.2 apresenta alguns exemplos de relações no domínio de geografia brasileira. Na descrição das relações, os elementos em negrito referem-se às entidades em um dado relacionamento descrito pelo termo em *itálico*.

Quadro 3.2: Exemplos de relações no domínio da geografia brasileira.

Relação	Descrição	Exemplo
Pertence(Cidade, Unidade Federativa)	Sobre uma cidade que está localizada em uma determinada Unidade Federativa , dizemos que a primeira <i>pertence</i> a esta última.	Pertence(Salvador, Bahia)
Tem_Prefeito(Cidade, Pessoa)	Uma pessoa que executa a função administrativa de gestão do executivo em nível municipal de uma dada cidade é denominada de seu(sua) <i>prefeito(a)</i> .	Tem_Prefeito(Salvador, Bruno Reis)
Fundação(Cidade, Data)	A data em que uma cidade foi fundada, é dita sua data de <i>fundação</i> .	Fundação(Salvador, 29 de março de 1549)

Nesse contexto, a delimitação de um escopo ou domínio de interesse, concentra-se na delimitação das relações a serem processadas, i.e. nos tipos de relacionamentos de interesse, assim como da natureza das entidades associadas por tais relações.



3.4.3 Extração Conjunta de Entidades e Relações

As tarefas de reconhecimento de entidades nomeadas e extração de relações são interdependentes, no sentido de que a definição do escopo a ser estudado delimita tanto as categorias e natureza das entidades a serem extraídas, como também as relações entre essas entidades. Também, note-se que, pelo fato de as relações serem comumente definidas entre entidades de tipo especificado, como o caso da relação *Tem_Prefeito* no Quadro 3.2 que ocorre entre entidades das classes **Cidade** e **Pessoa**, tanto as informações das entidades mencionadas no texto são úteis para a extração de relações, quanto a informação das relações identificadas pode ser útil ao processo de identificação de entidades.

De fato, na literatura recente, existem vários trabalhos que consideram a tarefa de extração conjunta de entidades e relações (ERE, do inglês *Entity and Relation Joint Extraction*), composta das tarefas de REN e ER (Agichtein; Gravano, 2000; Shaowei et al., 2022; Yuan et al., 2021b). Enquanto normalmente abordagens estruturam suas soluções de forma sequencial, usualmente realizando REN inicialmente e, posteriormente, realizando ER, como nos trabalhos de (Hasegawa et al., 2004) e de (Socher et al., 2012), a literatura recente aponta para as vantagens da identificação conjunta ao permitir um melhor aprendizado de restrições para identificação de entidades e relações, c.f. o recente *survey* realizado por (Shaowei et al., 2022) sobre métodos para tal tarefa.

3.4.4 Métodos empregados para EI na literatura

Várias abordagens foram adotadas para o problema de EI durante seu desenvolvimento histórico. Enquanto abordagens iniciais privilegiavam métodos ricos em conhecimento, como regras e recursos linguísticos e de conhecimento de mundo, a literatura recente na área privilegia métodos baseados em dados, como o aprendizado de máquina, com o recente emprego de arquiteturas neurais aos problemas.

A seguir faremos uma breve apresentação das abordagens descritas na literatura para os problemas de EI.

3.4.4.1 REN

As abordagens iniciais para REN baseavam-se, majoritariamente, no emprego de regras léxico-sintáticas e consulta a almanaques (*gazetteers*). Tais abordagens dependem da construção de listas de nomes próprios como antropônimos, topônimos etc., e outras palavras, como “Ltda.”, “Jr.” etc., que auxiliam no processo de identificação e classificação de ENs complexas ou desconhecidas. Essa é, por exemplo, a abordagem empregada por Wolinski et al. (1995) que combina almanaques e regras para a identificação e classificação de ENs. Posteriormente, almanaques foram também empregados em conjunção com métodos baseados em dados, como o trabalho de Florian et al. (2003) que os emprega aliados aos classificadores, enquanto Liu et al. (2019a) os utilizam durante o treinamento de uma rede neural, como um sinal de treinamento (parte da função de perda, ou *loss* em inglês).

Muitos trabalhos debruçaram-se também sobre o problema de construção automática ou semi-automática de almanaques, dos quais os trabalhos de Nadeau (2007), de Riloff et al. (1999) e de Etzioni et al. (2005) são alguns dos mais importantes.

Enquanto as abordagens iniciais para o problema baseavam-se em regras, com a disponibilidade de dados anotados para a tarefa, tais métodos foram rapidamente suplantados por métodos baseados em dados, tais como: os métodos baseados em classificação (Asahara;

Matsumoto, 2003; Sekine, 1998) e classificação sequencial (Bikel et al., 1999; McCallum; Li, 2003).

A redução de REN à tarefa de classificação sequencial merece destaque pelos bons resultados obtidos. Tal redução se dá através de um esquema de codificação do problema que nos permite representar fragmentos textuais e sua classificação como um problema de rotulação ou etiquetagem.

Partindo-se do pressuposto de que os fragmentos textuais descrevendo entidades nomeadas são contíguos, podemos codificar a tarefa de delimitação de entidades como classificação sequencial empregando rótulos que descrevem os limites de uma EN, e.g. o esquema BIO com os rótulos **B** (do inglês, *begin*) para designar a palavra inicial de uma EN, **I** (do inglês, *inside*) para designar palavras que fazem parte da EN mas não a iniciam e **O** (do inglês, *outside*) para designar palavras que não pertencem a uma entidade. Da mesma forma, podemos estender nosso esquema de codificação para incluir as classes de interesse. Assim, seguindo o esquema BIO, teremos os rótulos **B-PER** e **I-PER** para descrever entidades da classe **Pessoa**.

A redução do problema de REN à classificação sequencial está ilustrada no Exemplo 3.1.

Exemplo 3.1:

Renata/B-PER Silva/I-PER e/O Maria/B-PER Costa/I-PER palestraram/O
na/O Universidade/B-ORG Federal/I-ORG da/I-ORG Bahia/I-ORG.

Recentemente, destacam-se na literatura abordagens baseadas em redes neurais profundas, com uma grande concentração nos últimos anos em modelos gerativos de linguagem, devido aos resultados positivos obtidos por tais arquiteturas em diversas tarefas.

Na literatura são de grande destaque os modelos recentes BART (Lewis et al., 2020), RoBERTa (Liu et al., 2019c), T5 (Raffel et al., 2020), BERT (Devlin et al., 2019) e GPT-3 (Brown et al., 2020c), conforme descritos no Capítulo [Modelos de linguagem](#).

Similarmente, na língua portuguesa, nas duas edições do HAREM (Mota; Santos, 2008; Santos; Cardoso, 2007b), o primeiro esforço sistemático de desenvolvimento de soluções para a tarefa na língua, a maioria dos sistemas participantes baseava-se em métodos ricos em conhecimento, como regras e almanaques. De fato, nas duas avaliações, somente os sistemas MALINCHE (Solorio, 2007), NEURA (Ferrández et al., 2007) e R3M (Mota, 2008) não se baseavam em regras. Métodos baseados em classificação sequencial se seguiram para a língua portuguesa, como o RELP-CRF (Amaral; Vieira, 2014) baseado em um classificador sequencial. Mais recentemente, abordagens baseadas em redes neurais e modelos de linguagem foram desenvolvidas tornando-se o estado da arte da tarefa na língua. A Tabela 3.1 apresenta o atual estado da arte em português, com base no *corpus* HAREM. A métrica de avaliação apresentada, medida F1, será discutida na Seção 3.6.

Tabela 3.1: Trabalhos estado da arte no REN em português

Modelo	Medida F1
BERT-CRF (Souza et al., 2020a)	83,70%
BiLSTM-CRF+FlairBBP (Santos et al., 2019b)	82,26%
LSTM-CRF (Castro et al., 2018)	76,27%
CharWNN (Santos; Guimarães, 2015)	65,41%

Souza et al. (2020a) desenvolveram um modelo BERT para o Português com 2,68 bilhões de *tokens* e aplicaram o modelo em um classificador CRF. Santos et al., avaliaram



o impacto do modelo contextualizado Flair Embeddings aplicado a tarefa de REN junto com uma rede neural BiLSTM-CRF. Os autores também desenvolveram um modelo Flair Embeddings para o português, o FlairBBP, treinado com 4,9 bilhões de *tokens* (Santos et al., 2019b). Castro et al. (2018) utilizou uma rede LSTM e um classificador CRF junto com modelos *word embeddings* pré-treinados. Santos; Guimarães (2015) desenvolveram uma rede neural convolucional capaz de capturar características a nível de caracteres e também de incorporar *word embeddings* pré-treinados.

3.4.4.1.1 Reconhecimento de Entidades em Domínios Específicos

O reconhecimento de entidades tem sido aplicado em muitas áreas específicas, como direito, saúde e geologia. Nesses casos há uma demanda de adaptação dos modelos preditivos de acordo com a nova linguagem especializada do domínio e um novo conjunto de rótulos que devem ser aprendidos. Da mesma forma, são necessários novos conjuntos de dados para o processo de aprendizado, uma vez que abordagens de aprendizado de máquina necessitam de exemplos anotados para se chegar a um modelo preditivo eficaz.

Muitos trabalhos endereçam domínios específicos, citamos exemplos em diversas línguas. Para o inglês, uma rede neural BiLSTM-CRF para o domínio biomédico é proposta em (Habibi et al., 2017).

Um conjunto de dados do domínio jurídico em língua alemã é apresentado por Leitner et al. (2019), que empregam redes neurais BiLSTM para a rotulação dos textos. Em (Qiu et al., 2019), uma rede neural BiLSTM-CRF com mecanismo de atenção é aplicada para reconhecer entidades geológicas para a língua chinesa.

Para o português, um *corpus* para detecção de eventos de quedas de pacientes em prontuários eletrônicos é descrito em (Santos et al., 2020a). Os autores usaram uma rede neural BiLSTM-CRF+Flair para gerar um modelo classificador de *tokens*. Um *corpus* no domínio jurídico, tendo categorias específicas como legislação e jurisprudência é proposto por Araujo et al. (2018), que usaram uma rede neural BiLSTM-CRF para criar um primeiro *baseline* para esse *corpus*. Ademais, Consoli et al. (2020) analisam um *corpus* no domínio de geologia usando uma rede neural BiLSTM-CRF com um modelo contextualizado Flair Embeddings.

Desde a terceira edição, este livro conta com um capítulo de reconhecimento de entidades nomeadas no domínio jurídico, o Capítulo 13.

3.4.4.2 Extração de Relações

As abordagens iniciais para o problema de ER baseavam-se na definição de gabaritos e regras de extração, com base em informação sintática obtida de analisadores sintáticos rasos ou profundos (Cowie, 1983; Sager, 1978). Tais métodos foram rapidamente suplantados por métodos baseados em dados e padrões obtidos de *corpora*, como os famosos padrões de Hearst (1992) para identificação de relações de hiponímia.

O trabalho de Hearst (1992) se baseou na definição de padrões léxico-sintáticos para expressão de relações de hiponímia e hiperonímia a partir de uma análise de *corpus*. Ao escolher a relação de hiponímia, que ocorre em todo domínio, e padrões gerais baseados em aspectos da língua, como os representados no Quadro 3.3, garante-se a generalização dos padrões obtidos para diversos domínios e aplicações.



Quadro 3.3: Exemplos de Padrões de Hearst para hiponímia

Padrão	Exemplo	Extração
NP , tais quais $\{NP\}^*NP$	... países, tais quais França, Brasil e China	Is_a(França, país)
$NP\{, NP\}^*$ e outros(as) NP	Contusões, feridas, osso quebrados e outras lesões	Is_a(contusão, lesão)

Devido à dificuldade de construção manual das regras, os métodos de Riloff et al. (1993), empregam heurísticas para geração de padrões baseadas em informação gramatical, e de Soderland et al. (1995), que se baseia numa semântica de quadros (*frames*) empregando um analisador semântico e medidas de qualidade de identificação de exemplos, baseado no percentual de acerto sobre relacionamentos previamente conhecidos, para identificação de quadros relevantes.

As abordagens baseadas em aprendizado de máquina, hoje as mais comuns e com melhor desempenho na literatura (Konstantinova, 2014; Nasar et al., 2021) dividem-se em abordagens que realizam reconhecimento de entidades e extração de relações de forma conjunta e separada.

Abordagens baseadas na realização de REN e ER de forma separada baseiam-se em um fluxo de processamento em que, em geral, as entidades são identificadas primeiro e a tarefa de ER se reduz a identificar quando uma sentença ou fragmento textual denota uma relação semântica entre duas entidades. Consideremos o Exemplo 3.2, retirado de (Socher et al., 2012):

Exemplo 3.2:

[Gripe aviária]_{e1} é uma doença infecciosa causada pelo vírus da [influenza tipo a]_{e2}

Podemos, então, reduzir o problema de identificar a relação Causa-Efeito($e1, e2$) a um problema de classificação textual, identificando se a sentença acima fornece indícios para a expressão da relação de interesse. As soluções propostas na literatura para o problema são variadas e baseadas em diferentes métodos.

Zelenko et al. (2003), por exemplo, propõem funções de *kernel* para árvores sintáticas rasas, i.e. funções que descrevem medidas de similaridade entre tais árvores. Eles empregam tais medidas para treinar um classificador de *perceptron* com votação (*voted perceptron*) sobre relações no domínio de organizações extraídas de um *corpus* de textos jornalísticos. De forma similar, Zhao; Grishman (2005) empregam diferentes funções de *kernel* sobre informações sintáticas relevantes para a identificação de relação e argumentos visando treinar um classificador SVM sobre o *corpus* de ER da conferência ACE.

Culotta et al. (2006), por outro lado, empregam um classificador sequencial baseado em modelos escondidos de Markov para identificação de relações em um texto. Ao restringir sua análise a textos biográficos, os autores reduzem o processo de identificar instâncias de relações à identificação de fragmento textual que delimita o argumento e sua classificação, tarefa para a qual a classificação sequencial já é comumente utilizada. Consideremos o Exemplo 3.3 sobre George W. Bush, retirado de (Culotta et al., 2006):



Exemplo 3.3:

George é filho de George H. W. Bush e Barbara Bush.
pai mãe

Ao identificar o papel de pai e mãe, os autores conseguem construir a relação Pai(George H. W. Bush, George W. Bush) e Mãe(Barbara Bush, George W. Bush).

Métodos baseados em redes neurais, de forma geral, costumam empregar técnicas de aprendizado de representação (Bengio et al., 2013) para aprender representações do conteúdo semântico dos fragmentos textuais e reduzem o problema de ER à classificação textual. É o caso de Socher et al. (2012), que propõem a MV-RNN, uma rede neural que constrói um espaço de representação baseado em matrizes e vetores com o objetivo de capturar a composicionalidade de sentido de sintagmas e sentenças e os aplica para ER. Similarmente, Zeng et al. (2014) e Wang et al. (2016) empregam redes neurais convolucionais para obter representações vetorais de sentenças que serão empregadas no processo de classificação quanto à relação expressa pela mesma.

3.4.4.3 Extração Conjunta de Entidades e Relações

Abordagens baseadas em identificação sequencial de entidades e relações possuem desvantagens observadas na literatura. Primeiramente, como a ER é guiada pelas entidades identificadas no processo de REN, a propagação de erros da primeira tarefa pode ter impacto considerável na performance dos sistemas desenvolvidos. Segundo, uma vez que o contexto determinado limita tanto as tarefas de REN, quanto as de ER, existe uma interdependência entre as tarefas. Assim, propostas visando realizar a extração de entidades e relações de forma conjunta começaram a surgir na literatura recente, ganhando certo interesse da comunidade.

As abordagens empregadas para tal tarefa são diversificadas, incluindo desde métodos de aprendizado relacional a redes neurais

Roth; Yih (2007) propõem a utilização de métodos de programação inteira ao problema, baseados na teoria estatística de aprendizado relacional. Os autores utilizam classificadores locais para a identificação de entidades e relações e um classificador global que combina as informações dos classificadores locais em uma predição que maximiza a qualidade da extração, codificada por meio de restrições em programação inteira. Também baseados em modelos estatísticos, Yu; Lam (2010) propõem o uso de modelos gráficos globais para identificação de um descritor de relação e uma segmentação do texto para identificação dos argumentos.

Li; Ji (2014) e Miwa; Bansal (2016), por sua vez, reduzem a tarefa de ERE à classificação sequencial, utilizando redes neurais recorrentes bidirecionais sequenciais e estruturadas com base na estrutura superficial e na árvore de dependências sintáticas da entrada para identificação conjunta de entidades e relações.

3.5 Extração de Informação Aberta

A Extração de Informação Aberta (EIA), também conhecida como *Open Information Extraction*, Open IE ou OIE em inglês, é a tarefa de extrair informações estruturadas de documentos sem necessitar da pré-definição do contexto da tarefa, i.e. das relações e tipos de entidade de interesse. A tarefa foi inicialmente proposta pelo trabalho de (Banko et

al., 2007) e ganhou popularidade nas últimas décadas devido à sua aplicabilidade para processar e estruturar o conhecimento a partir de grandes volumes de dados disponíveis na Web, seguindo o paradigma da Web como um *Corpus* (WaC) (Meyer et al., 2003).

A EIA surge visando generalizar a tarefa de Extração de Relações. A principal diferença entre as duas abordagens, porém, reside na dependência da ER de uma especificação prévia do domínio de aplicação, bem como das relações alvo a serem identificadas, que a EIA visa eliminar.

Seguindo o trabalho original de Banko et al. (2007), que propôs o sistema TextRunner, vários métodos e sistemas para EIA foram propostos na literatura (Del Corro; Gemulla, 2013; Fader et al., 2011; Xavier et al., 2015), mas, como observado por Glauber; Claro (2018), os principais avanços na área se concentraram principalmente no idioma inglês.

A EIA para a língua portuguesa tem uma história bastante recente. A partir dos trabalhos de Souza; Claro (2014), Pereira; Pinheiro (2015) e de (Barbosa et al., 2016), têm crescido o número de estudos sobre a tarefa assim como os resultados obtidos por esses estudos, com recentes desenvolvimentos de métodos (Oliveira et al., 2022c; Sena et al., 2017; Sena; Claro, 2019, 2020; Souza et al., 2018), construção do *corpus* (Glauber et al., 2018) e avaliação dos sistemas disponíveis (Glauber et al., 2019a, 2019b; Malenchini et al., 2019).

Embora a área tenha visto um crescimento recente para o desenvolvimento de métodos para línguas como o inglês, principalmente com a aplicação de métodos supervisionados e redes neurais, esses avanços ainda não foram incorporados na literatura sobre EIA para a língua portuguesa. A razão para isso é principalmente a falta de recursos linguísticos disponíveis para orientar o desenvolvimento de pesquisas para a língua. Embora o foco no idioma inglês possa ser devido ao seu uso generalizado em todo o mundo, foi reconhecido pela comunidade científica que esse foco no inglês com suas características particulares pode introduzir algum viés na área (Bender, 2009).

Assim, esta seção aborda EIA para a língua portuguesa, incluindo uma formalização e a evolução das abordagens da área.

3.5.1 Formalização

A tarefa de EIA pode ser formalmente definida sendo $X = \langle x_1, x_2, \dots, x_n \rangle$ uma sentença composta de *tokens* x_i . Um extrator EIA é uma função que mapeia X em um conjunto $Y = \langle y_1, y_2, \dots, y_j \rangle$ como um conjunto de tuplas $y_i = \langle rel_i, arg1_i, arg2_i, \dots, argn_i \rangle$, que descrevem as informações expressas na sentença X . Neste capítulo, consideramos que as tuplas estão sempre no formato $y = (arg1, rel, arg2)$, onde $arg1$ e $arg2$ são sintagmas nominais, não necessariamente formados por *tokens* presentes em X , e rel é um descritor de um relacionamento entre $arg1$ e $arg2$. Não consideraremos extrações formadas por mais de dois argumentos neste capítulo.

3.5.2 Abordagens

Os primeiros métodos de EIA empregavam padrões de inspiração linguística para extração, como ArgOE (Gamallo; Garcia, 2015), ou adaptação de métodos para a língua inglesa, como SGS (Souza et al., 2018), InferReVerbPT Sena et al. (2017) e RePort Pereira; Pinheiro (2015). Os trabalhos são principalmente influenciados por métodos baseados no inglês da chamada segunda geração de EIA (Fader et al., 2011).

O primeiro sistema de EIA para o português de que temos conhecimento foi o DePOE (Gamallo et al., 2012). Ele executa a extração aberta multilíngue de triplas (inglês,



espanhol, português e galego) usando o analisador sintático de dependências baseado em regras *DepPattern*. No entanto, nenhuma avaliação ou resultados são relatados para a língua portuguesa. Os autores apresentam somente uma comparação dos seus resultados com *Reverb* na língua inglesa.

Souza; Claro (2014) se propuseram a analisar o conjunto de características mais representativas da língua portuguesa para a identificação de extrações válidas no contexto de EIA, tal qual empregado na língua inglesa com o sistema ReVerb (Fader et al., 2011).

O sistema RePort (Pereira; Pinheiro, 2015), por outro lado, é uma adaptação do ReVerb para a língua portuguesa baseada em análise sintática rasa com regras sintáticas e lexicais. Os autores relatam que suas extrações apresentam grande similaridade com suas correlatas extraídas pelo ReVerb (dos textos traduzidos para o inglês).

O RELP, proposto por Abreu; Vieira (2017), é um sistema aberto de extração de relações que extrai relações entre entidades nomeadas em um domínio de organização aplicando classificação sequencial com CRF (*Conditional Random Fields*). O sistema RelP extrai qualquer descritor de relação que expressa um relacionamento entre pares de entidades nomeadas (Organização, Pessoa ou Lugar), caracterizando-o como uma abordagem híbrida da REN com a EIA.

O InferReVerbPT desenvolvido por Sena et al. (2017) baseia-se numa adaptação do sistema ReVerb para a língua portuguesa, expandindo-o com a extração de relacionamentos implícitos obtidos por inferência por propriedades de simetria e transitividade das relações com inferência transitiva e simétrica. Um classificador SVM foi empregado para realizar a inferência baseado nas propriedades semânticas do verbo central no descritor de relação.

Souza et al. (2018) analisaram que a maior desvantagem dos estudos baseados em recursos linguísticos, como dados anotados, reside na escassez de tais recursos na maioria dos idiomas além do inglês. Assim, para mitigar esse problema, eles propõem um método de classificação de fatos baseado na similaridade de estruturas gramaticais (SGS). Sua abordagem modela estruturas morfosintáticas dos fatos (triplos descrevendo relacionamentos) para identificar padrões de semelhanças que podem ser usados para distinguir entre fatos válidos e inválidos. Eles aplicaram algoritmos de isomorfismo de grafos para detectar subgrafos descrevendo tais padrões.

Um novo sistema de EIA baseado em análise de dependência foi proposto por Gamallo; Garcia (2015), chamado ArgOE. Tal sistema é multilíngue, baseado em heurísticas e utiliza a informação de dependência sintáticas do texto para analisar a estrutura de dependência do verbo, bem como um conjunto de regras para gerar os relacionamentos. A introdução de um Analisador de Dependência em sistemas de EIA focados inteiramente na língua portuguesa foi feita pelos autores Oliveira et al. (2022c). O DptOIE é baseado em análise de dependência e regras elaboradas manualmente. As sentenças são pré-processadas por meio de um tokenizador, um PoS *Tagger* e um analisador de dependências. Os autores propõem um acoplamento de três módulos para tratar casos particulares: conjunções coordenadas, orações subordinadas e aposto.

Com a evolução dos métodos de EIA para a língua inglesa utilizando os modelos neurais, novas abordagens foram propostas também para a língua portuguesa.

O primeiro trabalho que utilizou aprendizado supervisionado com rede neural profunda para o português foi o de Ro et al. (2020) que descreve o sistema **Multi2OIE**. Os autores utilizaram o modelo de linguagem BERT multilíngue (Devlin et al., 2019) para obter representações vetoriais das palavras e reduzem a tarefa de EIA à classificação sequencial, identificado os fragmentos do texto que determinam os argumentos (arg_1, arg_2) e o descritor de relação (rel). Seu sistema foi capaz de produzir extrações para vários idiomas



(inglês, português e espanhol), treinados, entretanto, sobre dados traduzidos do inglês.

Stanovsky et al. (2018) propuseram uma abordagem de EIA para a língua inglesa baseada em triplas. Os mesmos fazem uso de uma classificação sequencial cuja limitação define uma tripla extraída para cada sentença. Este método utiliza uma arquitetura de Redes Neurais Recursivas (RNN) para realizar EIA. A EIA é formulada como uma tarefa de rotulagem de sequências, utilizando estratégias semelhantes às que foram aplicadas anteriormente a tarefas como o Reconhecimento de Entidades Nomeadas. Já os autores em Cui et al. (2018) e Zhang et al. (2017) propõem modelar o problema da EIA como um problema de aprendizado sequência a sequência (*seq2seq*). Eles definem uma estrutura *encoder-decoder* para aprender argumentos e tuplas de relação inicializadas a partir de um sistema de EIA.

Seguindo o trabalho de (Stanovsky et al., 2018), em 2022, Cabral et al. (2022) propuseram **PortNOIE**, uma arquitetura neural para EIA em português que combina representações contextuais de palavras com codificadores neurais para extrair relacionamentos baseado em classificação sequencial iterativa. Diferente de outros métodos de classificação sequencial para EIA, os autores focam na extração de múltiplas triplas de uma mesma sentença.

3.6 Avaliação

A avaliação sistemática de sistemas de EI foi estabelecida primeiramente nas conferências MUC, em particular na sua segunda edição, com o estabelecimento de gabaritos-padrão que deveriam ser utilizados por todos os sistemas participantes e a adoção de métricas de qualidade, baseadas naquelas usadas na área de recuperação de informação, que foram abordadas no Capítulo 2. Para avaliar a tarefa de extração de relações, a MUC-2 estabeleceu como métricas de qualidade do sistema as medidas de precisão e cobertura, também denominada de *Recall* ou Revocação.

A precisão de um sistema reflete a qualidade de suas extrações, i.e., quantas das extrações realizadas estão corretas, dado um *corpus* de teste. A medida de precisão pode ser calculada como:

$$P = \frac{\#(\text{relacionamentos corretamente extraídos})}{\#(\text{relacionamentos extraídos pelo sistema})} \quad (3.1)$$

A cobertura também conhecida como revocação, reflete quão abrangente um sistema é em suas extrações, i.e., quantas das extrações a serem realizadas em um *corpus* de teste, o sistema é capaz de realizar. A medida de cobertura pode ser calculada como:

$$R = \frac{\#(\text{relacionamentos extraídos})}{\#(\text{relacionamentos no corpus})} \quad (3.2)$$

Enquanto a MUC-3 adicionou duas novas métricas de avaliação, a saber sobre-geração (*overgeneration*) e sub-geração (*fallout*), tais métricas receberam pouco interesse na literatura. De fato, Lehnert; Sundheim (1991) argumentam que tais métricas foram pouco informativas ou difíceis de calcular para a tarefa de EI e, portanto, abandonadas. Foi também empregado nessa conferência um sistema automático de avaliação disponibilizado às equipes participantes que permitiu uma maior compreensão do modelo de avaliação e, como discutem Lehnert; Sundheim (1991), um avanço qualitativo nos sistemas gerados.

Além das medidas de precisão e cobertura, assim como em tarefas de classificação de texto e recuperação de informação, utilizamos a média harmônica entre essas medidas,



chamada medida F1, a fim de condensar a informação contida nas duas. A medida F1 pode ser calculada como:

$$F1 = \frac{2 * P * R}{P + R} \quad (3.3)$$

A avaliação da tarefa de REN segue padrões semelhantes aos aplicados à tarefa de ER. De fato, desde a MUC-6 (Grishman; Sundheim, 1996), as medidas de precisão, cobertura e F1 tem sido usada consistentemente como métricas de avaliação da tarefa de REN em diversos esforços de avaliação, como a CoNNL (Sang; De Meulder, 2003), para a língua inglesa, e das duas edições do HAREM (Gonçalo Oliveira et al., 2008; Santos et al., 2007), com exceção à ACE (Doddington et al., 2004) que apresenta uma combinação da tarefa de REN com reconhecimento de co-referência entre entidades e utiliza um sistema de pontuação próprio.

A avaliação de sistemas de EIA, por sua vez, possui algumas peculiaridades que precisam ser discutidas. Uma vez que a tarefa é postulada por Banko et al. (2007) como a extração de todas as relações identificadas em um dado fragmento textual, sem limitação de domínio de interesse, tal tarefa impõe imensa dificuldade aos esforços de avaliação.

De fato, Glauber et al. (2018) relatam um esforço de anotação de dados para a tarefa em língua portuguesa em que foram identificados por anotadores humanos mais de 400 relacionamentos em um *corpus* de 25 sentenças retiradas de textos jornalísticos e de enciclopédia. Assim, a avaliação de EIA deu-se, em grande parte de seu desenvolvimento e maturação, em conjuntos de dados não anotados, recorrendo a avaliações qualitativas das saídas dos sistemas e comparação direta por humanos das extrações obtidas.

Nesses esforços de avaliação, a precisão do sistema pode ser mensurada a partir da avaliação humana das saídas. Não é possível, entretanto, avaliar medidas como cobertura e F1, dada a inexistência de uma referência do conjunto total de relacionamentos a serem identificados. Assim, os autores da área propuseram diferentes métricas a fim de estimar tais valores, como a métrica rendimento (*yield*) (Fader et al., 2011; Schmitz et al., 2012).

A métrica de rendimento consiste no número de extrações válidas, i.e. corretas, de um dado sistema. Como calcular tal medida é, na maioria dos casos, impraticável dada a grande quantidade de extrações realizadas pelos sistemas, ela pode ser estimada a partir da precisão do sistema calculada sobre uma amostra aleatória das extrações realizadas (P'). Assim, podemos estimar o rendimento como:

$$Y = P' \cdot \#(\text{extrações realizadas}) \quad (3.4)$$

Foi também explorada a estratégia de criação (semi-)automática de conjuntos de dados usando vários sistemas (Del Corro; Gemulla, 2013), estratégias de supervisão fraca (Smirnova; Cudré-Mauroux, 2018), ou a geração de *corpora* para a tarefa a partir da transformação de anotações de tarefas próximas, como identificação de papéis temáticos (*Semantic Role Labeling*) por (Stanovsky et al., 2018). *Corpora* gerados de forma semi-automática vêm ganhando atenção na literatura recente, particularmente para a língua inglesa, devido a necessidade de dados anotados para se utilizar técnicas de aprendizado de máquina e redes neurais em EIA. *Corpora* como o OIE2016 (Stanovsky et al., 2018), Wire57 (Léchelle et al., 2018) e CARB (Bhardwaj et al., 2019) vêm se tornando *corpora* de referência em língua inglesa para o problema, apesar dos problemas existentes na construção de tais recursos – a não exaustividade das relações anotadas.

Para a língua portuguesa, foram propostas algumas iniciativas para avaliar os sistemas da OIE. Uma avaliação conjunta foi promovida durante o Fórum Ibérico de Avaliação de



Línguas (IberLEF) em 2019 (Collovini et al., 2019). A avaliação foi feita usando o *corpus* proposto por Glauber et al. (2018), que é composto por 442 relacionamentos extraídos de 25 frases de fontes como a seção em português da Wikipédia, o *corpus* CETENFolha, resenhas de filmes do portal Adoro Cinema² e o *corpus* Europarl. Apesar desta tarefa ter contemplado quatro cenários de avaliação, a avaliação geral dos sistemas permaneceu consistente nos diferentes cenários, indicando robustez nos resultados da avaliação. No geral, os sistemas DPTOIE (Oliveira et al., 2022c) e Linguakit (Gamallo; Garcia, 2015) tiveram o melhor desempenho, com o Linguakit2 dominando as avaliações de correspondência exata e o DPTOIE as avaliações de correspondências parciais (Collovini et al., 2019).

Outra abordagem de avaliação foi idealizada por (Malenchini et al., 2019). Seu foco foi a avaliação extrínseca dos sistemas de EIA através de sua contribuição na tarefa de respostas automáticas a perguntas. Os autores apresentaram um conjunto de dados de referência (*benchmark*) para avaliação extrínseca de sistemas de EIA em textos de língua portuguesa. Os sistemas que alcançaram os melhores valores na avaliação realizada pelos autores foram os sistemas ArgOE (Gamallo; Garcia, 2015), DependentIE (Glauber et al., 2019a) e DptOIE (Oliveira et al., 2022c).

3.7 Considerações finais

Este capítulo descreveu uma visão geral da área de Extração de Informação, apresentando a Extração de Informação Tradicional e a Extração de Informação Aberta. Transversalmente, apresentamos as formalizações necessárias e os conceitos fundamentais para a compreensão da EIA, assim como a avaliação da área e as heranças de outras áreas afins, tais como RI.

Nessa primeira versão, este capítulo descreveu de maneira bem sucinta as abordagens propostas para EI e EIA durante seu desenvolvimento histórico e as abordagens atuais da literatura, como as utilizando modelos de linguagens. Especificamente, a utilização da arquitetura Transformers, descritas no Capítulo [Modelos de linguagem](#) para as tarefas de EI e EIA tem sido bastante difundida para a língua inglesa e tem atuado em diversas áreas da PLN.

Agradecimentos

Agradecemos as colaborações dos autores deste Capítulo e suas indicações, assim como agradecemos a Adriana Pagano e Aline Macohin pela revisão e comentários.

²<https://www.adorocinema.com/>



Capítulo 4

Tradução Automática

Abordagens e Avaliação

Sheila Castilho

Helena de Medeiros Caseli

Publicado em: 26/09/2023

4.1 Introdução

A tradução automática (TA), também conhecida como tradução de máquina (em inglês, *machine translation* ou MT), refere-se à tradução de um texto eletrônico por um computador de uma língua para outra sem intervenção humana. Nesse sentido, convencionou-se chamar de língua (ou texto) fonte a língua de partida (origem) e língua (ou texto) alvo a língua de chegada (destino ou saída). Além de envolver a análise e interpretação (NLU) da língua-fonte e a geração (NLG) da língua-alvo, há a premissa fundamental de gerar uma saída que seja semanticamente equivalente (transmite o mesmo significado) à entrada.

Nos últimos anos, a TA evoluiu significativamente com o avanço de modelos estatísticos e neurais. Atualmente, ela é amplamente utilizada em todo o mundo por governos, indústria da tradução, consumidores finais e em pesquisas em uma variedade de aplicações.

Os primeiros sistemas bem-sucedidos de TA datam do final dos anos 1950 e início dos anos 1960¹, com os experimentos de Georgetown. No entanto, é possível encontrar referências a tentativas de tradução automática no século XVII (Hutchins, 2001). Desde então, diferentes abordagens para a TA foram desenvolvidas, incluindo abordagens baseadas em regras, exemplos, estatísticas e, mais recentemente, a TA neural, apresentadas brevemente nas diversas seções deste capítulo.

Hoje em dia, a TA desempenha um papel importante não apenas no âmbito comercial, mas também no âmbito social e político. Ela é amplamente utilizada em diversas aplicações de comunicação, que incluem²:

- Texto para texto: o usuário insere um texto-fonte e obtém uma versão traduzida em formato de texto;
- Texto para fala: o usuário insere um texto-fonte e obtém uma versão falada na língua-alvo;
- Fala para texto: o usuário fala no idioma fonte e obtém uma versão traduzida em formato de texto;
- Fala para fala: o usuário fala no idioma fonte e obtém uma versão falada na língua-alvo;

¹Para uma descrição abrangente da história da TA, sugere-se consultar alguns livros e capítulos sobre o assunto: (Hutchins, 2001), (Koehn, 2009), (Hutchins, 2010) e (Koehn, 2020).

²Todas as aplicações citadas estão disponíveis no Google tradutor. Disponível em: <https://translate.google.com.br/>.



- Imagem (de palavras) para texto: o usuário insere uma imagem contendo um texto e obtém a tradução desse texto.

Devido à ampla utilização da TA atualmente, seu impacto pode ser observado em nossa sociedade. Por essa razão, a avaliação da TA (tema da Seção 4.3) tornou-se mais importante, visando garantir a qualidade da tradução. Além da avaliação, este capítulo descreve as principais abordagens para a TA (Seção 4.2). No decorrer deste capítulo, alguns conceitos-chave são explicados para que você possa acompanhar os desenvolvimentos.

4.2 Abordagens

A tradução automática pode ser realizada de diversas maneiras, desde a mais simples (tradução direta), que envolve a tradução palavra-a-palavra (ou sequência de palavras), até a mais utilizada na atualidade, que é a tradução baseada em redes neurais artificiais (tradução neural). Na trajetória entre a tradução direta e a tradução neural, explicaremos também abordagens intermediárias, como a baseada em regras, a tradução por interlíngua e a tradução estatística.

Para tanto, vamos traduzir a sentença “A casa do meu avô é linda.” para o inglês. Como isso aconteceria de acordo com as abordagens mais utilizadas (ou tradicionais) na tradução automática?

4.2.1 Tradução direta

Na **tradução direta** ocorre o mapeamento direto de palavras-fonte para palavras-alvo sem passar por outros níveis de análise³. Assim, no nosso exemplo, os caracteres seriam combinados em palavras e cada uma seria mapeada diretamente para seu equivalente em inglês usando, por exemplo, um léxico bilíngue. Utilizando o léxico bilíngue disponível no github do MUSE⁴ e a tradução palavra-a-palavra, nossa saída seria como apresentado em Exemplo 4.1:

Exemplo 4.1:

Entrada: A casa do meu avô é linda.

Saída: `_ house _ my granddad _ beautiful .`

onde as palavras para as quais não se encontrou um equivalente no léxico consultado foram substituídas por `_`⁵.

Vejam que neste processo de tradução não há nenhum processamento referente às línguas envolvidas, uma vez que o resultado é obtido via **casamento de padrão**, seguido da substituição de uma palavra por outra com base em uma lista de pares de palavras.

³Vejam que aqui estamos usando “palavra” para denotar uma unidade lexical bastante comum no PLN. Contudo, vale ressaltar que outras unidades lexicais (n-grama ou expressão multipalavra, por exemplo) também podem ser usadas na tradução direta. Para saber mais sobre as unidades de processamento veja Capítulo [Sequência de caracteres e palavras](#).

⁴<https://dl.fbaipublicfiles.com/arrival/dictionaries/pt-en.txt>

⁵Isso porque palavras como “a” e “do” não têm entrada no léxico consultado, provavelmente por serem *stop words*. Para as demais palavras, a tradução foi gerada considerando apenas a primeira ocorrência de equivalente em inglês encontrada. Por exemplo, para “casa” existem as opções “house” e “home”, nesta ordem, e como “house” aparece primeiro, ela foi a escolhida para gerar a saída neste exemplo.

Obviamente a abordagem de tradução direta apresenta diversas **limitações**, como **não ser capaz de lidar com a estrutura (sintaxe) da língua**, que, como pode ser visto no Capítulo **Sequência de caracteres e palavras**, é fundamental para o tratamento adequado da língua. A tradução direta foi uma das primeiras abordagens a serem investigadas e não é mais utilizada nos tradutores atuais.

4.2.2 Tradução Automática Baseada em Regras

Como o nome sugere, os sistemas de TA baseados em regras (em inglês, *Rule-based Machine Translation* ou RBMT) são sistemas baseados em conhecimento desenvolvidos por meio da especificação de regras linguísticas, que levam em consideração morfologia, sintaxe e semântica das línguas envolvidas, além dos léxicos bilíngues de ambos os idiomas de origem e de destino⁶. Essas regras e léxicos são formuladas e criados manualmente por especialistas em linguagem.

A partir desses recursos, a RBMT é capaz de realizar mapeamentos mais complexos como os que consideram a sintaxe das línguas fonte e alvo. Assim, no nosso exemplo, uma possível regra que poderia ser aplicada é a que indica a inversão da posição do sujeito que possui a casa com o uso do “apóstrofo s” gerando “*My grandfather’s house is beautiful.*” ao invés de “*The house of my grandfather is beautiful.*”

Assim, vê-se que o processamento automático necessário para a tradução baseada em regras é um pouco mais complexo do que aquele aplicado na tradução direta, uma vez que agora é preciso saber o papel de cada palavra na sentença-fonte, ou seja, saber que “casa” é um substantivo e que “do” é uma combinação da preposição “de” e determinante (artigo definido “o”) para que a regra apresentada em Quadro 4.1 seja aplicada corretamente. A regra especifica que sempre que for encontrado, na sequência-fonte, um substantivo (<SUB>) seguido da preposição (<PREP>) “de” combinada (+) com os artigos (<DET>) “a” ou (|) “o” deve-se gerar como saída o apóstrofo (') seguido de “s” e o substantivo equivalente na língua-alvo. O símbolo “=>” separa o que deve ser considerado na língua-fonte (à esquerda) e o que deve ser gerado na língua-alvo (à direita).

Quadro 4.1: Exemplo de regra para a tradução automática baseada em regras

<SUB> <PREP/de>+<DET/[a|o]> => 's <SUB>

Para se determinar o “papel” de cada palavra na sentença o processamento necessário é o da etiquetagem morfossintática ou, do inglês, *part-of-speech tagging* descrito no Capítulo **Sequência de caracteres e palavras**.

Apesar de realizar um processamento automático um pouco mais complexo, a **desvantagem** da tradução baseada em regras não está aí, mas sim na necessidade de mapear o conhecimento linguístico em regras corretas, genéricas e abrangentes o suficiente para que sejam aplicáveis a vários exemplos. Vejam que esse mapeamento envolve, necessariamente, o conhecimento da língua-fonte, da língua-alvo, e de como o processo de tradução de uma para a outra deve ocorrer. Além de um **processo trabalhoso**, a geração de regras é também **limitada**, pois, como a língua está em constante mudança, o conjunto de regras gerado tem que ser constantemente atualizado e revisado. Além disso, a tradução

⁶A tradução baseada em regras enquadra-se na estratégia de tradução por transferência, na qual o mapeamento é realizado com base em uma análise da língua-fonte, seguida da aplicação de regras que fazem o mapeamento e a geração de equivalentes na língua-alvo.

de/para outra língua necessita de um novo conjunto de regras. Isso porque a tradução por transferência entre duas línguas requer que a representação do conhecimento extraído da língua-fonte, e que vai ser mapeado para a língua-alvo, seja capaz de abrigar todas as características de ambas as línguas, tornando-a específica para aquele par. Analogamente, o desenvolvedor tem que ter muito conhecimento de ambas as línguas ou a equipe deve contar com linguistas/tradutores, o que torna os **sistemas caros** de se implementar. Além disso, a saída dos sistemas de regras pode apresentar **pouca fluência**, já que as traduções são fornecidas por meio de regras.

A grande **vantagem** dos sistemas de regras é que, como não são necessários textos bilíngues para seu treinamento, esses sistemas são excelentes para **traduções de idiomas com recursos limitados**. Além disso, esses sistemas permitem que o desenvolvedor tenha um **controle maior**, sendo possível identificar exatamente onde estão os problemas, e a saída (texto-alvo) é relativamente previsível. Tanto regras quanto léxicos podem ser **refinados e personalizados**, com a adição de mais (novas) regras e entradas bilíngues para aprimorar a tradução. Outro ponto interessante, é que o conhecimento é **legível por seres humanos** o que facilita a manutenção. Os sistemas de TA baseados em regras foram os primeiros sistemas comerciais de TA na década de 1970 e abriram caminho para mais pesquisas em TA após o relatório ALPAC, que cortou os fundos (Seção 4.3.1).

4.2.3 Tradução por interlíngua

Mas será que esse conceito de transferência entre línguas não pode ser estendido para um número maior de línguas, ou seja, considerando um cenário multilíngue? Sim, essa é a ideia da **tradução por interlíngua**, que se propõe a usar uma língua intermediária – metalíngua – que é independente das línguas envolvidas na TA e ao mesmo tempo é capaz de representar informações de qualquer outra língua. Essa metalíngua, de natureza artificial, é não ambígua e, portanto, mais simples de processar do que qualquer linguagem natural. Assim, o processo de tradução entre duas línguas quaisquer é composto de duas etapas de traduções supostamente mais simples: uma realizada entre a língua-fonte e a metalíngua, e outra realizada entre esta metalíngua e a língua-alvo.

Mas será que essa abordagem simplifica o processo? Sim. Para ficar claro, imagine o seguinte cenário onde existem documentos escritos em n línguas distintas e queremos traduzi-los de e para quaisquer dessas línguas. Poderíamos construir arranjos de n tradutores⁷ automáticos para cada par de línguas (n_1 - n_2 , n_1 - n_3 , n_1 - n_4 , ..., n_2 - n_1 , n_2 - n_3 , n_2 - n_4 , ...) seguindo o modelo de transferência (tradução automática baseada em regras). Mas a tradução por interlíngua mostra-se mais vantajosa. Podemos definir uma metalíngua e dividir a tarefa em n grupos de especialistas/tradutores, cada um responsável por uma única língua, l_i (de preferência sua língua materna). Caberá a cada grupo construir um tradutor (lembre-se: bem mais simples!), ou codificador, de l_i para a metalíngua, e outro, decodificador, da metalíngua para l_i . Evidentemente todos os grupos compartilham o mesmo conhecimento sobre a língua intermediária.

Considerando que cada grupo fará sua parte, ao final, as possíveis combinações desses módulos de tradução darão origem a todos os tradutores almejados. Por exemplo, para traduzir l_i para l_j , juntamos o módulo codificador de l_i com o módulo decodificador para

⁷Para calcular a quantidade total de tradutores necessários, usamos a fórmula de arranjo de n línguas para combinação em pares ($k = 2$), $A_{n,2}$ que é $(n!)/(n-2)!$. Assim, para 5 línguas teríamos que construir 20 tradutores no sentido direita-esquerda e mais 20 tradutores no sentido esquerda-direita, totalizando 40 tradutores distintos!

l_j. Terão sido construídos $2n$ módulos de traduções mais simples, portanto, menos esforço do que o exigido para os tradutores bilíngues. Uma outra consequência é a possibilidade de se avaliar as traduções por meio da tradução inversa, já que os módulos independentes permitem a tradução nos dois sentidos.

Esse ideal foi compartilhado algumas vezes, no passado, por vários grupos de pesquisa acadêmica, mas infelizmente a prática evidenciou vários **problemas**. O maior deles é, segundo os críticos, a **ingenuidade** em se acreditar possível criar uma linguagem capaz de representar o significado de todas as outras, portanto, universal. Um outro problema – ou decorrente deste – é a adoção unânime de uma dada língua intermediária pelos grupos de línguas distintas, onde cada um deles reivindica alterações e adaptações, escancarando a **inexistência de um núcleo verdadeiramente universal**. Por essas e outras razões é que esse modelo não substituiu o modelo por transferência bilíngue.

No final dos anos 1990, o português brasileiro estava representado, pelo NILC⁸, numa iniciativa da ONU para construção de tradutores para as línguas mais faladas no mundo, o Projeto UNL⁹. Esse projeto tinha por objetivo o desenvolvimento de um sistema multilíngue de tradução automática baseada numa interlíngua de natureza semântica – a *Universal Networking Language* (UNL) – desenvolvida por pesquisadores vinculados à Universidade das Nações Unidas, órgão da ONU, em Tóquio¹⁰.

O **paradigma linguístico** (baseado em regras e interlíngua), no qual o conhecimento linguístico é explicitamente mapeado em recursos como regras, dominou o cenário da tradução automática **até a década de 1980**, quando abordagens baseadas em *corpus* (empíricas) surgiram. Aliadas à motivação de tentar superar as limitações da tradução baseada em regras, essas abordagens foram impulsionadas por dois fatores: (1) os **avanços no hardware** necessário para processamentos computacionais mais pesados, e (2) a **disponibilidade maior de recursos bilíngues**, em especial os *corpus* paralelos. As próximas seções tratam das abordagens baseadas em *corpus*: a tradução baseada em exemplos, a tradução estatística e a tradução neural.

4.2.4 Tradução Automática Baseada em Exemplos

Os sistemas de TA baseada em exemplos (do inglês, *Example-based Machine Translation* ou EBMT), também conhecidos como tradução por analogia, estão frequentemente associados à publicação do artigo de Nagao (1984), no qual o autor propõe um modelo baseado na imitação de exemplos de tradução de frases semelhantes, buscando utilizar a ideia de aprender a traduzir a partir de exemplos existentes (Koehn, 2020). Os sistemas de exemplos utilizam informações extraídas (sequências de palavras) de exemplos em *corpora* bilíngues de pares de tradução, alinhados em nível de sentença, ao qual convencionou-se chamar de *corpora* paralelos.

Por meio dessa abordagem, exemplos como os ilustrados no Quadro 4.2 serviriam de base para o sistema aprender traduções de trechos de texto, como as ilustradas em Quadro 4.3.

⁸<http://www.nilc.icmc.usp.br/>

⁹Esse projeto deu origem à *UNDL Foundation*.

¹⁰Mais informações sobre a linguagem UNL podem ser obtidas em (Uchida et al., 1999). Detalhes sobre o Projeto UNL-Brazil podem ser encontrados em (Martins et al., 2000; Nunes et al., 2003) e <http://www.nilc.icmc.usp.br/nilc/projects/unl.htm>.



Quadro 4.2: Exemplos para a tradução baseada em exemplos

A casa é muito bonita.	<i>The house is very beautiful.</i>
Meu avô foi internado ontem.	<i>My grandfather was hospitalized yesterday.</i>
Eu comprei uma jaqueta linda.	<i>I bought a beautiful jacket.</i>

Quadro 4.3: Trechos aprendidos

a casa	<i>the house</i>
meu avô	<i>my grandfather</i>
linda	<i>beautiful</i>

A partir dos trechos aprendidos, o sistema baseado em exemplos seria capaz de combiná-los para, a partir da entrada, gerar a saída apresentada em Exemplo 4.2:

Exemplo 4.2:

Entrada: A casa do meu avô é linda.
Saída: *The house _ my grandfather _ beautiful.*

Devido ao uso de diferentes métodos e técnicas, de acordo com Hutchins (2005, p. 63): “não parece haver um consenso claro sobre o que é ou o que não é um sistema de exemplos”. Para Carl; Way (2003, p. xix), uma definição analítica dos sistemas de exemplos era difícil, pois, segundo ele, tais sistemas “assumem uma posição entre os sistemas baseado em regras e os estatísticos” ao utilizar abordagens tanto baseadas em regras quanto orientadas por dados¹¹. No entanto, tanto os sistemas baseados em exemplos como os estatísticos se enquadram no paradigma de TA baseada em *corpus*. Enquanto alguns autores veem os sistemas baseados em exemplos como um paradigma em si mesmo, outros consideram os sistemas estatísticos como um tipo de sistema de exemplos, já que os primeiros sistemas estatísticos surgiram no final da década de 1980.

4.2.5 Tradução Automática Estatística

Os sistemas de TA estatísticos (em inglês, *Statistical Machine Translation* ou SMT) foram apresentados pela primeira vez por Brown et al. (1988); no entanto, a ideia de usar métodos estatísticos para traduções automáticas foi introduzida pela primeira vez por Weaver em 1949 (Brown et al., 1988, p. 71). Desde a primeira publicação de Brown et al., a equipe da IBM desenvolveu para a empresa o primeiro sistema estatístico funcional e houve um aumento drástico na pesquisa em TA estatística na área.

A ideia geral dos sistemas estatísticos é usar modelos estatísticos para extrair pares de tradução de *corpora* bilíngues. Podem ser encontradas três abordagens principais para a TA estatística:

¹¹Para uma descrição mais abrangente e para a história dos sistemas de TA baseados em exemplos, consulte (Carl; Way, 2003).

- TA estatística baseada em palavras (*Word-based Statistical Machine Translation*): alinha¹² palavras individuais no texto-fonte a palavras no texto-alvo e calcula a probabilidade da tradução. Também permite a exclusão e inserção de palavras.
- TA estatística baseada em frases (em inglês, *Phrase-based Statistical Machine Translation* ou PBSMT): alinha frases (não frases linguísticas, mas fragmentos de frases e palavras) no texto-fonte a frases no texto-alvo, comparando frases e seus vizinhos frasais ao considerar uma tradução. Essas frases também são chamadas de n-gramas, que são sequências contínuas de n palavras em sequência, ou seja, um unigrama é uma palavra, um bigrama são duas palavras, um trigramas são três palavras etc. A TA estatística baseada em frases é o tipo de abordagem estatística mais utilizado.
- TA estatística baseada em sintaxe: esses modelos traduzem unidades sintáticas usando árvores sintáticas geradas por analisadores sintáticos (Capítulo [Sequência de caracteres e palavras](#)).

Independente da estratégia escolhida, na **tradução estatística** a **probabilidade** determina como um texto-fonte deve ser traduzido para um texto-alvo. De acordo com a estratégia escolhida, essa probabilidade pode ser calculada considerando apenas palavras ou também sequências de palavras (frases). Essas frases são sequências de *tokens* (não necessariamente palavras) como “a casa do” ou “linda .” (onde o ponto final faz parte da frase). Seja considerando apenas palavras ou frases, a tradução é realizada com base em dois modelos computacionais: (1) um **modelo de tradução** que especifica como mapear texto-fonte em texto-alvo e (2) um **modelo de língua** que especifica como gerar um texto-alvo fluente. Desse modo, o modelo de tradução tenta maximizar a acurácia da tradução, enquanto o modelo de língua tenta maximizar a fluência da sentença gerada na língua-alvo (Seção 4.3).

Para tentar tornar esses conceitos menos abstratos, vamos retomar nosso exemplo da sentença “A casa do meu avô é linda.”. Nesse caso, o modelo de tradução poderia ser baseado em probabilidades de tradução de frases como os gerados para o *corpus* FAPESP (Aziz; Specia, 2011) com o auxílio do Moses¹³, como ilustrado na Tabela 4.1.

Tabela 4.1: Exemplos de frases fonte e alvo, e suas probabilidades, presentes em um modelo de tradução gerado a partir do *Corpus* FAPESP (Aziz; Specia, 2011).

id	frase-fonte (português)	frase-alvo (inglês)	probabilidade
1	a casa do	' house	0.0207779
2	a casa do	the house from	0.0623338
3	a casa	the house	0.297619
4	a casa	the home	0.0646474
5	do meu	of my	0.0813954
6	do meu	that of my	0.191576
7	meu avô	my grandfather	0.662453
8	meu avô , meu pai , eu	my grandfather , my father , me	0.0623338
9	avô	grandfather	0.916667
10	é	is	0.611613
11	é	é	0.794943

¹²No contexto da TA estatística, o alinhamento é uma tarefa de encontrar as correspondências entre texto-fonte e texto-alvo. Esse alinhamento pode se dar em nível de palavras (alinhamento lexical), de sentenças (alinhamento sentencial) entre outros.

¹³<https://www.statmt.org/moses/> e <https://github.com/moses-smt/mosesdecoder>



id	frase-fonte (português)	frase-alvo (inglês)	probabilidade
12	linda	beautiful	0.0389678
13	linda	pretty	0.00259724

Com base nas probabilidades da Tabela 4.1, diversas opções de tradução poderiam ser geradas como as apresentadas no Exemplo 4.3.

Exemplo 4.3:

- (a) *The house from my grandfather is beautiful.*¹⁴
- (b) *The house that of my grandfather is pretty.*¹⁵
- (c) *The home of my grandfather é beautiful.*¹⁶

E qual dessas sentenças o sistema escolheria como saída? Isso depende de alguns fatores que não vamos detalhar aqui, mas podemos dizer que o modelo de linguagem (Capítulo Modelos de linguagem) tem um papel fundamental na seleção da melhor sentença candidata. Nesse caso, o modelo de linguagem diz qual é a melhor sentença com base na probabilidade de ela ser encontrada na língua-alvo, ou melhor, no *corpus* de treinamento usado para gerar o modelo de língua-alvo.

Assim, tanto o modelo de tradução quanto o modelo de linguagem são treinados a partir de *corpus*. No caso da tradução estatística, as probabilidades são determinadas contando-se as frequências de ocorrência das palavras em grandes quantidades de textos (os *corpora*) paralelos. Um exemplo de *corpus* paralelo português-inglês é o coletado por Aziz; Specia (2011), o Corpus FAPESP¹⁷, que serviu de base para a geração das probabilidades apresentadas na Tabela 4.1. Para a geração do modelo de tradução, a probabilidade de uma frase em português ser traduzida para uma frase em inglês é calculada com base na co-ocorrência dessas frases no *corpus* paralelo. Para a geração do modelo de língua, a probabilidade de uma frase em inglês é calculada com base na ocorrência dessa frase na parte em inglês do *corpus* paralelo (ou de outro *corpus* monolíngue na língua-alvo).

Geralmente os modelos de tradução e de língua consideram frases de um tamanho máximo definido como parâmetro do treinamento. Quanto maior o tamanho da frase, maior é o contexto sendo considerado e, como consequência, mais coerente poderá ser a sentença gerada (Jurafsky; Martin, 2023). Porém, quanto maior o tamanho da frase, maiores serão o tempo e a quantidade de processamento necessários para realizar o treinamento dos modelos.

As **vantagens** dos modelos estatísticos são que, com **mais dados** utilizados no treinamento dos sistemas, não apenas a qualidade geral aumenta, mas, com o uso de um modelo de linguagem, as traduções estatísticas ganharam **fluência** em relação às abordagens anteriores. Além disso, a TA estatística permitiu um **uso mais eficiente de recursos humanos e de dados**. As **desvantagens** dessa abordagem são o **custo de criação de corpora paralelos**, especialmente para idiomas com recursos limitados. Além disso, a TA estatística tende a ter **dificuldade com pares de idiomas com ordem de palavras diferentes**.

Em 2007, o sistema *open-source* PBSMT mais famoso, desenvolvido por Koehn et al. (2007), foi lançado: o Moses SMT *toolkit*. Ao mesmo tempo, o Google lançou seu famoso

¹⁴Obtida por meio da combinação das frases 2, 7, 10 e 12.

¹⁵Obtida por meio da combinação das frases 3, 6, 9, 10 e 13.

¹⁶Obtida por meio da combinação das frases 4, 5, 9, 11 e 12.

¹⁷<http://www.nilc.icmc.usp.br/nilc/tools/Fapesp%20Corpora.htm>



Google Tradutor com abordagens estatísticas. Vale ressaltar que os modelos estatísticos conseguiram obter grande sucesso devido ao “aumento do poder de computação e armazenamento de dados, juntamente com a disponibilidade cada vez maior de recursos de texto digital como consequência do crescimento da Internet” (Koehn, 2009, p. 18). Devido à eficiência e precisão da abordagem estatística em relação às anteriores, ela se tornou a abordagem mais amplamente utilizada naquela época. Sistemas de tradução estatística baseada em frases (PBSMT) como os de Koehn et al. (2003) e Och; Ney (2004) eram o estado da arte até serem sucedidos pela tradução neural, a partir de 2015. De fato, a estratégia por trás do tradutor automático do Google foi a PBSMT por uma década (aproximadamente de 2006/2007 até 2016/2017)¹⁸. Atualmente, o Google e praticamente todos os sistemas de tradução online, bem como pesquisas nesta área usam a tradução neural (*neural machine translation*, NMT) ou algum sistema híbrido (estatístico e neural).

4.2.6 Tradução Automática Neural

Os sistemas de TA neural (em inglês, *Neural Machine Translation* ou NMT) foram introduzidos pela primeira vez na década de 1990 com alguns artigos sugerindo como redes neurais poderiam ser usadas para TA¹⁹ (Way; Forcada, 2018). No entanto, a quantidade dos dados usados para treinar esses modelos não era suficiente para produzir resultados razoáveis e, além disso, “a complexidade computacional envolvida excedia em muito os recursos computacionais daquela época, e, portanto, a ideia foi abandonada por quase duas décadas” (Koehn, 2020, p. 39).

Em geral, os modelos neurais consistem na construção de redes neurais *end-to-end* que mapeiam textos paralelos alinhados e são treinados para maximizar a probabilidade de uma sequência alvo Y , dada uma frase de origem X , sem informações linguísticas externas adicionais (Castilho et al., 2017b). Os sistemas neurais podem ser construídos com apenas uma rede em vez de uma sequência de tarefas separadas, como seu predecessor (a tradução estatística).

Com a publicação de resultados impressionantes em avaliação automática (Bahdanau et al., 2015; Bojar et al., 2016; Sennrich et al., 2016), os sistemas neurais geraram grande expectativa, especialmente porque a indústria de tradução busca melhorar a qualidade da TA para minimizar custos (Moorkens, 2017). A adoção dos sistemas neurais nos últimos anos tem sido extensiva, com um número crescente de provedores de TA e grupos de pesquisa concentrando seus esforços e recursos no desenvolvimento e implantação de sistemas NMT (Castilho et al., 2019a).

Na tradução neural, redes neurais artificiais são usadas para fazer a tradução de uma sentença-fonte para uma sentença-alvo. Uma rede neural artificial pode ser entendida como uma composição de diversas unidades de processamento (os neurônios artificiais) conectadas entre si, em camadas. Cada unidade de processamento recebe uma entrada numérica e gera uma saída numérica. A saída é calculada de acordo com os “pesos” (w) e as “entradas” (x) associados à unidade e uma função que determina como a saída deve ser calculada. Por exemplo, vamos supor que um neurônio artificial seja governado pela função x^2 . Nesse caso, se a entrada para esse neurônio for o número 2 então a saída será 4, se for 3 a saída será 9, se for -1 a saída será 1 e assim por diante. Os pesos são usados para

¹⁸Segundo informações disponíveis em: <https://ai.googleblog.com/2016/09/a-neural-network-for-machine.html>.

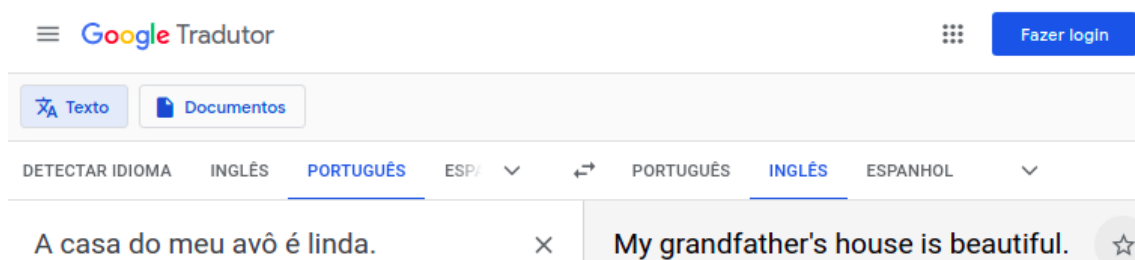
¹⁹Artigos como (Chalmers, 1992), (Chrisman, 1991), (Castano; Casacuberta, 1997), (Forcada; Neco, 1997) e (Neco; Forcada, 1997).



ajustar o aprendizado do neurônio e são uma das partes mais importantes da definição de uma rede neural artificial.

Na **tradução neural**, diversas camadas de neurônios são usadas para aprender como traduzir uma sentença-fonte em uma sentença-alvo a partir de um *corpus* paralelo. Esse aprendizado geralmente **demandava muito poder computacional²⁰ e tempo**, pois é realizado a passos pequenos em diversos ciclos de processamento²¹ das sentenças paralelas. E qual é a principal diferença metodológica da tradução neural para a estatística? Na NMT toda a sentença-fonte é considerada no aprendizado, de uma vez, e nos dois sentidos (da esquerda para a direita e da direita para a esquerda), ou seja, não há a quebra em frases como ocorria na PBSMT, nem a divisão clara entre modelo de tradução e modelo de língua. Dessa forma, a tradução gerada por um sistema NMT tende a ser mais fluente e natural, como ilustrado na Figura 4.1²².

Figura 4.1: Tradução gerada pelo Google tradutor (tradução neural)



Os modelos de tradução neural baseiam-se fortemente em duas tecnologias que se tornaram bastante usuais em PLN: *embeddings* e modelo de atenção. As *embeddings* (Capítulo [Representação vetorial e semântica distribucional](#)), são formas de representação de unidades lexicais (geralmente palavras) nas quais as unidades são mapeadas para vetores em um espaço de n (100, 300 ou mais) dimensões. Ao representar palavras como vetores densos notou-se que é possível mapear características linguísticas (morfológicas, sintáticas e semânticas) nesse espaço vetorial. Por exemplo, na Figura 4.2 é possível observar a proximidade semântica da palavra “avô” com outras palavras a partir das *word embeddings* do NILC²³ como “pai”, “tio”, “sobrinho” etc.

Usando as *word embeddings* bilíngues português e inglês do MUSE²⁴ é possível observar as similaridades entre línguas como ilustrado na Figura 4.3, onde as palavras em português aparecem em vermelho e as palavras em inglês, em azul. Essas *word embeddings* são usadas como forma de representação da língua nos modelos neurais de tradução automática.

Assim, a tradução neural não se baseia na combinação dos modelos de tradução e de língua, como faz a tradução estatística, mas sim em um modelo sequencial que prediz uma palavra por vez. O potencial deste modelo sequencial está na maneira como ele prediz as palavras: considerando toda a sentença-fonte e também o que já foi produzido para a sentença-alvo. Desde sua proposição, a modelagem sequencial neural passou por várias arquiteturas, indo desde as redes neurais recorrentes (em inglês, *recurrent neural network* ou RNN) usadas para codificação (em inglês, *encoder*) e decodificação (em inglês, *decoder*)

²⁰Geralmente são necessárias placas GPU (*Graphics Processing Units*), originalmente projetadas para processamento gráfico, capazes de fazer diversos cálculos em segundos.

²¹No contexto das redes neurais artificiais, um ciclo de processamento é chamado de época.

²²Tradução gerada pelo <https://translate.google.com.br/> em 23 de agosto de 2023.

²³<http://www.nilc.icmc.usp.br/embeddings>

²⁴<https://github.com/facebookresearch/MUSE>

Figura 4.2: Vizinhos mais próximos da palavra “avô” obtidos via consulta às *word embeddings* do NILC geradas usando o GloVe e dimensão 300.

```
[('pai', 0.7426180839538574),
 ('tio', 0.7307775020599365),
 ('sobrinho', 0.6814965009689331),
 ('irmão', 0.6784138679504395),
 ('avó', 0.6334453821182251),
 ('filho', 0.6296581029891968),
 ('paterno', 0.628169059753418),
 ('bisavô', 0.6196369528770447),
 ('sogro', 0.5976346731185913),
 ('amigo', 0.5973160862922668)]
```

até os mecanismos de atenção (em inglês, *attention mechanism*) (Bahdanau et al., 2015) que permitem ao decodificador focar em partes específicas da sentença de entrada em seu processo de geração da saída.

Nesse momento, os Transformers (Vaswani et al., 2017b) são o estado da arte na tradução. A Figura 4.4 ilustra a tradução da sentença de exemplo, em português, para inglês usando Transformers²⁵. Nessa ilustração, quanto mais clara (amarelo, verde claro, azul claro etc.) a célula que une a linha da palavra em inglês com a coluna da palavra em português, maior a “força” da relação entre elas. Por exemplo, observa-se uma forte relação entre “my” e “meu”, “grandfather” e “avô”, “home” e “casa”, e “beautiful” e “linda”.

Contudo, assim como todas as demais abordagens, a tradução neural também tem suas **limitações**. Uma delas é que, diferentemente da PBSMT onde é possível “olhar” para os modelos aprendidos e entender o que foi usado na tradução (como as frases da Tabela 4.1), a tradução neural é considerada uma **caixa-preta** (*black box*): entender o que pode ter sido considerado para gerar a tradução depende de desvendar uma visualização do modelo de atenção (como o da Figura 4.4), já que as previsões dos modelos neurais consistem em milhões de parâmetros. Isso **dificulta a extensão dos modelos** previamente treinados e coloca em dúvida a **robustez do sistema**. Além disso, por ser uma abordagem relativamente nova, a tradução neural ainda enfrenta alguns desafios, como o **desempenho ruim em condições fora do domínio** e **para idiomas com recursos limitados**.

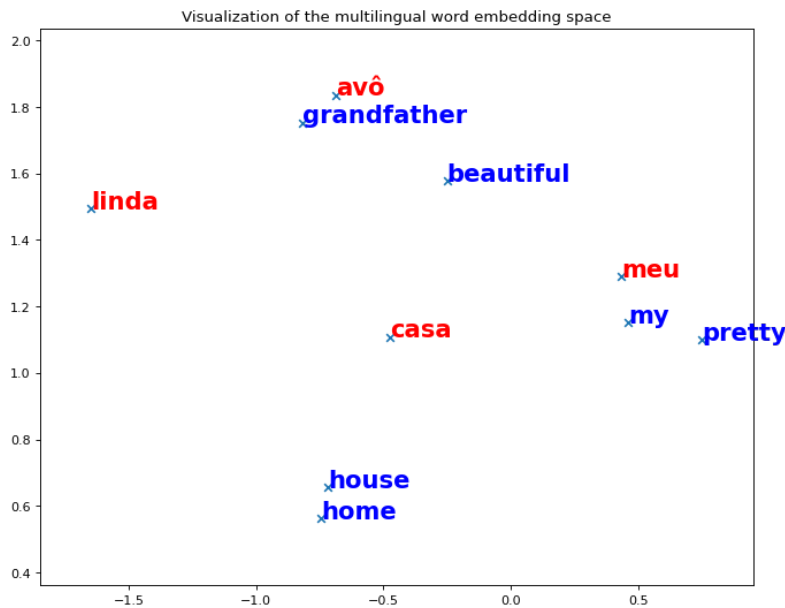
Além disso, é possível observar que as estratégias de tradução baseadas em *corpus* são **fortemente influenciadas pelo corpus usado no treinamento**. Por exemplo, os modelos estatísticos só terão a capacidade de traduzir uma palavra se ela tiver ocorrido um número significativo de vezes no *corpus* de treinamento, caso contrário não haverá uma frase correspondente contendo essa palavra e o sistema não conseguirá gerar uma tradução completa para a sentença original. No caso da tradução neural, isso é um pouco amenizado pelo uso de *embeddings* de unidades menores (em inglês, *subword units*) do que as palavras, as quais conseguem aproximar palavras desconhecidas às possíveis correspondências conhecidas²⁶. Por exemplo, se o “a” for esquecido no “linda” da Figura 4.1 o Google tradutor consegue gerar a mesma saída, sem problema. O tratamento de palavras e *subwords* é abordado no Capítulo **Sequência de caracteres e palavras**.

Outro ponto a se observar é que os sistemas neurais **precisam de um corpus maior**

²⁵O modelo foi treinado usando o Google Colab e o código disponível em [neste link](#) do GitHub do Brasileiras em PLN.

²⁶<https://ai.googleblog.com/2020/06/recent-advances-in-google-translate.html>

Figura 4.3: Visualização, em duas dimensões, das palavras em português (em vermelho) das palavras que ocorrem na sentença de exemplo e algumas possíveis traduções para o inglês (em azul).



e de melhor qualidade do que os estatísticos, pois eles são rápidos em memorizar exemplos mal-formados (Khayrallah; Koehn, 2018). Por isso, para algumas línguas com menos recursos (em inglês, *low-resourced languages*) os sistemas estatísticos ainda podem apresentar um desempenho melhor do que alguns sistemas neurais. E, por esse motivo, muitas pesquisas atuais têm enfatizado o desenvolvimento de técnicas de aumento de dados (em inglês, *data augmentation*) para sistemas neurais.

Apesar disso, os sistemas neurais são o estado da arte na área de tradução automática (agora em 2023), apresentando, especialmente, uma **fluência muito superior** aos sistemas estatísticos, o que dificulta a avaliação humana da tradução, a qual deve ser mais cuidadosa aos erros de acurácia.

Atualmente, existem várias técnicas de aprendizado profundo para os sistemas neurais, diferentes ramos, orientações de pesquisa e tendências²⁷.

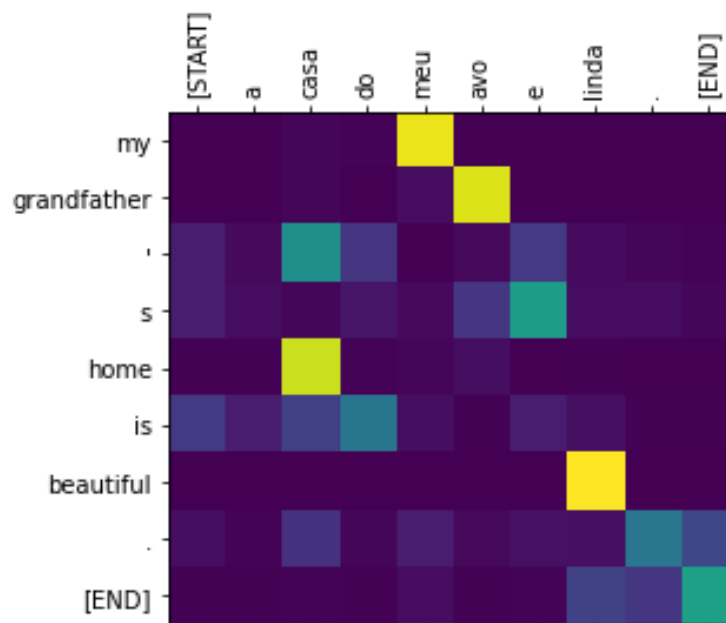
4.3 Avaliação da Tradução Automática

Devido à importância da tradução no mundo globalizado de hoje, o interesse na avaliação da qualidade da tradução (AQT – do inglês, *Translation Quality Assessment* ou TQA) cresceu a ponto de a avaliação da TA (ATA) se tornar um subcampo em rápido crescimento dentro da TA²⁸. No entanto, como a tradução é um processo multifacetado que envolve fatores cognitivos, linguísticos, sociais, culturais e técnicos, definir e medir a qualidade da tradução também reflete essa complexidade (Castilho et al., 2018). A AQT tem sido um tópico muito discutido em estudos de tradução, tecnologia da tradução e na indústria de

²⁷Para uma descrição abrangente do estado da arte da tradução automática neural, veja “Neural Machine Translation” de Koehn (2020).

²⁸Veja mais sobre Avaliação de sistemas de PLN no Capítulo [Avaliação de tecnologias de linguagem](#).

Figura 4.4: Visualização de um modelo de atenção usado para traduzir a sentença de exemplo.



tradução e localização, mas ainda não há muito consenso sobre o que é e como ela deve ser feita (Castilho et al., 2018) ²⁹.

Aqui, faremos a distinção entre a AQT e a ATA: enquanto a AQT abrange a avaliação tanto das traduções humanas quanto das traduções automáticas, a ATA se concentra exclusivamente na avaliação da qualidade dos sistemas de TA. Nesta Seção, iremos definir a avaliação da TA, apresentar diferentes abordagens e discutir algumas das avaliações mais influentes na sua história, destacando a importância de realizar a avaliação dos sistemas de TA.

4.3.1 O que é Avaliação da Tradução Automática?

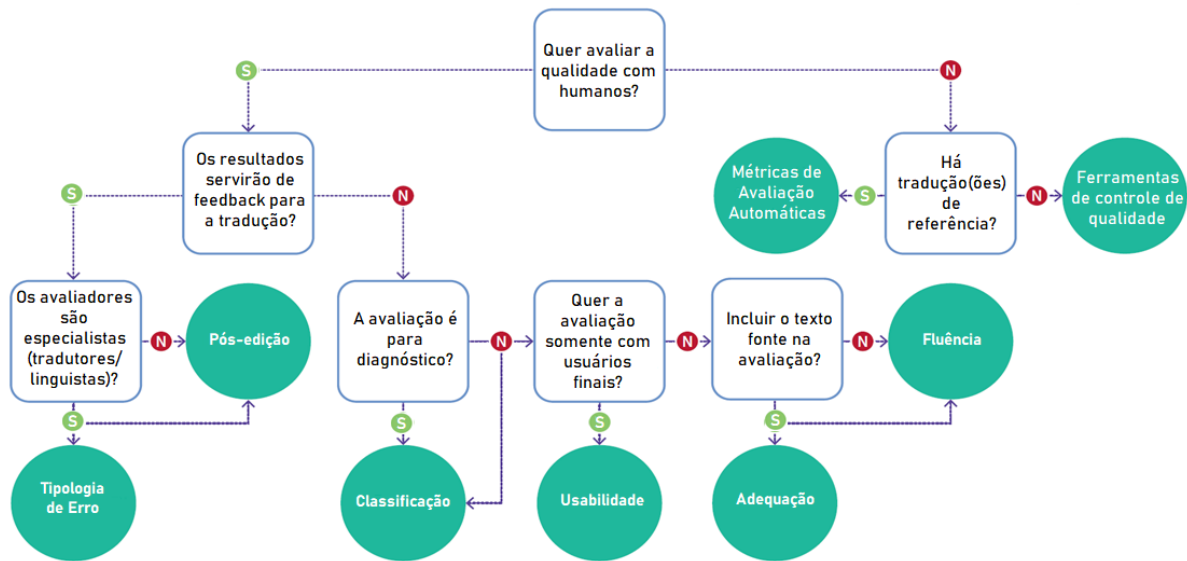
A avaliação da Tradução Automática (ATA) é a prática de analisar a saída de tradução de um sistema (ou sistemas) de TA e julgar a qualidade dessa tradução com base em critérios estabelecidos. As abordagens para a ATA incluem **avaliação automática**, usando métricas automáticas Seção 4.3.3, ou **avaliação manual humana** Seção 4.3.4, realizada por pessoas, e às vezes uma combinação das duas. O fluxograma da Figura 4.5 foi proposto por Doherty et al. (2018) para ajudar educadores e tradutores nos vários tipos de ATA.

Devido à falta de consenso no campo dos estudos de tradução e na indústria da tradução sobre o que constitui uma “boa” tradução, várias abordagens para a avaliação da TA surgiram. Portanto, a ATA pode ser realizada de várias maneiras diferentes, com várias abordagens, e não existe um único método que seja suficiente para abordar todos os propósitos de avaliação (Hovy et al., 2002).

Tradicionalmente, a avaliação da TA foi dividida em dois paradigmas: avaliação glass-box (caixa de vidro) e avaliação black-box (caixa-preta). Enquanto a primeira se preo-

²⁹Veja mais sobre a avaliação da tradução automática [neste vídeo](#).

Figura 4.5: Fluxograma da ATA baseado em Doherty et al. (2018)



cupada com “a qualidade de um sistema com base em suas propriedades internas” (Dorr et al., 2011, p. 744) e foi amplamente utilizada com sistemas baseados em regras (Seção 4.2.2), a última “mede a qualidade de um sistema apenas com base em sua saída, sem considerar os mecanismos internos do sistema de tradução” (ibid). As abordagens de avaliação black-box são o foco desta Seção.

A avaliação desempenha um papel essencial na TA, pois fornece informações sobre o funcionamento do sistema, quais as partes são eficazes e quais as áreas que precisam de melhorias. No entanto, a avaliação é um problema complexo, pois não existe uma única tradução correta para uma determinada fonte, e “pode haver várias traduções corretas possíveis” (ibid).

Para julgar a qualidade de uma determinada saída de TA, é necessário definir o que a “qualidade” significa para essa tarefa de tradução especificamente. Mas quem define o que é qualidade da tradução? A resposta para essa pergunta, definindo o que constitui uma tradução “boa” ou “ruim”, tem sido e ainda é amplamente debatida nos diferentes campos relacionados à tradução. Para O’Brien et al. (2011, p. 55), a qualidade está relacionada à opinião do cliente, mas na indústria da tradução, a avaliação de qualidade “é gerenciada por intermediários na cadeia de fornecimento e demanda”, geralmente usando uma abordagem única para todos (em inglês, *one size fits all*). Para a ATA, a avaliação de qualidade depende do uso pretendido da tradução. Portanto, qualidade pode significar uma tradução *fluente* que pareça ter sido escrita por um falante nativo; ou uma tradução *precisa* que transmita todo o significado expresso no texto de origem; ou talvez qualidade signifique uma tradução que seja ao mesmo tempo fluente e adequada. Além disso, qualidade pode ser definida como uma tradução fácil de ser pós-editada por alguns, enquanto outros podem definir qualidade como uma tradução que os usuários finais possam usar. E alguns podem ainda querer que todos esses critérios sejam atendidos para se ter uma tradução de qualidade, enquanto outros podem definir qualidade da tradução com outros critérios totalmente diferentes.

Em resumo, ao avaliar a saída da TA, essencialmente estamos tentando determinar graus de “qualidade” e “não qualidade” dessa tradução para um público específico, uma vez que

um critério de qualidade que seja crucial em determinada cenário, pode ser irrelevante em outros (Dorr et al., 2011).

Nesse sentido, alguns projetos de tradução que merecem referência são:

- **ALPAC Report (1966)**: examinou a qualidade e eficácia dos sistemas de TA nos Estados Unidos. O relatório concluiu que a TA não era útil devido à baixa qualidade, o que resultou em uma redução significativa no financiamento para o processamento de linguagem natural (PLN) e TA naquela época. O relatório também desencorajou discussões sobre avaliação dentro da comunidade de PLN por muitos anos.
- **DARPA Initiative (1992)**: uma das primeiras iniciativas na avaliação de TA nos anos 1990. A metodologia de avaliação era baseada em julgamentos humanos, onde os avaliadores atribuíram uma pontuação de 1 a 5 para as frases traduzidas automaticamente, comparando-as ao texto original e a traduções de referência humanas produzidas por tradutores profissionais.
- **FEMTI (2002)**: foi um esforço para padronizar o processo de avaliação de TA, organizando os métodos existentes de ATA e relacionando-os com o propósito e contexto dos sistemas, fornecendo “diretrizes para a seleção de características de qualidade a serem avaliadas, dependendo da tarefa esperada, dos usuários e das características de entrada de um sistema de tradução automática” (Estrella et al., 2009).
- **GALE (2005)**: teve como objetivo desenvolver e aplicar tecnologias de software para TA de grandes volumes de fala e texto em várias línguas. Foi estabelecido um protocolo de avaliação que utilizava métodos automáticos, humanos, baseados em tarefas e semi-automáticos, com foco na taxa de erro de tradução mediada por humanos (HTER).
- **EuroMatrix (2006)**: financiado pela Comissão Europeia, esta iniciativa construiu sistemas de TA estatísticos e híbridos para diversos idiomas europeus. A avaliação dos sistemas de tradução automática incluiu avaliações automáticas com base em onze métricas, incluindo BLEU (ver Seção 14.2.2), METEOR, Translate Error Rate, Word Error Rate, entre outros (Seção 4.3.3); e avaliações humanas que consistiram em classificações de fluência e adequação, *ranking*, tempo de leitura, tempo de pós-edição, teste de preenchimento de lacunas, clareza e informatividade (“Euromatrix. Survey of Machine Translation Evaluation”, 2007).
- **WMT (2006)**: começou como o Workshop de Tradução Automática Estatística e progrediu para a Conferência de Tradução Automática em 2016, tem sido uma grande contribuinte para o avanço da pesquisa em avaliação de TA. Ao longo dos anos, a WMT tem utilizado diferentes metodologias de avaliação. Desde 2008, a WMT realiza uma tarefa compartilhada, ou esforço colaborativo, (em inglês, *shared task*) de avaliação que compara uma ou mais traduções de referência com as saídas dos sistemas de TA e tem utilizado, principalmente, a métrica BLEU, embora essa tenha sido superada por diversas métricas diferentes. Em relação à avaliação manual, a WMT experimentou e implementou uma ampla gama de metodologias, incluindo julgamentos de fluência e adequação (2006 e 2007), diferentes avaliações de classificação (2007-2016), pós-edição e compreensão de frases (2009-2010), e avaliação direta e conjuntos de testes (desde 2016). Além disso, a *shared task* da WMT vem realizando avaliações em nível de documento desde 2019.
- **TAUS (2005)**: fundada em 2005 como uma rede de dados linguísticos com um extenso repositório de dados e uma rede de engenharia de linguagem humana, a

TAUS tem estado na vanguarda das tentativas de estabelecer indicadores para a avaliação da qualidade de tradução, desenvolvendo o Framework de Avaliação Dinâmica de Qualidade (DQF) em 2011. O modelo é uma estrutura de avaliação baseada em diferentes tipos de conteúdo, propósito do uso, ferramentas, processos e outros aspectos.

- **QTLaunchPad (2012):** projeto colaborativo financiado pela União Europeia (2012-2014) que reuniu e forneceu dados e ferramentas para a avaliação da qualidade de tradução, além de métricas de qualidade compartilhadas para tradução humana e automática (Uszkoreit; Lommel, 2013). O projeto contribuiu para o projeto de TA QT21 (2015-2018), cujo objetivo era desenvolver uma avaliação aprimorada. O QTLaunchPad desenvolveu o *framework* de Métricas de Qualidade Multidimensionais (MQM), que descreve e define métricas personalizadas de qualidade de tradução para avaliar a qualidade de textos traduzidos. MQM é amplamente utilizado como um padrão para métricas de avaliação humana em MTEval. Como parte do QT21, o MQM e o DQF (da TAUS) foram harmonizados, e atualmente, a Tipologia de Erros DQF da TAUS é um subconjunto reconhecido do MQM.

4.3.2 A importância de uma avaliação replicável

Como visto neste Capítulo, a avaliação da TA tem sido realizada desde o surgimento da própria TA. No entanto, o relatório ALPAC tornou a avaliação quase um tópico proibido na comunidade de PLN (Paroubek et al., 2007), e até hoje, a avaliação de TA, especialmente a avaliação humana, é considerada “desnecessária” em alguns momentos. Não é raro encontrar artigos de pesquisa usando uma única métrica para avaliar seu sistema (geralmente pontuações BLEU – ver Seção 4.3.3, e ainda mais comum são os artigos de pesquisa com avaliação humana muito limitada e falha, e até mesmo artigos de pesquisa sem avaliação humana alguma (Marie et al., 2021).

O avanço dos sistemas de TA gerou muita expectativa na comunidade, especialmente porque a indústria de tradução buscava uma melhoria na qualidade da TA para reduzir custos (Moorkens, 2017). À medida que os sistemas de TA neural se tornaram líderes no mercado, com uma clara melhoria na qualidade das traduções, surgiram reivindicações exageradas de que esses sistemas estavam “preenchendo as lacunas entre a tradução humana e a tradução automática” (Wu et al., 2016). Diante disso, pesquisadores na área de avaliação de TA alertaram a comunidade para ter cautela e não fazer promessas exageradas, além de enfatizarem a necessidade de mais pesquisas para lidar com as limitações dos sistemas de TA e realizar avaliações humanas mais abrangentes (Castilho et al., 2017b).

No entanto, as afirmações exageradas continuaram, com alguns declarando que seu sistema de TA neural havia atingido a “paridade humana” (em inglês, *human parity*) (Hassan et al., 2018) e outros alegando que a TA é um problema “resolvido” com uma qualidade de tradução “quase perfeita”. Como resposta, a fim de verificar essas alegações, dois estudos independentes (Toral et al., 2018) e (Läubli et al., 2018) reavaliaram os dados utilizados por Hassan et al. (2018) e descobriram que a escolha dos avaliadores, o contexto linguístico e a criação de traduções de referência têm um impacto significativo na avaliação de qualidade, apontando para falhas nas práticas atuais em avaliação de TA.

Como resultado, os pesquisadores de avaliação TA passaram a se envolver em discussões mais aprofundadas sobre a necessidade de aprimorar continuamente a metodologia de avaliação, “com o objetivo de superar as limitações das métricas automáticas e das abordagens humanas, evitar superestimativas na capacidade da TA e explicar os resultados



aparentemente contraditórios da TA de forma mais abrangente” (Castilho et al., 2019a, p. 2). Além disso, os pesquisadores começaram a buscar diretrizes para avaliar a “paridade humana” na avaliação da TA (Läubli et al., 2020).

Então, quem precisa de avaliação de TA? Em resumo, todos os campos diferentes relacionados à tradução. Algumas questões precisam ser abordadas para que a avaliação seja confiável, tais como:

- O que a **qualidade** significa nesta avaliação? Como vimos, há um grande debate sobre a definição de qualidade, então, antes de avaliar um sistema de TA, o pesquisador precisa definir o que seria uma tradução “boa” e “ruim” nesse cenário.
- **Que tipo de sistema** está sendo avaliado? Dependendo do par de idiomas, os sistemas estatísticos podem mostrar um problema específico. Já os sistemas neurais são conhecidos por serem fluentes e conhecer a similaridade entre palavras, o que torna mais difícil detectar erros.
- Qual é o **objetivo** a ser alcançado com esta avaliação? Saber por que o sistema de TA está sendo avaliado é importante para decidir que tipo de avaliação precisa ser realizada. Por exemplo, se alguém quer saber se as implementações realizadas em um sistema de TA feito para lidar com expressões multipalavras resultaram em traduções corretas dessas expressões, uma análise linguística aprofundada da saída seria mais apropriada do que avaliar a fluência da saída.

Como podemos ver, a avaliação de TA é um problema complexo, e não é surpresa que tenha se estabelecido como um campo independente. Com os avanços na qualidade dos sistemas de TA atuais, a comunidade de avaliação precisa estar atualizada com os procedimentos que possam fornecer avaliações mais sólidas, capazes de confirmar ou refutar as alegações feitas, como afirmou Carl Sagan: “Alegações extraordinárias exigem evidências extraordinárias”.

4.3.3 Métricas Automáticas

Como vimos anteriormente, a ATA assume uma complexidade intrínseca devido a uma multiplicidade de fatores. Tipicamente, encontramos duas vertentes de avaliação: a avaliação automática e a avaliação humana (manual), ocasionalmente mescladas para compor uma abordagem híbrida. Nesta Seção, abordaremos as métricas automáticas mais predominantes, reconhecendo a sua influência no domínio da TA.

As métricas automáticas pioneiras empregadas na TA tiveram origem em outras áreas da PLN. Por exemplo, temos a *Word Error Rate* (WER), introduzida por Su et al. (1992), que originou-se do campo de reconhecimento da fala (Capítulo [Texto ou fala?](#)). Por sua vez, a ROUGE, desenvolvida por Lin (2004), teve sua origem na sumarização automática. Outra métrica muito usada anteriormente foi a *F-measure*, empregada em recuperação de informação (Capítulo [2](#)) e em diversas outras áreas do PLN, também encontrou aplicabilidade nesse contexto.

As Métricas de Avaliação Automática (MAA) são adotadas na TA quando se busca evitar a intervenção humana direta (Figura [4.5](#)). As MAAs atuam como programas computacionais, recebendo as traduções de um sistema de TA e as traduções de referência (TR) como entrada, e produzindo uma pontuação numérica que reflete a similaridade entre as traduções de TA e TR.

A maioria das MAAs é classificada como métricas de referência (em inglês, *reference-based metrics*), exigindo a disponibilidade da TR, isto é, a tradução humana do texto em



avaliação, a fim de serem empregadas como ponto de comparação. No entanto, abordagens mais recentes incorporam modelos de linguagem pré-treinados para medir a semelhança entre uma tradução gerada e um conjunto de referências. Nesse contexto, essas métricas medem a similaridade entre as representações semânticas das palavras e as frases presentes tanto nas traduções geradas quanto nas referências, utilizando recursos linguísticos capturados durante o pré-treinamento desses modelos, em detrimento de uma comparação direta com traduções humanas específicas.

As MAAs preponderantes na área da TA são aquelas que operam sobre características lexicais e dispensam a necessidade de treinamento (em inglês, *untrained*). Essas métricas, geralmente baseadas em similaridades (em inglês, *matching*) e diferença (ou distância) de edições (em inglês, *edit distance*) entre o resultado da TA e a TR, avaliam a **sobreposição** entre a hipótese (resultado da TA) e a TR. Tal avaliação contempla tanto a precisão quanto a abrangência dos elementos lexicais (Lee et al., 2023). Duas vertentes são identificadas entre as métricas lexicais: as *word-based* (baseadas em palavras), que analisam as similaridades entre palavras; e as *character-based* (baseadas em caracteres), que investigam a similaridade entre caracteres.

As métricas lexicais *word-based* mais amplamente empregadas permitem medir tanto a similaridade dos n -gramas quanto a distância de edição (*edit distance*). Dentre as métricas baseadas em n -grama, destacam-se as amplamente conhecidas BLEU (Papineni et al., 2002), METEOR (Banerjee; Lavie, 2005) e NIST (Doddington, 2002). Por outro lado, as métricas que calculam a distância de edição e que têm destaque são TER (e HTER) (Snover et al., 2006) e WER (Su et al., 1992). Vale mencionar a singularidade da métrica chrF (Popović, 2015), a qual, além de ser *character-based*, também mede a similaridade dos n -gramas.

Mais recentemente, métricas treinadas em modelos baseados em redes neurais usando a arquitetura Transformer foram propostas. Dentre essas, há as métricas supervisionadas (*supervised-metrics*) e as não-supervisionadas, dependendo da técnica de aprendizado, ambas categorias com *word-embeddings* e *contextual-embeddings* (Lee et al., 2023). Entre as métricas não-supervisionadas mais recorrentes, destacam-se MEANT (*word embedding*) (Lo; Wu, 2011), BERTscore (Zhang et al., 2020), Yisi (Lo, 2019) e BARTscore (Yuan et al., 2021a) (*contextual-embedding*). Entre as supervisionadas, estão a BEER (Stanojevic; Sima'an, 2014) e BLEND (Ma et al., 2017) (ambas *word-embeddings*), BERT for MTE (Shimanaka et al., 2019), BLEURT (Sellam et al., 2020) e COMET (Rei et al., 2020).

As **vantagens** das MAAs são que elas são **eficientes**, **econômicas** e fornecem avaliações **consistentes**, ou seja, se a métrica for computada para a mesma tradução várias vezes, todas elas vão dar o mesmo resultado. No entanto, uma **preocupação** é a **dependência** exclusiva das similaridades entre a saída do sistema e as referências. Primeiramente, não há somente uma única tradução correta para um texto, sendo assim, o significado do texto pode ser traduzido de várias maneiras. Mas seriam todas as traduções “igualmente boas”? E o que “boa” significa nesse determinado contexto da tradução? Nesse caso, usar múltiplas TRs seria essencial para se ter uma avaliação mais justa. Segundo, as MAAs não oferecem insights detalhados sobre erros de tradução, pontos fortes e fragilidades de um sistema. Elas não dizem o que funciona no sistema, o que precisa ser melhorado; sendo o único objetivo medir a semelhança com a(s) referência(s), e consequentemente, as melhorias específicas decorrentes de modificações no sistema de tradução permanecem obscuras. E finalmente, o sistema com uma pontuação menor pode ser melhor na prática do que um sistema com uma pontuação mais alta. Enquanto as MAAs servem como ferramentas quantitativas valiosas, elas não revelam completamente as complexidades da qualidade da

tradução. Uma abordagem mais abrangente, combinando MAAs com avaliações humanas e análises qualitativas, oferece uma compreensão mais profunda do desempenho dos sistemas de TA.

Embora as MAAs não se revelem apropriadas para mensurar a qualidade final dos sistemas, impulsionam o avanço da pesquisa em TA, uma vez que podem ser empregadas de forma constante durante o desenvolvimento e a implementação desses sistemas. Em essência, as MAAs são medidas úteis na comparação entre sistemas de TA ou de versões de um mesmo sistema de TA, mas são limitadas na predição da qualidade da tradução.

4.3.4 Métricas Humanas

O processo de avaliar a qualidade da TA por meio da intervenção humana é essencial. Embora as MAAs proporcionem uma avaliação quantitativa, a avaliação humana oferece uma visão mais detalhada e uma análise mais ampla de fenômenos linguísticos complexos subjacentes ao desempenho dos sistemas de tradução, sendo assim imprescindível em uma compreensão mais abrangente dos sistemas de TA.

A avaliação humana pode ser feita através de diversos paradigmas, sendo os mais comuns o paradigma de **fluência-adequação** e **pós-edição**. A abordagem de **ranqueamento** de segmentos (em inglês, *ranking*) também é comumente empregada para a comparação dos sistemas de tradução, e possibilita a avaliação comparativa de diversos sistemas, fornecendo insights sobre a eficácia relativa de suas saídas. Igualmente, a **anotação de erros**, sob a forma de marcações específicas, oferece um feedback valioso sobre os sistemas em análise.

Outras abordagens incluem métricas secundárias, tais como legibilidade, compreensibilidade e usabilidade das traduções resultantes. Vale mencionar métricas centradas no usuário, que são avaliadas com testes de compreensão, por exemplo, que podem ser utilizados para aferir não apenas a fidelidade à tradução, mas também a transmissão efetiva da mensagem subjacente. Essa abordagem complementar permite uma apreciação mais holística da eficácia das traduções geradas.

Importante ressaltar que a avaliação humana é realizada por uma variedade de avaliadores, incluindo tanto profissionais quanto amadores. Essa diversidade de perspectivas pode contribuir para uma avaliação mais robusta e representativa da real eficácia dos sistemas de TA se empregados na mesma avaliação. Porém, também pode resultar em conclusões diferentes se a avaliação é feita só com um ou com o outro (só amadores ou só profissionais, por exemplo). Essa Seção vai abordar as métricas humanas mais comuns na avaliação de TA, assim como a importância de se calcular a concordância entre anotadores.

4.3.4.1 Adequação: Fidelidade Semântica

A **adequação** (em inglês, “*accuracy*” ou “*fidelity*”), ou também “acurácia” ou “exatidão”, é uma métrica importante na avaliação humana da TA. Ela foca na **fidelidade semântica** entre o texto-fonte e a tradução, mostrando a profundidade e a precisão da transferência de significado, permitindo determinar se a mensagem que está sendo transmitida e seu sentido foram preservados de maneira precisa e fiel.

Essencialmente, a adequação investiga até que ponto a tradução transmite o significado do texto de origem para o texto-alvo. Nesse contexto, uma escala Likert³⁰ é utilizada

³⁰A escala Likert é um método de medição criado por Likert (1932) que apresenta ao respondente uma afirmação ou pergunta e solicita que o respondente avalie o grau em que concorda com ela. A escala envolve uma série de itens ou afirmações aos quais os respondentes atribuem níveis de concordância ou discordância.



para classificar o nível de transferência semântica, onde geralmente uma pergunta é feita para o avaliador: “Até que ponto a tradução transfere o significado do texto-fonte para o texto-alvo?”. Como resposta, o avaliador pode escolher uma opção em uma escala que varia desde “Nada” até “Tudo”, com os graus intermediários de “Pouco” e “Muito”.

As **limitações** da adequação é que ela **não fornece informações sobre a fluência** da tradução, deixando uma lacuna importante na avaliação. Em determinados casos, o foco reside exclusivamente no significado da sentença de origem, tornando a fluência uma preocupação secundária. Além disso, a adequação não oferece detalhes precisos sobre os erros presentes na tradução, o que pode dificultar a identificação de pontos específicos para a melhoria do sistema.

4.3.4.2 Fluência: Naturalidade e Estrutura

A **fluência**, ou “inteligibilidade” (em inglês, “*fluency*” ou “*intelligibility*”), é outra métrica importante na avaliação humana da TA, que se preocupa com a **naturalidade e a estrutura** do texto-alvo, revelando o grau de fluência e adaptabilidade da saída da TA às normas linguísticas e socioculturais da língua-alvo.

Essa avaliação foca diretamente no **texto-alvo**, priorizando a avaliação da gramática e dos aspectos estruturais da tradução. Essencialmente, ela investiga o quão natural e fluido é o fluxo do texto-alvo dentro do contexto da língua-alvo, considerando suas normas linguísticas e socioculturais específicas. Uma característica distintiva da dimensão de fluência é que ela **pode ser avaliada independentemente do texto-fonte**, uma vez que se concentra exclusivamente no resultado final da tradução.

Também medida numa escala Likert, uma pergunta típica dirigida ao avaliador poderia ser: “Quão fluente está o texto-alvo, ou seja, como está o fluxo e a naturalidade do texto-alvo no contexto da língua-alvo e suas normas linguísticas e socioculturais em um dado contexto?”. A escala Likert pode variar desde “Sem fluência” até “Nativo”, incluindo os graus intermediários de “Pouca fluência” e “Quase nativo”.

Essa métrica oferece informação sobre a naturalidade da tradução e se ela soa natural e fluída para um falante nativo ou se exibe características de “linguagem quebrada”, prejudicando assim a experiência de leitura e compreensão.

Assim como a adequação, a avaliação da fluência também apresenta **limitações**. Ela **não proporciona informações sobre a adequação** da tradução, pois o foco está exclusivamente na fluência da sentença de destino, tornando a adequação uma preocupação secundária. Além disso, a fluência também **não oferece detalhes precisos sobre os erros presentes na tradução**.

Por esse motivo, é comum que as avaliações de adequação e fluência caminhem juntas, uma vez que é mais intuitivo avaliar uma em relação à outra. No entanto, há momentos em que pode ser necessário priorizar uma em detrimento da outra. Documentações técnicas, por exemplo, podem demandar uma maior ênfase na adequação, priorizando a transmissão precisa do significado.

Algumas **considerações sobre o uso de escalas Likert** na avaliação humana são importantes. Vale a pena ressaltar que as escalas Likert podem apresentar complexidades na sua aplicação. Diversos tipos podem ser utilizados, como escalas numéricas, de janela deslizante (em inglês, *sliding window*) e afirmações (concordo/discordo). Embora essas escalas sejam facilmente compreensíveis e quantificáveis, elas carregam uma natureza subjetiva, pois falham em medir as atitudes reais dos respondentes, levantando a questão: qual a diferença exata entre uma pontuação 5 e uma 4? (Um erro? Dois erros?). Além

disso, a presença de um número par de opções pode levar os participantes a escolherem o centro, demonstrando a delicadeza desse tipo de avaliação.

4.3.4.3 Ranqueamento: Hierarquia de Traduções

O **ranqueamento** (em inglês, *ranking*) tem como propósito classificar e comparar duas ou mais traduções, com o intuito de estabelecer uma **hierarquia de qualidade** entre elas. As comparações podem ser efetuadas tanto entre as traduções geradas por diferentes sistemas de TA quanto entre traduções geradas por humanos. Essa abordagem permite identificar **nuances de qualidade**, ressaltando as distinções entre as opções.

Além disso, o ranqueamento pode incorporar a **possibilidade de empates**, quando duas ou mais traduções são avaliadas como equivalentes em qualidade, sendo categorizadas como “igualmente boas” ou “igualmente ruins”. A categorização desses empates enriquece a análise, oferecendo insights sobre o grau de qualidade comparativa.

No âmbito de diagnósticos, esse tipo de classificação oferece a capacidade de indicar melhorias ao comparar o sistema avaliado com uma linha de base (*baseline*). Essa perspectiva não somente permite avaliar o progresso alcançado mas também identificar áreas específicas de aprimoramento, promovendo a constante evolução do sistema de TA.

Uma aplicação prática do ranqueamento é a seleção do sistema mais adequado para um projeto específico. A análise hierárquica das traduções permite a **escolha embasada no desempenho**, assegurando que a tradução atenda de forma eficiente aos requisitos e objetivos do projeto em questão.

O uso do ranqueamento apresenta suas **limitações** a serem consideradas. Ele **não oferece uma avaliação refinada**, não detalhando os erros presentes nas traduções. Quando empates não são permitidos, traduções igualmente boas ou ruins podem ser classificadas de maneira diferente, ressaltando uma inconsistência na hierarquia.

4.3.4.4 Anotação de Erros: Taxonomias

A anotação de erros se destaca como um método essencial para identificar e classificar imperfeições presentes em textos traduzidos. Diversas taxonomias foram propostas para essa finalidade, como Vilar et al. (2006), Font Llitjós et al. (2005), Federico et al. (2014), Costa et al. (2015), DQF de TAUS (O’Brien et al., 2011) e MQM (Lommel; Melby, 2018) by QT212 (Doherty et al., 2013). Para o português brasileiro, Martins (2014) e Martins; Caseli (2015) trazem a adaptação das categorias de erros de Popovic; Burchardt (2011) e Vilar et al. (2006) para traduções português-inglês.

As tipologias de erros frequentemente abrangem uma série de aspectos, incluindo palavras ausentes (em inglês, *missing words*), palavras adicionadas (em inglês, *added/extra words*), ordem errada das palavras (em inglês, *word order*), traduções literais (em inglês, *literal translation*), traduções erradas (em inglês, *mistranslation*), palavras incorretas (em inglês, *incorrect words*), formas inadequadas (em inglês, *incorrect form*), pontuação inadequada (em inglês, *punctuation*), entre outros, que podem incluir outras subcategorias específicas.

A adoção de taxonomias de erros na ATA oferece diversos **benefícios**, tais como identificar **tipos específicos** de erros nas saídas de TA, fornecer **relatórios detalhados de erros** para o aprimoramento dos sistemas, e fornecer informações aos clientes sobre a **qualidade da tradução**. Além disso, provedores de serviços linguísticos utilizam taxo-



nomias e avaliações de severidade³¹ para monitorar o trabalho de tradutores. A anotação de erro também ajuda a investigar as **relações entre tipos específicos de erros e as preferências de usuários ou pós-editores**, bem como avaliar o **impacto** de diferentes tipos de erros em várias etapas do processo de pós-edição.

Contudo, entre as principais **limitações** dessa estratégia, destaca-se que a anotação manual de erros é um **processo caro e demorado**, demandando um investimento significativo de tempo. Além disso, essa avaliação **nem sempre é uma tarefa simples**, especialmente quando se trata de diferenciar entre categorias como “tradução literal” e “tradução errada”.

Outra complexidade está associada à **dependência da língua**. Diferentes idiomas possuem particularidades que podem tornar a identificação e a classificação de erros uma tarefa mais desafiadora. Também é relevante considerar que a eficácia da anotação de erros pode variar de acordo com o **tipo de abordagem de TA**: enquanto ela pode ser mais adequada para sistemas de tradução baseados em regras (RBMT), pode não ser tão precisa para sistemas de tradução estatística (SMT) ou de tradução neural (NMT). Nesse contexto, a seleção da abordagem de avaliação mais adequada torna-se um ponto de reflexão. Além disso, a **falta de consenso entre avaliadores** é uma questão importante, frequentemente requerendo treinamento e prática para alcançar um nível satisfatório de concordância (Capítulo **Conjunto de dados, dataset e corpus**).

4.3.4.5 Pós-Edição na Avaliação de TA

A pós-edição (PE) (do inglês, *post-editing*) é definida como “a correção da saída da tradução automática bruta por um tradutor humano, de acordo com instruções e critérios de qualidade específicos” (O’Brien, 2011, p. 197). A PE emerge como uma ferramenta fundamental na avaliação de sistemas de TA, oferecendo uma perspectiva mais aprofundada sobre o esforço envolvido nesse processo. A medição desse esforço pode ser abordada de diferentes perspectivas (Krings, 2001), proporcionando uma visão mais abrangente do desempenho do sistema e do impacto da TA no fluxo de trabalho.

- **Esforço Temporal:** mede o ritmo de pós-edição. Avaliando o tempo gasto na pós-edição por palavras por segundos é possível compreender a velocidade desse processo. Nesse contexto, uma eficiência temporal maior, ou seja, menos tempo gasto na pós-edição, pode indicar uma melhor qualidade da saída da TA, influenciando a produtividade.
- **Esforço Técnico:** mede o número de operações de edição realizadas, como inserções (*insertions*), remoções (*deletions*), e trocas (*shifts*). Nesse sentido, a métrica hTER (Seção 4.3.3) é frequentemente utilizada como uma estimativa aproximada do esforço técnico. Uma menor quantidade de edições necessárias está diretamente correlacionada a uma melhor qualidade da TA, uma vez que está ligada ao tempo de esforço e, conseqüentemente, à produtividade.
- **Esforço Cognitivo:** pode ser medido por meio de diferentes abordagens, incluindo o rastreamento ocular (*eye-tracking*). A redução desse esforço cognitivo durante o processo de pós-edição é indicativa de uma qualidade superior da TA, e tal esforço tem sido correlacionado a outras métricas de avaliação humana.

³¹As severidades geralmente são classificadas como “Crítico” (*critical*), “Grave” (*major*), e “Mínimo” (*minor*).



A utilização da PE na avaliação da TA é motivada por diversos fatores. Além de avaliar a utilidade do sistema de TA em produção, ela **permite identificar erros comuns e gerar novos dados de treinamento ou teste**. Contudo, é importante ressaltar que as medidas de esforço de PE **tendem a variar entre avaliadores** novatos (estudantes) e profissionais, bem como entre o público em geral e profissionais experientes.

Alguns trabalhos com pós-edição com o português incluem: De Sousa et al. (2011); Almeida (2013), Castilho et al. (2014), Moorkens et al. (2015), Castilho et al. (2017a), Silva et al. (2017a), Castilho et al. (2019b), Castilho; Resende (2022). Há também os trabalhos que investigaram a automatização do processo de pós-edição para o português: Caseli; Inácio (2020).

4.3.4.6 Considerações Finais sobre Avaliação Humana na Avaliação de TA

Ao explorar a avaliação humana na avaliação de TA, encontramos uma série de questões metodológicas e pragmáticas que merecem reflexão:

- Precisamos sempre avaliar tanto a adequação quanto à fluência?
- Quantos avaliadores são necessários e quais as competências linguísticas eles devem possuir?
- Devemos envolver tradutores, linguistas ou especialistas no domínio? O viés em cada escolha é um ponto a ser considerado.
- Quantos pontos devem ser incluídos em uma escala Likert de avaliação?
- Qual o grau de concordância entre avaliadores (Capítulo **Conjunto de dados, dataset e corpus**) e a consistência nas avaliações individuais?

Considerações pragmáticas também emergem, incluindo o custo associado aos avaliadores, à geração de textos de referência traduzidos por humanos, e à qualidade dessas referências. O tempo investido, a baixa concordância intra e inter-avaliadores, e a questão de saber se o objetivo da avaliação é apenas avaliar melhorias em um sistema são fatores preponderantes.

4.3.5 Avaliação dependente de contexto

Como vimos nas seções anteriores, a avaliação de TA começou desde o princípio da área com projetos como DARPA, que avaliaram a qualidade das traduções com métricas humanas. À medida que a área avançou, métricas automáticas foram gradualmente incorporadas, inicialmente provenientes de outros domínios do PLN, e posteriormente desenvolvidas especificamente para o campo da TA.

E apesar de ambas as MAAs e as métricas humanas serem o estado da arte na avaliação, à medida que os sistemas neurais de TA evoluíram, tornando-se mais complexos e produzindo traduções de maior qualidade, surgiu a necessidade de uma **avaliação mais abrangente e rigorosa** que levasse em conta fatores diversos. Adicionalmente, com o aumento nos esforços direcionados à incorporação de contexto nos sistemas de TA neurais, viu-se a necessidade de também se ter uma avaliação com contexto, uma vez que, os resultados obtidos na avaliação com sentenças eram limitados, pois ela não é capaz de identificar as melhorias desses sistemas (Läubli et al., 2018; Toral et al., 2018). Ademais, as MAAs subestimam a qualidade dos sistemas NMT (Shterionov et al., 2018), e a credibilidade dessas métricas para sistemas em nível de documento também tem sido objeto de críticas (Smith, 2017). Diante disso, surgiu a necessidade de avaliar a TA considerando



um contexto mais amplo, possibilitando uma análise mais abrangente do contexto em questão. Entretanto, a metodologia para essa avaliação ainda está em sua fase inicial e poucos estudos foram realizados nesse sentido.

Em 2018, alegações de “paridade humana” (em inglês, *human parity*) na qualidade da TA (Hassan et al., 2018) foram rebatidas por Toral et al. (2018) e Läubli et al. (2018), os quais apontaram que essa paridade não se replicava quando se considerava o **contexto** ou outros fatores, como a direção da tradução ou a experiência do anotador.

A Conferência de Tradução Automática (WMT), realizada desde 2006 e que até o ano de 2019 restringiu suas avaliações à análise de frases individuais, começou sua primeira tentativa de conduzir avaliações humanas em nível de documento no domínio de notícias no ano de 2019 (Barrault et al., 2019), em resposta às críticas apresentadas por Toral et al. (2018) e Läubli et al. (2018). Adotando uma abordagem direta de avaliação (Graham et al., 2016), a conferência solicitou que avaliadores de multidão³² atribuísem uma pontuação (de 0 a 100) a cada sentença. Os avaliadores foram instruídos a avaliar: (i) textos completos, (ii) segmentos individuais consecutivos na ordem original, e (iii) frases individuais selecionadas aleatoriamente. No ano subsequente, na edição WMT20, houve uma mudança de abordagem, expandindo o âmbito de avaliação para abranger artigos completos, demandando dos avaliadores a análise de segmentos específicos enquanto visualizavam o documento completo, bem como a avaliação da tradução do conteúdo (Barrault et al., 2020).

Castilho (2020) e Castilho (2021) testou as diferenças na concordância inter-anotadores (CIA) entre duas metodologias de avaliação: (i) uma centrada em sentenças individuais e (ii) outra com contexto, para o português brasileiro. No estudo de Castilho (2020), tradutores avaliaram a saída de TA considerando critérios de fluência, adequação (usando uma escala Likert), ranqueamento e anotação de erros. Essa avaliação foi conduzida em duas configurações distintas onde os tradutores atribuíram: (i) uma pontuação para cada sentença isolada, e (ii) uma pontuação para o documento como um todo. Os resultados demonstraram que os níveis de CIA para a metodologia em nível de documento atingiram níveis negativos, enquanto a satisfação dos tradutores com essa metodologia foi bastante reduzida. No entanto, esse enfoque evitou situações de avaliação incorreta (*misevaluation*) que são recorrentes quando se analisam sentenças isoladamente.

Continuando esse trabalho, Castilho (2021) modifica a configuração em nível de documento e repete o experimento com mais tradutores, onde ela compara a CIA na avaliação de (i) sentenças únicas aleatórias, (ii) avaliação de sentenças individuais em que os tradutores têm acesso à fonte completa e à saída de TA, e (iii) avaliação de documentos completos. Os resultados mostraram que uma metodologia em que os tradutores avaliam sentenças individuais no contexto de um documento gera um bom nível de CIA em comparação com a metodologia de sentença única aleatória, enquanto uma metodologia em que os tradutores atribuem uma pontuação por documento mostra um nível muito baixo de CIA. A autora afirma que atribuir uma nota por sentença no contexto evita casos de avaliação incorreta que são extremamente comuns nas configurações de avaliação de frases aleatórias. Além disso, a autora postula que o maior acordo de CIA na configuração de sentença única aleatória ocorre porque “os avaliadores tendem a aceitar a tradução quando a adequação é ambígua, mas a tradução está correta, especialmente se for fluente” (Castilho, 2021, p. 42), e afirma que **o método de avaliação de sentença única aleatória deve ser evitado**, pois o problema de avaliação incorreta é especialmente problemático ao avaliar a qualidade

³²Avaliadores contratados via plataformas como [Mechanical Turk](#).



de sistemas NMT, uma vez que eles apresentam um nível aprimorado de fluência.

Após isso, em Castilho (2022) foi demonstrado que o contexto necessário para resolver questões de avaliação é influenciado pelo domínio, sem parecer estar intrinsecamente ligado ao comprimento das sentenças presentes no *corpus* envolvendo os idiomas inglês, português, irlandês, chinês e alemão. Em consequência disso, a pesquisa de Castilho et al. (2023) revelou que o impacto da extensão do contexto não parece influenciar significativamente os resultados, porém a estruturação da pontuação desempenha um papel crucial. Isso se deve ao fato de que sentenças conectadas tendem a gerar resultados mais diversos, com abordagens mais acuradas para resolver ambiguidades lexicais quando comparadas aos cenários de pontuação normais. Além disso, o estudo apontou que os sistemas GPT demonstraram proporcionar traduções mais precisas do que os sistemas de Tradução Automática.

Diante desse panorama, a avaliação de TA com contexto encontra-se em sua infância, com diversas questões em aberto. O futuro da avaliação de TA deve considerar se as métricas automatizadas e as avaliações humanas atuais conseguem capturar de forma realista a qualidade dos sistemas de nível de documento, e se é necessário modificar ou criar novas abordagens. Especial atenção deve ser dada aos modelos de linguagem como o GPT, conhecidos por gerar traduções fluentes e coesas, uma vez que a avaliação deve incorporar precisão da informação, fidelidade ao conteúdo original e coerência global, evitando a introdução de informações imprecisas ou divergentes.

Ademais, a avaliação de documentos traduzidos não deve se limitar a métricas automáticas, mas também usar a avaliação humana. Os avaliadores humanos desempenham um papel crucial em identificar nuances de qualidade que as métricas automáticas podem não capturar, como aspectos culturais, ambiguidades e sutilezas linguísticas. Portanto, a **combinação de métricas automáticas com avaliações humanas** se mostra uma abordagem fundamental para obter uma compreensão abrangente da qualidade da tradução de documentos.

4.4 O Futuro da Tradução Automática

O futuro da TA parece muito promissor. Com a globalização e a internet, mais conteúdo é criado todos os dias, e, portanto, estão surgindo cada vez mais casos de uso nos quais a TA pode ser útil (Way, 2018). Segundo a “Slator 2019 Language Industry Market Report” (2019, p. 14), “a TA está bem encaminhada para se tornar a tecnologia mais importante para aprimorar a produtividade dos tradutores humanos”.

O aumento impressionante na qualidade com o surgimento da TA neural (NMT) em comparação com seu antecessor, o PSMT, foi exagerado pela mídia (Läubli et al., 2018; Toral et al., 2018), mas é incontestável que a NMT tenha sido, de fato, uma mudança de paradigma na área. No entanto, o entusiasmo em torno da TA diminuiu, com empresas de tradução de grande e médio porte relatando que, embora o uso da TA tenha aumentado, os benefícios percebidos têm se estabilizado (Sarah Hickey, 2020) em termos de grandes avanços na qualidade.

No entanto, com o aumento da qualidade, é possível abordar uma variedade maior de tipos de documentos e públicos. Isso significa que há muito espaço para **personalização de sistemas de TA** projetados para casos de uso e contextos específicos, melhorando a precisão. Para a TAUS (2020, p. 16), a NMT será “aplicada de forma útil em ambientes de tradução de fala” e, na verdade, em todo discurso falado, já que lida melhor com conteúdo gerado pelo usuário. Além disso, “a NMT ajudará na expansão adicional de tecnologias de



tradução de fala [...] disponíveis principalmente como sistemas monolíngues baseados em inglês, [...] transformando-os em sistemas multilíngues”, o que implicará “muitas mudanças profundas e caras”.

Mais recentemente, no fim de 2022, os modelos de linguagem em larga escala (Capítulo *Modelos de linguagem*), como o GPT-3 da OpenAI³³, têm desempenhado um papel importante no campo da tradução automática e prometem desempenhar um papel ainda maior no futuro. Esses modelos surgiram com o avanço das redes neurais e do aprendizado profundo. Desde sua introdução, os LLMs têm sido amplamente utilizados na TA, proporcionando melhorias significativas na qualidade e na fluidez das traduções geradas. Eles têm sido capazes de lidar com nuances linguísticas, contexto e ambiguidades, resultando em traduções mais precisas e naturais. Com o contínuo avanço da tecnologia, espera-se que os LLMs sejam capazes de melhorar a personalização das traduções, adaptando-se a estilos de escrita específicos e preferências individuais.

No entanto, embora os LLMs tenham apresentado avanços significativos na área da TA, ainda existem **desafios** a serem superados. A qualidade da tradução depende de vários fatores, como a disponibilidade de dados de treinamento de alta qualidade e a compreensão do contexto e nuances linguísticas. Além disso, os LLMs podem ser sensíveis a preconceitos (bias) presentes nos dados de treinamento, resultando em traduções imprecisas ou enviesadas.

Segundo a pesquisa da CSA, “a pós-edição como serviço diminuirá ao longo do tempo, sendo substituída pela tradução automática adaptativa em software de tradução mais dinâmico” (p. 23), e haverá uma demanda crescente por linguistas profissionais que possam interagir com a saída da tradução automática “Slator 2019 Language Industry Market Report” (2019, p. 22). Além disso, o relatório da Slator afirma que agora, com os altos níveis de qualidade e a ampla disponibilidade de ferramentas gratuitas de tradução automática, os clientes corporativos esperam mais do que uma tradução automática “apenas boa” e estão buscando “soluções personalizadas, adaptadas ao seu conteúdo, que possam ser adaptadas para seus fluxos de trabalho e preferências estilísticas específicas” “Slator 2019 Language Industry Market Report” (2019, p. 22).

Vale ressaltar que todos os relatórios afirmam que a maioria da indústria “ainda não espera que a qualidade da tradução automática atinja os níveis da tradução humana em um futuro próximo” TAUS (2020, p. 16), e, portanto, tanto a tradução humana quanto a interação humana com a tradução automática ainda são altamente demandadas.

Como podemos ver, os sistemas de TA estão atingindo níveis de qualidade significativamente altos e, por isso, estão sendo cada vez mais utilizados em diversas áreas de negócio. Com a TA se tornando ubíqua em nosso dia a dia, a necessidade de testar a qualidade desses sistemas se tornou essencial (Castilho et al., 2019a). Há muito espaço para a TA melhorar, e, portanto, uma boa prática na avaliação da TA é essencial para evitar afirmações exageradas e fornecer aos usuários um feedback honesto.

³³<https://openai.com/>



Capítulo 5

Sumarização Automática

Jackson Wilke da Cruz Souza
Paula Christina Figueira Cardoso
Crysttian Arantes Paixão
Pedro Henrique Paiola
Gabriel Lino Garcia
João Vitor Mariano Correia
João Renato Ribeiro Manesco
Danilo Samuel Jodas
Douglas Rodrigues
João Paulo Papa

Publicado em: 13/03/2024

Atualizado em: 16/04/2026

 <https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/>

5.1 Introdução

Talvez resumir seja uma das tarefas mais comuns em nosso cotidiano. Não são raras as vezes que nos pegamos reportando algo a alguém: quando compartilhamos sobre o último filme que assistimos, ou o livro que lemos, ou quem sabe quando contamos para alguém sobre aquela história ou notícia que vimos ou vivenciamos.

Nessas tarefas, partimos do princípio que não conseguiremos reproduzir exatamente o que aconteceu, mas faremos o máximo de esforço para transmitir uma mensagem mais próxima do original, valorizando o **conteúdo** mesmo em detrimento da **forma**. Assim, por exemplo, podemos optar por reestruturar o fato acontecido ou vivenciado dentro da narrativa sintetizada: selecionamos o que julgamos ser o mais importante do evento; lançamos mão de inversões dos fatos, prevendo melhorar a compreensão do ouvinte/leitor da narrativa; ou ainda trocamos as palavras e construções frasais originais por outras semelhantes ou sinônimas.

Outra consideração importante nesse processo é levar em conta o tempo ou limite (textual) que temos à disposição para a construção da narrativa que foi resumida. Saber disso de antemão ditará o quão detalhista ou generalista precisaremos ser na síntese dos fatos, tendo em vista que quanto mais tempo e/ou espaço tivermos, mais próximo à realidade estará a narrativa sintetizada.

Ao chegar nesse ponto do nosso exemplo é necessário nos questionar sobre algo bastante relevante: o quanto das percepções pessoais e autorais ficam na narrativa quando comparada ao evento real? O fato de haver liberdade estilística na produção da síntese, não significa que a narrativa deva veicular conteúdo falso, já que ela precisa representar o evento real. Apesar disso, é possível, em alguns casos, acrescentar e evidenciar nossas avaliações e perspectivas. O esforço do autor é pensar em diferentes operações textuais



que incidem sobre o efeito de sentido de verossimilhança ao conteúdo, como a paráfrase ou ainda a cópia.

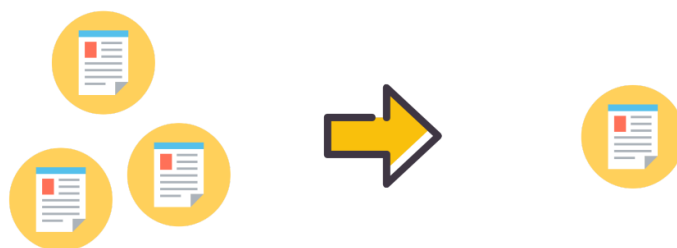
Do ponto de vista linguístico, os estudos sobre as operações associadas à transformação de texto-fonte em um texto-alvo sintetizado remetem ao conceito de **reescrita**. Para Fiad (2013), esse conceito é tido como o trabalho que é feito por um autor (com ou sem auxílio de um mediador) por meio de operações na linguagem que abrangem diferentes aspectos do texto conduzindo-o à sua modificação. Assim, podemos dizer que o processo de resumir algo está compreendido na reescrita.

Já do ponto de vista computacional, essa atividade que nos é tão comum e feita de maneira, por vezes, intuitiva, pode ser realizada automaticamente. Em processamento de linguagem natural (PLN), essa tarefa é conhecida como **sumarização automática** (do-ravante, SA). A SA pode ser definida como a produção automatizada de versões reduzidas de outros textos, resultando em sumários (ou “resumos”). Rino; Pardo (2003) elucidam algumas premissas para a SA, a saber:

1. Existência de um texto que deve ser condensado, tido como *texto-fonte*;
2. O texto-fonte dispõe de (i) uma ideia ou tópico principal, (ii) um conjunto de informações relacionadas entre si, (iii) um propósito comunicacional que organiza e relaciona as informações entre si e (iv) uma narrativa visando a informatividade de maneira coerente;
3. Identificação do conteúdo relevante no texto-fonte e construção de uma nova narrativa coerente no sumário, preservando a ideia inicial.

A partir dessas premissas, apresentamos um modelo de representação do processo de SA, na Figura 5.1.

Figura 5.1: Processo genérico de Sumarização Automática



Na Figura 5.1, a SA se inicia na consulta de um ou mais textos-fonte, que será(ão) submetido(s) a determinada análise para fazer a verificação do que é mais ou menos importante. Quando o processo de sumarização se baseia em apenas um texto-fonte, diz-se que a sumarização é **monodocumento**; quando baseado em dois ou mais textos-fonte, **multidocumento**. Como resultado, é elaborado o sumário, caracterizado por ser uma versão menor do(s) texto(s)-fonte. Mais adiante, detalharemos esses e outros aspectos e conceitos importantes à SA, como quantidade de textos-fonte a resumir e de línguas nesse processo, além de tendências e perspectivas metodológicas.

Com relação às tendências de **abordagens da SA**, ao longo do tempo, percebemos que as pesquisas em SA do começo dos anos 70 até o início da década de 90 tinham como motivação a recuperação de informação em textos. Por conta disso, os métodos utilizados para a sumarização nesse período se baseiam quase que exclusivamente em informações na

superfície textual, como a ocorrência de palavras-chave no título e no corpo do texto. No começo dos anos 90, percebeu-se que os métodos utilizados anteriormente não conseguiam dar conta de problemas linguísticos mais complexos, como resolução de correferência (Capítulo [Resolução de correferência](#)). Assim, métodos linguisticamente motivados começaram a ser implementados em sistemas de SA.

Já nas duas primeiras décadas dos anos 2000, notou-se uma tendência sócio-comportamental que também influenciou a SA: a democratização da Web. Por conta disso, nesse período, a sumarização se popularizou devido ao fato de haver um grande volume de informação produzida e circulante em relação ao pouco tempo que as pessoas tinham disponível para consumi-la. Mais recentemente, vimos a influência dos *Large Language Models* (LLMs) (Capítulo [Modelos de linguagem](#)) em PLN. Com relação à SA, a implementação desse tipo de abordagem apontou que é possível produzir sumários mais coerentes, coesos, mesmo em contextos que exigem manobras de reescrita dos textos. Em PLN há diferentes motivações que fomentam interseções com a SA. Diversas aplicações e subáreas, como compreensão de textos, recuperação da informação, indexação, mineração de textos e sistemas de pergunta-resposta, partem de texto(s)-fonte que necessita(m) ter seu(s) conteúdo(s) classificado(s) entre mais e menos importante e, a partir disso, propor uma nova versão do texto, ou em um tamanho menor ou um outro conteúdo. Além disso, os sumários também podem beneficiar pessoas que precisam ler biografias ou coleções de documentos em que seja necessário ter acesso à informação de maneira resumida.

Ao longo deste capítulo queremos introduzir conceitos importantes para a SA, bem como propor um material que você possa utilizar na construção de seus primeiros sumários. Além disso, nesta nova edição do capítulo você contará também com o acréscimo de novas reflexões sobre a utilização de LLMs na SA, sobretudo na abstrativa. Você irá perceber que em alguns momentos remontamos a história da área, pois entender as limitações que sistemas e métodos apresentaram ao longo da trajetória da SA nos permite entender quais os desafios atuais impostos ora pelas tecnologias desenvolvidas, ora pela própria língua em relação aos sistemas computacionais. Ainda, apresentamos exemplos de algoritmos introdutórios e simples para que você possa praticar a sumarização e, quem sabe, poder aplicá-la em outros cenários.

5.2 Por trás do processo

5.2.1 Sobre conceitos

Antes de discutir métodos e sistemas de geração e avaliação da SA, é necessário conhecer alguns conceitos importantes. Para tanto, destacamos as seguintes características da SA: (i) função, (ii) público-alvo, (iii) tipo, (iv) abordagem, (v) fonte, e (vi) idioma.

A versão sumarizada do texto-fonte é conhecida como sumário, que pode ser classificado como informativo, indicativo ou crítico, a depender da **função** comunicativa que exerce (Mani; Maybury, 1999). Os sumários informativos são aqueles que apresentam a informação original de maneira a preservá-la a ponto de dispensar a leitura do texto-fonte. Os sumários indicativos são orientações genéricas sobre o texto original, fazendo com que o leitor tenha um panorama sobre o conteúdo, sem ter acesso a seus detalhes, como o índice de um livro. Por fim, os sumários críticos são aqueles que permeiam entre a síntese e a avaliação do conteúdo, aproximando-se do gênero resenha.

Quando pensamos no **público-alvo**, os sumários podem ser caracterizados em genéricos ou específicos. Sumários genéricos são aqueles que extraem a informação do texto-fonte



levando em consideração apenas critérios técnicos, como importância do conteúdo. Já os sumários específicos, além de considerarem aspectos técnicos também devem levar em conta elementos relacionados ao leitor que afetam a construção do sumário, como o conhecimento prévio ou o interesse do público-alvo sobre o assunto do texto-fonte.

No que se refere ao **tipo**, os sumários podem ser classificados em extrativos e abstrativos (do inglês *abstracts*). Os sumários extrativos têm a característica de serem compostos por sentenças dos textos-fonte. Nesse sentido, o conteúdo não é submetido a nenhum tipo de edição linguística, mas apenas selecionado a partir de determinados critérios (como tamanho disponível do sumário e a importância do conteúdo, por exemplo). Já os sumários abstrativos são caracterizados por seu conteúdo (parcial ou integral) ter sido submetido a alguma operação linguística de reescrita, como a paráfrase.

Com relação à **abordagem**, em PLN há três grandes vertentes, organizadas em função da quantidade e profundidade de informação linguística utilizada no processo de sumarização. A abordagem profunda utiliza muito conhecimento linguístico na elaboração dos sumários. Nesse caso, podem ser levados em conta aspectos pertinentes à semântica, ao discurso e à pragmática, por exemplo, para tratar determinados fenômenos linguísticos. Já a abordagem superficial utiliza pouco conhecimento linguístico; quando o faz, utiliza majoritariamente informações dos níveis morfológico, morfossintático e sintático. Esse tipo de abordagem se baseia em conhecimento estatístico e empírico. Por fim, a abordagem híbrida seria a combinação das duas abordagens anteriores.

É importante destacar que há possibilidade de elaborar diferentes formatos de sumários a partir da abordagem escolhida. Os sumários produzidos sob a abordagem superficial tendem a ser extrativos, já que utilizam, na maioria das vezes, técnicas de contagem de *tokens* e/ou identificação de informações que estão na superfície textual. Ao passo que abordagens profundas e híbridas tendem a produzir sumários abstrativos, pois demonstram estratégias linguísticas de retextualização que podem resultar em reelaboração da forma com que o conteúdo está expresso.

Quanto à **fonte**, os sumários podem ser classificados em mono ou multidocumento. Os sumários monodocumento são resultantes apenas de um único texto que serviu de base para a sumarização. No caso dos sumários elaborados a partir de dois ou mais textos que dissertam sobre o mesmo assunto, tem-se a SA multidocumento. Este último tipo de sumarização, em especial, precisa lidar com fenômenos linguísticos que tornam o processo mais desafiador, como “redundância”, “complementaridade” e “contradição” provenientes de fontes produzidas com estilos, vieses e perspectivas autorais diferentes durante a escrita dos textos¹.

Quanto ao **idioma**, os sumários podem ser produzidos a partir de uma única língua (monolíngue) ou com duas ou mais línguas (multilíngue).

Outro conceito importante na SA é a **taxa de compressão** dos sumários, isto é, quanto de informação será incluída no sumário. Essa quantidade é estabelecida em função da taxa de compressão, que é a razão entre o tamanho do sumário e o tamanho do texto-fonte (Mani, 2001). Um sumário com taxa de compressão de 70% apresenta tamanho equivalente a 30% do tamanho do texto-fonte, medido ou em função da quantidade de palavras ou de sentenças. Assim, se o texto-fonte possuir 1000 palavras e deseja-se construir um sumário com uma taxa de compressão de 0,7, o sumário terá 300 palavras. No caso da SA multidocumento, a taxa de compressão pode ser definida em relação ao maior texto ou à soma total de palavras ou de sentenças de todos os textos. Porém, em um cenário de

¹Sugerimos a leitura de (Souza et al., 2011a), (Souza; Felippo, 2018), (Souza, 2019) e (Silva; Di Felippo, 2014) que se aprofundam nos referidos fenômenos linguísticos.



sumarização abstrativa, é preciso ponderar que nem sempre os valores estabelecidos nas taxas de compressão farão segmentações de sentenças de maneira devida. Nesse sentido, com relação à informatividade, é preciso avaliar o conteúdo que já foi escolhido para compor o sumário, caso a taxa de compressão pretendida exceda a quantidade de palavras estimada para o sumário.

5.2.2 Sobre o processo de sumarização

A seguir, apresentamos uma possível análise de um sumário retirado do *corpus* CSTNews² (Cardoso et al., 2011). Os textos do Quadro 5.1 serviram de fonte de informação para a elaboração de sumários extrativos multidocumento do Quadro 5.2.

Quadro 5.1: Textos-fonte extraídos do *corpus* CSTNews

Texto-fonte 1	Texto-fonte 2
O médico pessoal do argentino Diego Maradona, Alfredo Cahe, revelou nesta segunda-feira que uma recaída da hepatite aguda de que sofre foi o motivo da nova internação do ex-craque. Maradona havia recebido alta no último dia 11, mas voltou a ser internado na sexta-feira e os boletins médicos não especificaram o que se passava com o ex-jogador –Cahe descartou pancreatite ou úlcera. “Maradona teve uma recaída na hepatite aguda. Agora está estável. Apesar de ter melhorado no domingo, deverá continuar internado”, disse Cahe, em declarações ao jornal “La Nación”. Maradona, 46, desenvolveu um hepatite tóxica por excesso de consumo de álcool, o que já o manteve internado durante 13 dias antes da primeira alta. Cahe disse ainda que Maradona não voltou a consumir bebidas alcoólicas e que as causas da recaída estão sendo investigadas.	BUENOS AIRES - Maradona voltou a ter problemas de saúde no fim de semana. Internado em um hospital em Buenos Aires, ele teve uma recaída e voltou a sentir dores devido a hepatite aguda que o atinge, segundo seu médico pessoal, Alfredo Cahe. “Agora está estável. Mesmo com esta melhora, ele continuará internado”, disse o médico, que descartou a possibilidade do ex-jogador ter uma pancreatite (inflamação do pâncreas, órgão situado atrás do estômago e que influencia na digestão). Cahe reforçou que Maradona ainda tem problemas. “Os valores hepáticos dele na avaliação não estão equilibrados e ele não está bem. Mas não é nada grave”, afirma, em entrevista ao diário La Nación. No domingo, Maradona assistiu ao empate por 1 a 1 no clássico Boca Juniors e River Plate pela televisão. Os torcedores do Boca, que compareceram em grande número ao Estádio La Bombonera, levaram muitas faixas e bandeiras com mensagens de apoio ao ídolo argentino. Sua filha, Dalma, foi ao estádio assistir ao jogo.

Os textos resumidos apresentam uma taxa de compressão de 70% do total de palavras do conjunto de textos-fonte. Assim, o humano poderia selecionar sentenças inteiras para compor seu sumário; caso a próxima sentença ultrapassasse muito o tamanho pré-estabelecido, permitia-se ter um sumário com menos palavras³.

²O *corpus* CSTNews é um repositório de textos jornalísticos publicados no ano de 2007, originalmente escritos em português. Ele está organizado em 50 conjuntos de textos, em que cada conjunto possui de dois a três textos sobre um mesmo assunto. Cada conjunto de textos inclui anotações linguísticas (como morfosintática, semântica e discursiva, por exemplo), além de apresentar sumários humanos e automáticos. Esse repositório serviu de base para muitas pesquisas em SA em língua portuguesa, e está disponível em: <http://nilc.icmc.usp.br/CSTNews/>.

³Para saber mais detalhes sobre a criação de mais sumários humanos para o *corpus* CSTNews, consulte

Quadro 5.2: Sumários extraídos do *corpus* CSTNews

Sumário humano multidocumento	Sumário automático multidocumento
BUENOS AIRES - Maradona voltou a ter problemas de saúde no fim de semana. <Texto-fonte 2> Internado em um hospital em Buenos Aires, ele teve uma recaída e voltou a sentir dores devido a hepatite aguda que o atinge, segundo seu médico pessoal, Alfredo Cahe. <Texto-fonte 2> Agora está estável. Apesar de ter melhorado no domingo, deverá continuar internado”, disse Cahe, em declarações ao jornal “La Nación”. <Texto-fonte 1> Maradona, 46, desenvolveu um hepatite tóxica por excesso de consumo de álcool, o que já o manteve internado durante 13 dias antes da primeira alta. <Texto-fonte 1> Cahe disse ainda que Maradona não voltou a consumir bebidas alcoólicas e que as causas da recaída estão sendo investigadas. <Texto-fonte 1>	Internado em um hospital em Buenos Aires, ele teve uma recaída e voltou a sentir dores devido a hepatite aguda que o atinge, segundo seu médico pessoal, Alfredo Cahe. <Texto-fonte 2> “Maradona teve uma recaída na hepatite aguda. Agora está estável. Apesar de ter melhorado no domingo, deverá continuar internado”, disse Cahe, em declarações ao jornal “La Nación”. <Texto-fonte 1> Cahe disse ainda que Maradona não voltou a consumir bebidas alcoólicas e que as causas da recaída estão sendo investigadas. <Texto-fonte 1>

Os textos-fonte do Quadro 5.1 noticiam o estado de saúde do ex-jogador de futebol Maradona e sobre seu processo de internação para tratar de um quadro de hepatite aguda. Apesar de serem textos jornalísticos, percebe-se que eles apresentam estilos de escrita diferentes, como o uso mais extensivo do discurso direto; ou ainda a construção narrativa dando ora mais detalhes, ora menos.

Os textos dos sumários extrativos também apresentam particularidades entre os sumarizadores: das 5 sentenças escolhidas pelo humano, 3 foram extraídas do Texto-fonte 1, admitindo que ele pode ter preferência por alguma fonte que julgue ser de sua confiança; ao passo que o sistema automático seleciona apenas 1 advinda do Texto-fonte 2, das 3 sentenças escolhidas. Diante disso, é possível inferir que o primeiro texto-fonte apresenta o conteúdo de maneira a ganhar a predileção de ambos os sumarizadores, seja pelo estilo, composição ou qualidade da informação.

Apesar de ambos os sumarizadores terem escolhido sentenças em comum, algo a se destacar é o fato de os textos sumarizados apresentarem estruturas narrativas discretamente diferentes. No sumário automático, a sentença inicial é apenas a segunda do texto sumariado por um humano. Assim, é possível inferir que na sumarização manual elegeu-se uma sentença para abrir o texto que pudesse introduzir melhor o assunto, antes de chegar ao cerne da informação.

Também é possível identificar que os textos sumarizados, em relação aos textos-fonte, mantêm uma ordenação da informação: o início do texto apresenta informações mais importantes em relação ao seu desfecho. Nesse caso, os sumarizadores precisaram de alguma forma identificar as informações mais relevantes dos textos-fonte e, em seguida, refletirem como esse conteúdo se organizaria com relação ao tempo e à progressão temática no próprio sumário.

A identificação de conteúdo relevante não é uma tarefa trivial. Desde o advento da SA, um dos maiores desafios para as pesquisas foi desenvolver maneiras de identificar a

(Dias et al., 2014).



informação tida como a mais relevante no(s) texto(s)-fonte. Conforme foram superados alguns desafios, outros foram identificados e passaram a pautar muitos estudos, como a SA multilíngue, de atualização (update summarization) e abstrativa.

5.3 Avaliação na SA

Usualmente, a avaliação de desempenho dos sistemas de SA é medida por meio da análise de seus sumários, considerando os critérios de **informatividade** e **qualidade** dos textos produzidos. A informatividade geralmente é calculada de forma automática e consiste em verificar quanto da informação relevante dos textos-fonte é preservada no sumário automático. A avaliação da qualidade, por sua vez, é tradicionalmente realizada por humanos, pois o foco reside na análise de aspectos relacionados à gramaticalidade, coesão e coerência, foco e clareza referencial – elementos para os quais ainda não há métodos automáticos plenamente eficazes de avaliação. Ainda, ao final deste capítulo discutiremos brevemente os avanços que vêm sendo obtidos por meio dos LLMs.

Para medir a informatividade, a métrica de avaliação amplamente utilizada é a ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) (Lin, 2004). Essa medida compara automaticamente a quantidade de n-gramas (conjunto de palavras em sequência) em comum entre um sumário automático e um ou mais sumários de referência, ou seja, produzido por um humano. Porém, em cenários em que não houver sumários de referência, o texto-fonte pode ser adotado como critério para avaliação (Louis; Nenkova, 2013). A ROUGE possui quatro variantes principais, descritas a seguir:

- a) ROUGE-N (*ROUGE-Ngram*): Avalia a sobreposição de n-gramas entre o sumário automático e o sumário de referência. A ROUGE-1, por exemplo, considera a sobreposição de unigramas (palavras individuais), enquanto o ROUGE-2 considera a sobreposição de bigramas (pares de palavras consecutivas), e assim por diante.
- b) ROUGE-L (*ROUGE-Longest Common Subsequence*): Mede a similaridade baseada na sequência de palavras comuns mais longa entre o sumário automático e o texto de referência. A ROUGE-L não exige correspondência consecutiva, mas sim a mesma ordem das palavras tanto no sumário automático quanto no sumário de referência.
- c) ROUGE-W (*ROUGE-Word*): Calcula as correspondências consecutivas mais longas entre o sumário automático e o sumário de referência.
- d) ROUGE-S (*ROUGE-Skip-gram*): Essa métrica considera pares de palavras com um certo número de palavras intermediárias “puladas”. Ela é útil para capturar dependências de longa distância entre palavras.

O resultado ROUGE é dado em termos de precisão (Equação 5.1), revocação (Equação 5.2) e medida-f (Equação 5.3), e possui correlação com a avaliação humana (Lin, 2004). A precisão (P) expressa a proporção de n-gramas coincidentes entre os sumários automático e de referência em relação ao número de n-gramas do sumário automático. A revocação (R) representa a proporção de n-gramas coincidentes entre os sumários automático e de referência em relação ao número de n-gramas do sumário de referência. Tais medidas são complementares e, por isso, costuma-se utilizar a medida-f (F) que representa a média harmônica entre precisão e revocação. Como precisão e revocação são inversamente relacionadas, uma tende a diminuir quando a outra sofre um aumento.



$$\text{Precisão} = \frac{\text{Número de n-gramas em comum com o sumário automático}}{\text{Número de n-gramas do sumário automático}} \quad (5.1)$$

$$\text{Revocação} = \frac{\text{Número de n-gramas em comum com o sumário automático}}{\text{Número de n-gramas do sumário referência}} \quad (5.2)$$

$$\text{Medida f} = \frac{2 * P * C}{P * C} \quad (5.3)$$

Para exemplificar, considere o seguinte fragmento como texto-fonte “o gato estava embaixo da cama” e a sentença candidata “o gato foi encontrado embaixo da cama”. Sem considerar os procedimentos de pré-processamento, as palavras em comum das duas sentenças são [o, gato, embaixo, da, cama]; então, os valores para revocação, precisão e medida-f considerando a ROUGE-1 são respectivamente: 0,83, 0,71 e 0,76.

Por ser rápida, barata e não sujeita à subjetividade, a ROUGE é uma das medidas mais populares para avaliar sumários (p.ex. (Cardoso, 2014), (Gambhir; Gupta, 2017), (Aliguliyev et al., 2019), (Parida; Motlicek, 2019), (Paiola et al., 2022)). A correlação dessa medida com o julgamento humano aumenta quando se utiliza vários sumários de referência. No entanto, a ROUGE aborda somente a capacidade de seleção de conteúdo dos sistemas, ignorando aspectos de qualidade linguística como coerência e gramaticalidade. Nesse caso, como o cálculo da pontuação ROUGE é realizado a partir da sobreposição de n-gramas, se um sumário apresentar o mesmo conteúdo do sumário de referência, contendo as mesmas ideias, mas utilizando expressões diferentes, a pontuação final não deve ser muito alta, apesar da qualidade do sumário. Essa, inclusive, é uma limitação da medida ROUGE na avaliação de sumários abstrativos (Gupta; Gupta, 2019).

Nesse sentido, em 2005, a DUC (*Document Understanding Conference*)⁴ introduziu cinco propriedades linguísticas para avaliar a qualidade dos sumários automáticos e que não utiliza sumários de referência, a saber⁵:

- (i) **gramaticalidade:** que diz respeito à ausência de erros de ortografia, pontuação e sintaxe;
- (ii) **não redundância:** que se refere à ausência de informações repetidas;
- (iii) **clareza referencial:** que diz respeito à clara identificação dos componentes da superfície textual que fazem remissão a outro(s) elemento(s) do sumário;
- (iv) **foco:** se refere ao fato de que as informações de uma sentença devem se relacionar com as informações do restante do sumário;
- (v) **estrutura e coerência:** que diz respeito à organização do sumário considerando sua textualidade.

Para avaliar de acordo com os critérios estabelecidos pela DUC, coleta-se a opinião de um grupo de juízes sobre um mesmo sumário e calcula-se a média para cada critério julgado. Cada anotador atribui uma nota que varia de 1 (muito ruim) a 5 (muito bom).

⁴Em 2008, a DUC passou a se chamar TAC (*Text Analysis Conference*). Disponível em: <https://tac.nist.gov/>.

⁵Original: <https://www-nlpir.nist.gov/projects/duc/duc2006/quality-questions.txt>.



Vale destacar que, embora esse tipo de avaliação não dependa do uso de sumário de referência, ela pode beneficiar sumários automáticos que sejam bastante diferentes daqueles de referência. Provavelmente, esses sumários automáticos teriam notas muito baixas pela ROUGE, mas, ainda assim, poderiam ser considerados informativos e coerentes. Por essa razão, alguns trabalhos de SA utilizaram tais medidas avaliativas, além da ROUGE (Ermakova et al., 2019; Pitler et al., 2010).

Conforme Luhn (1958), a produção de sumários é uma tarefa intelectual e que sofre influência da familiaridade com o assunto, atitude e disposição do produtor. O autor também sugere que a produção de sumários de referência pode depender dos interesses de quem o produz, dos interesses dos leitores e da importância subjetiva que ele atribui às informações textuais. Assim, algumas ponderações podem ser feitas a respeito disso: os humanos divergem na escolha de informações importantes; e existe a possibilidade de que a mesma pessoa, ao resumir novamente um texto-fonte, crie um sumário totalmente diferente do anterior (Hailu et al., 2020; Luhn, 1958).

Contudo, a necessidade de anotações ou avaliações manuais muitas vezes inviabiliza suas aplicações na avaliação de métodos de sumarização, em particular para bases de dados com uma quantidade massiva de textos. Por esse motivo, mesmo para trabalhos voltados à SA abstrativa é comum encontrar o uso de medidas mais simples, que podem ser calculadas de forma automática, como a ROUGE. Embora seja factual, há esforços recentes para desenvolver métricas mais sensíveis ao nível semântico, a partir de modelos de linguagem como avaliadores.

Além disso, no cenário atual em que a grande área de geração de texto em linguagem natural evolui rapidamente, críticas surgem a medidas de avaliação que dependem estritamente da sobreposição de unidades de n-gramas entre sumários de referência e candidatos, o que não é adequado para medir a qualidade de sumários abstrativos. Assim, têm surgido frentes de trabalhos que buscam criar protocolos de avaliação desde as métricas aplicadas até a qualidade dos conjuntos de dados utilizados.

Algumas métricas recentes buscam avaliar a similaridade entre textos com base em seus significados, aproximando a análise da semântica. Essas métricas utilizam representações vetoriais de palavras e sentenças (os chamados *embeddings*), o que permitem captar relações mais sutis entre textos com vocabulários diferentes, mas significados semelhantes. Um exemplo é o BERTScore (Zhang et al., 2020), que calcula a similaridade entre cada *token* do sumário automático e do sumário de referência utilizando *embeddings* extraídos de modelos como o BERT. Nesse sentido, essa métrica pode reconhecer paráfrases ou reformulações que mantêm o conteúdo original, superando algumas das limitações da ROUGE. Ainda que métricas como o BERTScore possam apresentar maior correlação com julgamentos humanos e serem mais robustas a variações linguísticas, seu uso ainda é restrito devido à necessidade maior de poder computacional para rodar modelos de linguagem para a extração de *embeddings*. Além disso, medidas de similaridade semântica ainda precisam de um sumário de referência, podendo penalizar um sumário candidato que apresente uma seleção de fatos diferente do apresentado no sumário de referência, mesmo que fosse considerado por um avaliador humano como um bom resumo do texto-fonte.

Mais recentemente, com a popularização dos LLMs, como o ChatGPT, têm ganhado espaço abordagens em que os próprios modelos são utilizados como avaliadores de sumários. Essa nova linha de pesquisa, conhecida como *LLM-as-a-Judge*, propõe que o modelo desempenhe o papel de juiz, avaliando diretamente a qualidade de um sumário com base em critérios específicos, como factualidade, coerência, fluência ou mesmo preferências humanas expressas em anotações anteriores. Um dos principais atrativos dessas abordagens



é a flexibilidade: com um único modelo, é possível avaliar diversos aspectos qualitativos do texto, com menor necessidade de métricas separadas ou protocolos elaborados (Gu et al., 2025).

Na SA, por exemplo, a utilização de LLMs já demonstrou que é possível avaliar sumários em múltiplas dimensões, como informatividade e coerência, por exemplo, atribuindo notas em escalas (por exemplo, de 1 a 5). As avaliações sobre esses métodos têm mostrado boa correlação com julgamentos humanos e podem ser utilizados mesmo quando há múltiplos sistemas sendo comparados (Gao et al., 2023). Além disso, a mesma abordagem pode ser estendida para domínios multimodais ou especializados, desde que os LLMs tenham sido ajustados para tais contextos.

Apesar do grande potencial, o uso de LLMs como avaliadores ainda enfrenta alguns desafios. Um deles é a necessidade de sumários de referência, especialmente em tarefas que requerem comparação com um texto ideal. Quando o sumário automático inclui informações implícitas ou inferenciais que não estão objetivamente no texto-fonte, o modelo pode ter dificuldade para julgar sua adequação sem informações externas ou conhecimento de mundo – o que pode levar a equívocos ou inconsistências. Outro ponto crítico diz respeito à possibilidade de alucinações: como os LLMs são treinados com grandes volumes de texto da Web, eles podem emitir julgamentos, correlações e/ou inferências com base em “conhecimentos” incorretos, incompletos ou não factuais. Por fim, ainda que tragam escalabilidade e redução de dependência de humanos, o uso de LLMs não é isento de custos: executar modelos de língua, sobretudo os comerciais, para avaliar milhares de sumários pode ser financeira e computacionalmente oneroso, principalmente em comparação com métricas tradicionais como a ROUGE.

Portanto, embora a utilização de LLMs como juízes na avaliação automática de sumários represente um avanço significativo, questões relacionadas à confiabilidade, à robustez e ao custo ainda precisam ser cuidadosamente consideradas. Como destacado no survey recente sobre LLM-as-a-Judge (Gu et al., 2025), há um esforço crescente da comunidade para padronizar metodologias, mitigar vieses e ampliar benchmarks específicos para avaliação automática. Ainda assim, o uso responsável e crítico dessas ferramentas é essencial para garantir que os sistemas de avaliação acompanhem de forma justa e precisa a evolução dos modelos de sumarização.

5.4 Aplicação em Sumarização Extrativa

De maneira geral, os sumários são produzidos a partir de três etapas⁶, conforme a Figura 5.2.

Na primeira etapa, conhecida como **análise**, os textos-fonte passam por um processo de interpretação, resultando em uma representação formal deles. Na fase seguinte, tida como **transformação**, o sumário é elaborado a partir do ranqueamento de segmentos dos textos-fonte em função de algum critério que indique a relevância, sendo selecionados os segmentos com maiores pontuações até que se atinja a taxa de compressão desejada. Na etapa de **síntese**, gera-se, então, o sumário em língua natural a partir do conteúdo selecionado. Nessa etapa podem ser utilizados métodos de tratamento de correferência, fusão, linearização, justaposição e ordenação de sentenças. Essas três fases contribuem umas com as outras, de modo que alguns métodos que ocorrem na síntese também poderiam

⁶Para mais detalhes sobre as três etapas dessa arquitetura genérica, indicamos (Jones, 1993) e (Mani, 2001).



Figura 5.2: Etapas do processo de sumarização



Fonte: Adaptado de (Sparck-Jones, 1998).

estar na fase de transformação, e vice-versa.

A partir dessa arquitetura genérica, apresentamos nesta seção dois algoritmos com estratégias distintas de seleção de conteúdo para gerar os sumários. É possível que para compreender algumas questões que levantaremos aqui sejam necessárias informações prévias. Pensando nisso, elaboramos o Apêndice A, em que discutimos aspectos relevantes ao pré-processamento dos textos para SA e que complementa o assunto abordado nesta seção. Ademais, para complementar a leitura, disponibilizamos um notebook⁷ virtual com os algoritmos⁸ apresentados neste capítulo. Sugerimos que a leitura desta seção seja acompanhada desse material suplementar.

5.4.1 Algoritmo baseado na frequência das palavras

Uma das unidades mínimas de análise linguística é a palavra, compreendida dentro do nível morfológico; quando combinadas entre si, é possível construir outros objetos de análise, como a sentença. Para a sumarização de textos escritos, as palavras e as sentenças podem ser o início do processo de identificação de conteúdo relevante. Os métodos clássicos de identificação de segmentos relevantes avaliam a importância de cada sentença em um texto-fonte com base em seu peso e, em seguida, selecionam aquelas com pesos mais elevados (acima de um limite mínimo) para formar o resumo (Baxendale, 1958; Edmundson, 1969; Luhn, 1958). As propostas mais recentes utilizam *word embeddings* (Hailu et al., 2020) e modelos de redes neurais profundas, como BERT (Liu; Lapata, 2019).

Nesta seção, apresentamos uma estratégia simples baseada na frequência de palavras para selecionar sentenças relevantes do texto-fonte para compor os sumários. Nessa perspectiva, quanto mais uma palavra se repete no texto, mais relevante ela é. Propusemos um algoritmo geral, organizado em sete passos, que seleciona as sentenças com maior número de palavras relevantes para compor um possível sumário. Vamos aplicar o algoritmo ao Texto-fonte 1 do Quadro 5.1 com o objetivo de determinar um *ranking* de sentenças para a composição do sumário.

Quadro 5.3: Algoritmo geral

⁷https://colab.research.google.com/drive/1qGfpxQJgnJy_UKKiHpKQSRACiFRwjyeP?usp=sharing Os resultados podem ser um pouco diferentes dos que são apresentados aqui, devido aos testes realizados durante a escrita do texto.

⁸*Grosso modo*, algoritmo pode ser definido como um conjunto de regras, compreendendo uma sequência finita de ações visando a resolução de alguma tarefa/problema.

Pré-processamento

Passo 1: Normalização do texto

Passo 2: Remoção de caracteres de pontuação

Passo 3: Remoção de *stopwords***Geração do sumário**

Passo 4: Cálculo das frequências das palavras que compõe o texto pré-processado

Passo 5: Determinação do peso de cada palavra

Passo 6: Cálculo do peso de cada sentença

Passo 7: Seleção das sentenças com maiores pesos até atingir a taxa de compressão

Ao executar o **Passo 1**, todo o texto será transformado para letras minúsculas para fins de padronização. Esse procedimento faz com que tenhamos palavras graficamente distintas a serem consideradas de maneira similar, como “hepatite” e “Hepatite”.

Após a normalização, no **Passo 2**, devem ser removidos os caracteres de pontuação. Essa operação se justifica pelo fato de que uma palavra, acompanhada de uma pontuação, difere-se de outra com a mesma grafia sem pontuação (como “cahe” e “cahe,”). Os caracteres de pontuação podem divergir de um método para outro, a partir de diferentes linguagens de programação ou *software*. Aqui, consideramos a pontuação sendo o ponto final, a vírgula, a sequência de dois hífen (--) e as aspas (” ou “).

Já no **Passo 3**, removemos as *stopwords* utilizando a lista do Natural Language Toolkit – NLTK – (Bird, 2006). Por exemplo, o trecho “o médico pessoal do argentino diego” torna-se “médico pessoal argentino diego” após a remoção dos *tokens* “o” e “do”.

É importante destacar que a remoção das *stopwords* se baseia na hipótese da distribuição (Capítulo [Representação vetorial e semântica distribucional](#)). Assim, na SA com abordagem lexical, acredita-se que após a remoção das *stopwords*, as palavras restantes representam melhor o conteúdo e aquelas que mais se repetem, são relevantes. Assim, no **Passo 4**, deve-se calcular a frequência de ocorrência dos *tokens*. Os cinco termos mais frequentes, em ordem alfabética do Texto-fonte 1 estão listados na Tabela 5.1.

Tabela 5.1: Cinco termos mais frequentes do Texto-fonte 1

Termo	Frequência	Peso
maradona	5	$5/5 = 1,0$
cahe	4	$4/5 = 0,8$
hepatite	3	$3/5 = 0,6$
internado	3	$3/5 = 0,6$
recaída	3	$3/5 = 0,6$

A partir da frequência das palavras, no **Passo 5** conseguimos determinar diferentes pesos para os *tokens* com base na frequência relativa. O cálculo é definido pela divisão da frequência de cada termo pela maior frequência calculada. No caso, o termo “maradona” é o que mais ocorre no texto, com frequência 5. Assim, todos os termos do texto devem ser divididos por 5, determinando o peso de cada um deles. Ressalta-se que essa não é a única forma de ponderar a relevância de uma palavra ou sentença em um texto. No exemplo, para fins de simplificação, optamos por considerar a fala do médico como uma única sentença por se tratar de um discurso direto compreendido entre aspas.

Uma vez calculados os pesos de cada uma das palavras, podemos estimar a pontuação das sentenças. Considerando o Texto-fonte 1, no **Passo 6**, definimos a pontuação por meio da soma dos pesos da frequência relativa de cada palavra restante em cada sentença.



Na Tabela 5.2, apresenta-se o peso final de cada sentença. Por exemplo, a Sentença 1 tem o peso 5,8, pois a soma dos pesos dos *tokens* (destacados no texto) resulta nesse valor.

Tabela 5.2: Sentenças do texto-fonte com seus respectivos pesos

Sentença	Marcações	Peso
1 O médico pessoal do argentino Diego Maradona, Alfredo Cahe, revelou nesta segunda-feira que uma recaída da Hepatite aguda de que sofre foi o motivo da nova internação do ex-craque.	O (médico - 0,2) (pessoal - 0,2) do (argentino - 0,2) (Diego - 0,2) (Maradona - 1,0), (Alfredo - 0,2) (Cahe - 0,8), (revelou - 0,2) (nesta - 0,2) (segunda-feira - 0,2) que uma (recaída - 0,6) da (Hepatite - 0,6) (aguda - 0,4) de que sofre fo (motivo - 0,2) da (nova - 0,2) (internação - 0,2) do (ex-craque - 0,2).	5,8
2 Maradona havia recebido alta no último dia 11, mas voltou a ser internado na sexta-feira e os boletins médicos não especificaram o que se passava com o ex-jogador – Cahe descartou pancreatite ou úlcera.	(Maradona - 1,0) (havia - 0,2) (recebido - 0,2) (alta - 0,4) no (último - 0,2) (dia - 0,2) (11 - 0,2), mas (voltou - 0,4) a ser (internado - 0,6) na (sexta-feira - 0,2) e os (boletins - 0,2) (médicos - 0,2) não (especificaram - 0,2) o que se (passava - 0,2) com o (ex-jogador - 0,2) (Cahe - 0,8) (descartou - 0,2) (pancreatite - 0,2) ou (úlcera - 0,2).	6,0
3 “Maradona teve uma recaída na hepatite aguda. Agora está estável. Apesar de ter melhorado no domingo, deverá continuar internado”, disse Cahe, em declarações ao jornal “La Nación”.	“(Maradona - 1,0) teve uma (recaída - 0,6) na (hepatite - 0,6) (aguda - 0,4). (Agora - 0,2) está (estável - 0,2). (Apesar - 0,2) de ter (melhorado - 0,2) no (domingo - 0,2), (deverá - 0,2) (continuar - 0,2) (internado - 0,6)”, (dis (Cahe - 0,8), em (declarações - 0,2) ao (jornal - 0,2) “(La - 0,2) (Nación - 0,2)”.	6,6
4 Maradona, 46, desenvolveu uma hepatite tóxica por excesso de consumo de álcool, o que já o manteve internado durante 13 dias antes da primeira alta.	(Maradona - 1,0), (46 - 0,2), (desenvolveu - 0,2) um (hepatite - 0,6) (tóxica - 0,2) por (excesso - 0,2) de (consumo - 0,2) de (álcool - 0,2), o que já o (manteve - 0,2) (internado - 0,6) (durante - 0,2) (13 - 0,2) (dias - 0,2) (antes - 0,2) da (primeira - 0,2) (alta - 0,4).	5,0
5 Cahe disse ainda que Maradona não voltou a consumir bebidas alcoólicas e que as causas da recaída estão sendo investigadas.	(Cahe - 0,8) (disse - 0,4) (ainda - 0,2) que (Maradona - 1,0) não (voltou - 0,4) a (consumir - 0,2) (bebidas - 0,2) (alcoólicas - 0,2) e que as (causas - 0,2) da (recaída - 0,6) estão (sendo - 0,2) (investigadas - 0,2).	4,6

No **Passo 7**, ocorre a seleção de conteúdo relevante. Para fins de exemplificação, a taxa de compressão utilizada foi de 60% em relação ao número de sentenças originais. Apesar disso, destacamos que geralmente a taxa de compressão é calculada em função do número de palavras do texto-fonte. Assim, o sumário poderá ter duas sentenças, conforme se mostra no Quadro 5.4.

Quadro 5.4: Sumário gerado com algoritmo de frequência de palavras

Maradona havia recebido alta no último dia 11, mas voltou a ser internado na sexta-feira e os boletins médicos não especificaram o que se passava com o ex-jogador – Cahe descartou pancreatite ou úlcera.

“Maradona teve uma recaída na hepatite aguda. Agora está estável. Apesar de ter melhorado no domingo, deverá continuar internado”, disse Cahe, em declarações ao jornal “La Nación”.

A partir desse algoritmo, é importante fazermos algumas considerações:

- **termos com maiores frequências:** a depender do estudo, talvez seja interessante remover os termos que ocorrem em maior frequência, pois podem não determinar exatamente a essência do texto. Suponha que estamos analisando textos com a palavra “covid”. Se o *corpus* for criado a partir desse termo, é quase certo que todos os textos que o compõem devam possuí-lo. O pesquisador deve avaliar se é interessante tê-lo como fator determinante ou inseri-lo em sua *stoplist*;
- **frases curtas *versus* longas:** avaliando essa comparação, as frases com mais palavras tendem a ter um maior peso, porém podem não ser tão relevantes. Conforme demonstrado na Tabela 5.2, por exemplo, a Frase 1 tem 29 palavras e possui peso 5,8. Já a Frase 2 possui 33 palavras com peso 6 e, a Frase 3 apresenta 27 palavras e o maior peso 6,6. A forma como o algoritmo determina o peso de uma frase deve ser criteriosamente avaliada;
- ***stoplist*:** a lista de *stopwords* pode interferir no resultado, uma vez que determina quais palavras devem ser removidas, influenciando nos pesos atribuídos pelos métodos aplicados. Portanto, os resultados dos métodos de SA podem variar conforme a *stopwlist* utilizada;
- **a ordenação do conteúdo selecionado:** a tarefa de ordenação de sentenças em sumarização multidocumento é difícil quando comparada com a sumarização monodocumento. O conteúdo é extraído de diferentes fontes e, portanto, nenhum documento pode fornecer a ordenação adequada (Barzilay et al., 2001). Como o sumário do Quadro 5.4 é monodocumento, o conteúdo selecionado foi organizado na ordem em que apareceu no texto-fonte.

5.4.2 Algoritmo baseado em Transformers

O modelo de *Bidirectional Encoder Representations from Transformers* (BERT) (Capítulo Modelos de linguagem) é considerado um dos melhores algoritmos para sumarização da atualidade, apresentando resultados expressivos (Rogers et al., 2021) em diferentes áreas. O algoritmo foi desenvolvido nos laboratórios da Google, a fim de implementar melhorias no algoritmo de busca, permitindo converter as intenções dos usuários (termo de busca) para a codificação adequada contribuindo para otimizar e melhorar a assertividade (LIVESO, 2020). Outra contribuição é a possibilidade do treinamento do algoritmo para atender a um determinado idioma. Algumas empresas direcionaram seus serviços para o treinamento dos idiomas e áreas específicas com demanda, como o português (NEURAL-MIND, 2020) e o meio empresarial (Julião, 2024). Em suma, o modelo BERT, por meio de técnicas avançadas de PLN, possibilitou uma melhor compreensão da informação textual



por parte dos modelos matemáticos, o que por sua vez permitiu melhorias e uma ampla gama de aplicações em diversas áreas.

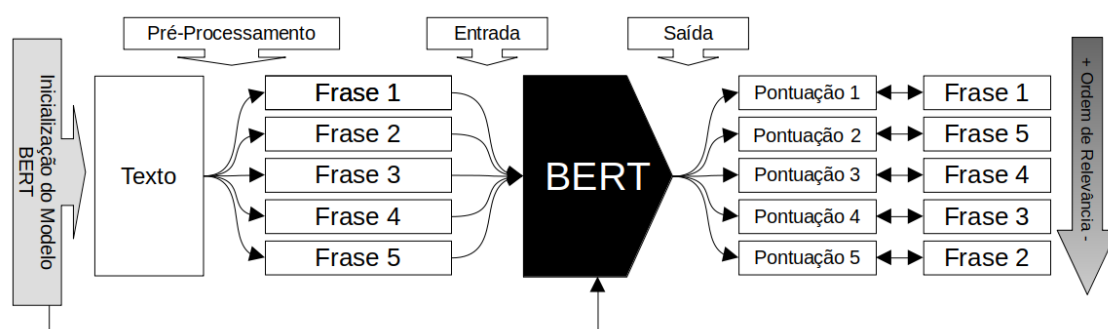
Diferente de seus antecessores, aqui o algoritmo se baseia em determinar as palavras-chave e variações para atribuir relevância aos trechos de um texto a fim de resumizá-lo. Os antecessores, na maioria, analisavam as palavras individualmente, ou às vezes, considerando os termos ao entorno da palavra-chave requisitada. O BERT mudou essa estratégia, permitindo mapear o texto de forma geral. Dessa forma, o algoritmo possibilita até prever termos a partir de um contexto, ampliando as possibilidades de aplicações.

O processamento dos primeiros algoritmos era fundamentado em processar o texto, independente de o texto estar repetido ou não entre a coleção de documentos, para gerar uma lista das palavras-chave e atribuir pontuação às sentenças. Nesse sentido, até então, o modelo não possuía uma memória, demandando um novo processamento a cada texto a ser analisado. Essa questão foi contornada após a criação do BERT e apresentado por Devlin et al. (2019), sendo implementado por meio da tecnologia *Transformers* (Vaswani et al., 2017b).

O BERT pode ser aplicado tanto na sumarização abstrativa⁹ quanto extrativa; porém, aqui, nos deteremos à segunda. O algoritmo é baseado em uma aplicação com treinamento bidirecional. Diferente dos métodos que focam na estrutura composicional das sentenças, que processam as palavras ora da esquerda para direita, ora da direita para esquerda (como os primeiros algoritmos de sumarização, com destaque para o Luhn (1958), Earl (1970) e Edmundson (1969)), no BERT o tratamento ocorre lendo os dados nos dois sentidos durante o treinamento. Desse modo, o algoritmo permite que a palavra seja analisada dentro de um “contexto”, levando em consideração o que acontece ao seu redor.

Vaswani et al. (2017b) ressaltam a importância do treinamento bidirecional e demonstram que, uma vez treinado o modelo para o idioma desejado, reduz-se a necessidade de arquiteturas específicas, permitindo uma generalização na solução de desafios, como a escolha do conteúdo mais relevante nos sumários. Portanto, uma vez realizado um pré-treinamento, basta realizar um ajuste fino na rede para atender a casos específicos. No caso da tarefa de SA, o modelo pode ser utilizado para classificar e extrair as sentenças mais relevantes de um texto. Aqui, organizamos o processo em cinco passos, conforme a Figura 5.3.

Figura 5.3: Ilustração do processo simplificado de sumarização do BERT



De acordo com a Figura 5.3, no **Passo 1**: o BERT deve ser inicializado, carregando

⁹Quanto à sumarização abstrativa, sugerimos a leitura de (Barros, 2022), (Nascimento, 2023) e (Silva, 2022).

algumas configurações do modelo já treinado (“Inicialização do Modelo BERT”). No **Passo 2**, o texto a ser fornecido ao BERT passa por um pré-processamento, que compreende segmentação em sentenças e tokenização. O pré-processamento pode incluir outras etapas, como visto no algoritmo apresentado na seção anterior. Após os ajustes iniciais, o texto é submetido ao modelo na camada de codificação (“Entrada”). Ao entrar na rede, os *tokens* passam a compor um vetor numérico e constituem um espaço vetorial¹⁰ para mapear o texto. Já no **Passo 3**, a partir dos vetores no espaço vetorial e do processamento do modelo BERT, em sua camada de decodificação (“Saída”), atribui-se um valor de importância de cada palavra no contexto do texto analisado. No **Passo 4**, com o valor de importância de cada palavra, torna-se possível atribuir um valor (ou “importância”) à sentença que a possui (“Pontuação”). Uma vez definido esse valor para cada sentença, é possível classificá-las em ordem de relevância. Por fim, no **Passo 5**, a partir do nível de compressão informado no início do processamento pelo usuário, o BERT retorna as sentenças para compor o sumário de acordo com a taxa de compressão escolhida. Para exemplificar, são apresentados dois sumários gerados pelo BERT. O Sumário 2 segue a taxa de compressão do algoritmo anterior que é de 60% em relação ao número de sentenças originais e o Sumário 3 coleta uma sentença a mais, isto é, três sentenças.

Caso o sumário produzido não seja condizente com as expectativas, o modelo BERT pode sofrer ajustes finos para aprimorar os resultados. A tecnologia do BERT já fornece um modelo treinado a partir de dois *corpora* volumosos (a saber, *BookCorpus*¹¹ e Wikipedia). O treinamento do BERT consiste em fornecer ao algoritmo um *corpus* específico da área requisitada, o qual deve ser modificado para se adequar a essa etapa. As adequações previstas são *texto-previsão* - a criação de um conjunto de texto (frases); *previsão da próxima frase*; e *máscara* - modelagem de linguagem mascarada. A adequação texto-previsão consiste em ajustar o texto de entrada e a respectiva frase que deverá ser prevista pelo algoritmo; com isso, parte do treinamento é executado. A segunda parte do treinamento consiste na adaptação máscara em ajustar o texto em *tokens* (usando o WordPiece¹²), fazendo a permuta de alguns termos por máscaras (como “###”) e os possíveis valores que podem substituí-las. Após o *corpus* ser ajustado, ele é submetido ao BERT para treinamento e geração do modelo adequado ao caso requisitado. De forma geral, esse ajuste fino ocorre em casos muito específicos, nos quais o texto pode ser de determinados domínios, como jurídico ou médico, onde podem ocorrer jargões e termos técnicos que não estão presentes na base geral de treinamento do modelo.

5.4.3 Avaliando os sumários

Até aqui, indicamos de maneira teórico-prática como funcionam os algoritmos de sumarização extrativa baseados em frequência de palavras e no modelo BERT. Agora, nosso objetivo é analisar como os sumários produzidos por esses algoritmos podem diferir entre si a partir de um texto-fonte com cinco sentenças.

¹⁰Para um melhor entendimento do espaço vetorial, sugere-se consultar o Capítulo [Representação vetorial e semântica distribucional](#).

¹¹<https://paperswithcode.com/dataset/bookcorpus>

¹²Wordpieces são partes de palavras, devendo não confundir-se com morfemas, já que não carregam nenhum significado. Wordpieces podem ser obtidas de maneira empírica e constituem um vocabulário induzido a partir de dados do texto. Para saber mais detalhes sobre esse conceito, confira o Capítulo [Texto ou fala?](#) e o Capítulo [Sequência de caracteres e palavras](#). Neste capítulo utilizamos o pacote WordPiece proposto pela Google, e disponível em: <https://huggingface.co/learn/nlp-course/chapter6/6?fw=pt>.



No Quadro 5.5 são apresentados os sumários extrativos gerados pelos métodos Frequência de Palavras e BERT. O texto-fonte é o mesmo do Quadro 5.1.

Quadro 5.5: Sumários gerados a partir dos métodos Luhn e BERT

Método	Texto	Qtd. sentenças	Qtd. palavras
Sumário 1 - Frequência de palavras	Maradona havia recebido alta no último dia 11, mas voltou a ser internado na sexta-feira e os boletins médicos não especificaram o que se passava com o ex-jogador – Cahe descartou pancreatite ou úlcera. “Maradona teve uma recaída na hepatite aguda. Agora está estável. Apesar de ter melhorado no domingo, deverá continuar internado”, disse Cahe, em declarações ao jornal “La Nación”.	2	60
Sumário 2 - BERT	O médico pessoal do argentino Diego Maradona, Alfredo Cahe, revelou nesta segunda-feira que uma recaída da hepatite aguda de que sofre foi o motivo da nova internação do ex-craque. Maradona, 46, desenvolveu uma hepatite tóxica por excesso de consumo de álcool, o que já o manteve internado durante 13 dias antes da primeira alta.	2	54
Sumário 3 - BERT - 3 sentenças	O médico pessoal do argentino Diego Maradona, Alfredo Cahe, revelou nesta segunda-feira que uma recaída da hepatite aguda de que sofre foi o motivo da nova internação do ex-craque. Maradona havia recebido alta no último dia 11, mas voltou a ser internado na sexta-feira e os boletins médicos não especificaram o que se passava com o ex-jogador – Cahe descartou pancreatite ou úlcera. Maradona teve uma recaída na hepatite aguda.	3	69

De maneira geral, observa-se que a taxa de compressão pode impactar diretamente a informatividade do sumário. Todos os três sumários gerados apresentam informações necessárias à compreensão do evento (no caso, o problema de saúde de Maradona e a recaída de seu quadro clínico). Apesar disso, destaca-se que, apesar de a informatividade ser um conceito, por vezes, subjetivo, ele está associado ao objetivo do sumário (se ele deve ser informativo, indicativo ou avaliativo, por exemplo).

Conforme já mencionado, é importante planejar a organização do conteúdo selecionado para que o texto final faça sentido e satisfaça a necessidade do usuário. Por exemplo, no Sumário 1, embora a segunda sentença tenha maior pontuação que a primeira, se ela iniciar o texto final, o leitor perderá o contexto de que Cahe é o médico de Maradona.

Num cenário em que comparamos os sumários com o texto-fonte, notamos que todos selecionaram sentenças do início e do meio do conteúdo original. Possivelmente essa escolha se baseou no fato de as sentenças que finalizam textos jornalísticos apresentarem

detalhamentos sobre o tópico principal, enquanto as primeiras sentenças apresentarem sumariamente o assunto.

Além de aspectos linguístico-textuais que devem ser observados nos sumários, como visto, há parâmetros estatísticos que podem ser utilizados durante a avaliação. Para tanto, os três sumários foram avaliados usando a medida ROUGE, mais especificamente ROUGE-1, ROUGE-2 e ROUGE-L. Os valores obtidos para cada sumário são apresentados na Tabela 5.3.

Tabela 5.3: Resultado da avaliação pelo método ROUGE

Algoritmos	Rouge 1			Rouge 2			Rouge L		
	R	P	F	R	P	F	R	P	F
Frequência de Palavras	0.447	1.00	0.618	0.443	1.00	0.614	0.447	1.00	0.618
Bert - 20%	0.394	1.00	0.566	0.390	1.00	0.561	0.394	1.00	0.566
Bert - 3 sentenças	0.519	1.00	0.683	0.516	1.00	0.681	0.619	1.00	0.683

Uma característica do texto jornalístico é que as informações localizadas no início do texto expressam o fato principal de uma notícia (Canavilhas, 2012), por isso, são selecionadas para compor o sumário. Tal consideração pode ser diferente para outros gêneros textuais. O resultado da métrica ROUGE, na comparação de sumários automáticos extrativos com seu texto-fonte (referência), justifica a precisão ser 1 para todos os casos. Se a comparação fosse em relação a um sumário humano com alguma reescrita, o resultado seria totalmente diferente.

Observa-se que os valores da medida-f não são muito diferentes entre si. Cada um dos sumários traz um contexto que pode ser importante para um tipo de leitor, mas não para outro: o Sumário 1 recupera a opinião do médico e o Sumário 2 apresenta o breve histórico da saúde de Maradona. Ao reduzir a taxa de compressão, o Sumário 3 foi mais abrangente que os demais, resultando em uma melhor avaliação em termos de medida-f.

5.5 Aplicação em Sumarização Abstrativa

A tarefa de sumarização abstrativa visa ir além da extração e concatenação de sentenças ou trechos do texto-fonte. Seu objetivo é gerar novas sentenças que expressem o conteúdo principal do texto, muitas vezes reescrevendo e reorganizando informações, prevendo maior fluidez e concisão. Embora mais flexível e com potencial para gerar sumários de qualidade superior, essa abordagem também apresenta maior complexidade e riscos, como a possibilidade de geração de informações incorretas (alucinações) ou desconectadas do conteúdo original.

Assim como na sumarização extrativa, o processo pode ser dividido em etapas. Tradicionalmente, considera-se o seguinte *pipeline*: extração de informações, seleção de conteúdo e geração de sentenças (Lin; Ng, 2019). A seguir, apresentamos brevemente dois grupos de abordagens para construção de sumarizadores abstrativos: as baseadas em estrutura e/ou semântica, e aquelas baseadas em modelos neurais, especialmente na arquitetura Transformer.



5.5.1 Algoritmos baseados em estrutura sintática e/ou semântica

Os algoritmos desta categoria buscam representar as informações mais importantes do texto por meio de estruturas como árvores de dependência, grafos semânticos ou ontologias, para, então, gerar sentenças-resumo. Em geral, esses métodos combinam extração de estruturas linguísticas e alguma técnica de geração textual (Gupta; Gupta, 2019).

Como exemplo, destacam-se abordagens baseadas em árvores de dependência, que buscam representar a estrutura sintática das sentenças (veja Capítulo [A ordem e a função das palavras em uma sentença](#)). Normalmente estes métodos partem de uma sumarização extrativa, com a seleção das sentenças consideradas mais importantes, então sentenças similares são identificadas a partir de um analisador e posteriormente são construídas árvores de dependência para estas sentenças, buscando fundir as árvores de sentenças semelhantes. Por fim, ocorre um processo de linearização da árvore resultante, convertendo-a para texto.

Árvores de dependência, porém, buscam capturar estritamente a estrutura sintática das sentenças, não sendo possível captar, por si só, nuances de contexto. Por exemplo, considere o seguinte exemplo: “A menina postou uma foto com o celular” apresentado no Capítulo [A ordem e a função das palavras em uma sentença](#). Esta sentença pode ser interpretada de, ao menos, duas maneiras: (i) a menina postou uma foto a partir do seu celular, ou (ii) ela postou uma foto em que ela aparece utilizando o celular. A sintaxe da sentença, por si só, além de poder ser ambígua, pode não expressar exatamente o que ocorreu, sendo necessário mais informações sobre o contexto.

Por outro lado Neste sentido, abordagens baseadas em semântica consistem em obter uma representação semântica do texto e, a partir disso, gerar um sumário. Um exemplo de representação semântica são os grafos AMR (*Abstract Meaning Representation*) (Banarescu et al., 2013), que modelam o conteúdo do texto por meio de entidades e suas relações. Na Figura 5.4 é apresentado um exemplo de um grafo AMR construído para uma determinada sentença. Nela, percebe-se que o verbo “querer”, por exemplo, possui dois argumentos, representados pelas arestas ARG0 (quem quer?) e ARG1 (o que quer?). Por não focar nos itens lexicais da sentença, no modelo AMR o verbo “desejar”, por exemplo, não precisa necessariamente ser mantido no grafo, desde que o significado da sentença seja preservado. Dessa forma, uma mesma sentença pode gerar diferentes representações semânticas em AMRs, assim como uma mesma representação AMR pode ser transformada em sentenças distintas.

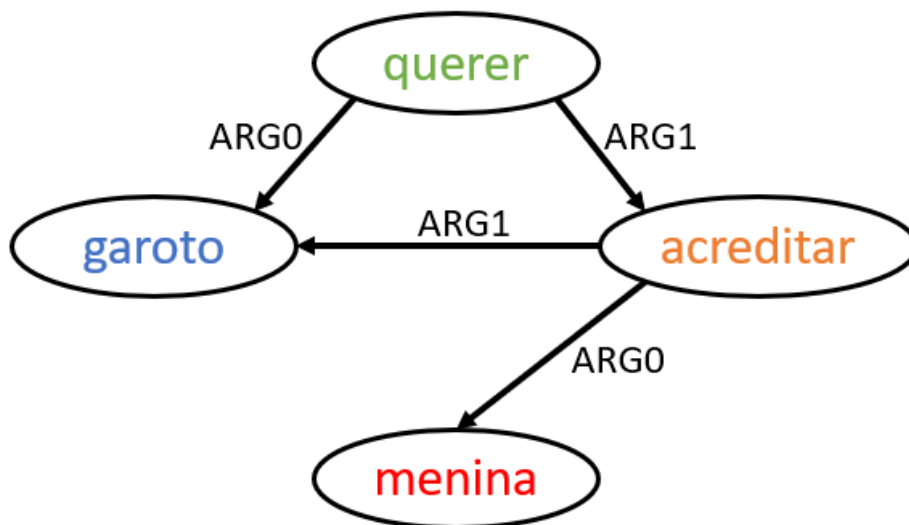
Em (Dohare et al., 2018), é apresentado um sistema de sumarização abstrativa baseado em grafos AMR. A abordagem dos autores se inspira no modo com que os humanos resumem textos. A partir da AMR do texto de entrada, são identificadas as entidades e eventos mais relevantes, seguidos pelas relações-chave entre essas entidades. Posteriormente, o sistema recupera informações de contexto relacionadas às relações selecionadas. O subgrafo resultante representa uma versão condensada da AMR original, e pode ser interpretado como uma sumarização semântica do texto, a partir da qual se gera o sumário final.

Cabe destacar que há um potencial interessante na combinação dessas abordagens com modelos de linguagem. Os Transformers, por exemplo, utilizam grafos semânticos como entrada para modelos gerativos (Hua et al., 2023; Yasunaga et al., 2022) ou como ferramenta para validação (Liu et al., 2024; Qiu et al., 2024; Roy et al., 2023) dos sumários produzidos. Essa integração, no entanto, ainda é uma área em desenvolvimento.



Figura 5.4: Um exemplo de sentença e sua AMR correspondente

O garoto deseja que a menina acredite nele.



Fonte: (Paiola, 2022)

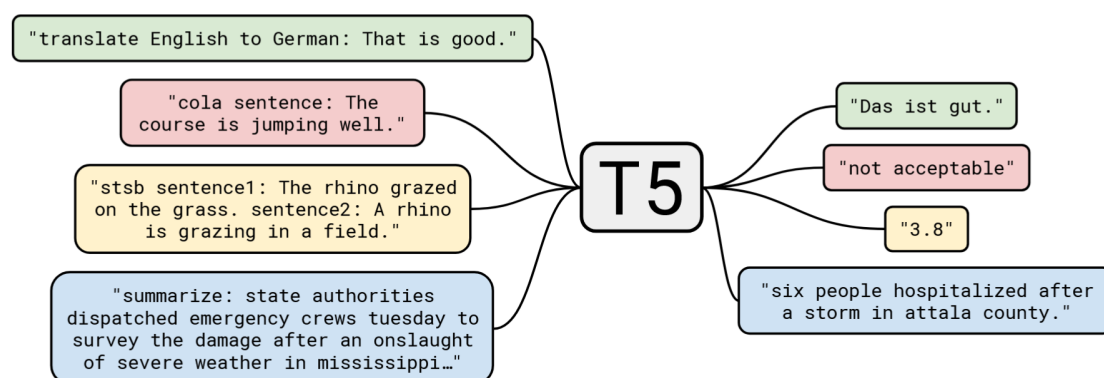
5.5.2 Algoritmos baseados em Transformers

Com o avanço da arquitetura Transformer (Vaswani et al., 2017b), os métodos de sumarização abstrativa passaram a ser dominados por modelos baseados em aprendizado profundo e, mais recentemente, por modelos pré-treinados de larga escala. A aplicação desses modelos permitiu representar melhor a semântica dos textos, possibilitando sumários mais coesos e coerentes.

Um dos modelos nessa linha é o T5 (*Text-To-Text Transfer Transformer*) (Raffel et al., 2020), cujo diferencial está em tratar toda tarefa de PLN como um problema de transformação de texto para texto. Na tarefa de sumarização, o modelo recebe como entrada o texto-fonte precedido de um prefixo que indica a tarefa (por exemplo, "summarize:", como pode ser visto na Figura 5.5), e gera como saída o sumário correspondente. Adaptando esse modelo ao português brasileiro, foi proposto o PTT5 (Carmo et al., 2020b), uma versão do T5 pré-treinada em um grande *corpus* em português, o BrWac (Wagner Filho et al., 2018), sendo posteriormente ajustada para diferentes tarefas, incluindo a sumarização automática, como o PTT5-Summ (Paiola et al., 2022).

Modelos como o PTT5 permitem o uso de diferentes tamanhos de arquitetura (*small*, *base* e *large*), sendo possível realizar o *fine-tuning* específico para a tarefa de sumarização, com custos computacionais variados. O desempenho desses modelos está intimamente ligado ao conjunto de dados utilizado para o ajuste fino. Modelos ajustados em bases com sumários muito curtos, por exemplo, tendem a produzir resumos menores, não podendo ser aplicados para os casos em que necessitemos de sumários mais longos. Do mesmo modo, se o modelo foi treinado para sumarizar notícias, dificilmente ele conseguirá sumarizar adequadamente outros tipos de texto.

Além dos modelos especialistas treinados especificamente para a tarefa de sumarização, também vêm sendo explorados os LLMs, que são mais generalistas, capazes de realizar

Figura 5.5: Um diagrama do *framework* de texto para texto proposto para o modelo T5

Fonte: (Raffel et al., 2020)

diversas funções, incluindo a própria sumarização. Tais modelos são geralmente treinados em múltiplas tarefas e idiomas, podendo ser aplicados diretamente à tarefa de sumarização mesmo sem ajuste fino, apenas com o uso de *prompts* específicos. Por exemplo, ao fornecer um comando como “Escreva um resumo em português com no máximo 150 palavras”, é possível obter um sumário coerente e informativo, ainda que genérico, do ponto de vista funcional.

De modo geral, os LLMs podem conseguir resultados melhores, inclusive possibilitando a personalização de resultados mais facilmente, apenas passando instruções via *prompt*. Por outro lado, há uma série de questões que podem ser consideradas, como o custo de utilizar um LLM, seja de processamento, ao rodar em *hardware* local; ou de utilizar algum modelo fechado via API; até a questão de sigilo de dados, considerando o uso de modelos fechados. Além disso, considerando estudos como o de Sarmento; Oliveira (2024), podemos perceber que, apesar de modelos maiores, como o próprio GPT, realmente mostrarem um desempenho consideravelmente maior que modelos pequenos como PTT5, o mesmo pode não acontecer com LLMs na faixa de 8 a 9 bilhões de parâmetros, que chegam a apresentar resultados próximos ou até inferiores aos modelos especialistas de menor tamanho. Nesse trabalho, o LLaMa 3.1 (Grattafiori et al., 2024) de 8 bilhões de parâmetros obteve uma ROUGE-L de 33,8% na base de dados CSTNews, enquanto o PTT5 *Small*, com *fine-tuning* realizado pelos autores, alcançou 39,3%. O fato é que este último modelo contém 60 milhões de parâmetros, sendo mais de 100 vezes menor que o LLaMa.

Isso não quer dizer que não se deve considerar a aplicação de LLMs para sumarização. Pelo contrário, esse ponto reforça que eles não simplesmente substituem os modelos de linguagem menores ou até outras técnicas de sumarização, mas são importantes para avaliar o contexto de aplicação, com todas as suas particularidades.

Esses pontos evidenciam um dilema atual: modelos especialistas, como o PTT5, quando bem ajustados, tendem a oferecer um bom equilíbrio entre custo e desempenho, especialmente em cenários com restrições computacionais. Por outro lado, os LLMs oferecem maior flexibilidade e qualidade textual, mas com custos mais elevados e maior sensibilidade a variações no *prompt* podendo gerar grandes mudanças no resultado final (Zhang et al., 2025).

Diante desses novos modelos, novos desafios também surgem. Entre eles, destacam-se:

- a necessidade de conjuntos de dados de qualidade para o ajuste fino e avaliação (veja Capítulo [Conjunto de dados, *dataset* e *corpus*](#));
- o risco de os modelos incorporarem vieses ou inserirem informações incorretas (veja Capítulo [Questões éticas em IA e PLN](#));
- e o custo envolvido, que nem sempre é viável em contextos práticos.

Ainda assim, a aplicação de Transformers na SA abstrativa continua sendo um campo promissor e ativo de pesquisa, com potencial de aplicação real em sistemas voltados à leitura assistida, curadoria de notícias, educação e outras áreas que demandam a síntese de grandes volumes de informação textual.

5.5.3 Sumarização abstrativa na prática

Para ilustrar as possibilidades e desafios da sumarização abstrativa, preparamos exemplos de geração de sumários usando modelos de linguagem, assim como diferentes avaliações desses sumários. Os códigos que utilizamos para esses experimentos estão disponíveis¹³.

5.5.3.1 Sobre o experimento

Os experimentos foram realizados com quatro modelos: **PTT5-Summ** (especialista em sumarização em português) e três LLMs generalistas (**Qwen3-8B**, **Sabiá-7B** e **Chatbode-7B**). Cada modelo recebeu textos jornalísticos como entrada, sendo solicitado a gerar um sumário abstrativo – ou seja, reescrevendo livremente com o objetivo de condensar a informação central.

Para a avaliação, utilizamos três perspectivas complementares:

- **Métricas automáticas:** ROUGE-1, ROUGE-2, ROUGE-L e BERTScore, que medem a sobreposição ou a similaridade semântica entre o sumário gerado e o sumário de referência.
- **Avaliação por LLM:** GPT-4o foi usado como “avaliador automático”, atribuindo nota de 0 a 10 com justificativa, a partir de um *prompt* simples e do texto-fonte.

Vale ressaltar que, como já discutido neste capítulo, avaliar sumarização abstrativa é um desafio: sumários diferentes podem ser igualmente válidos, e métricas automáticas nem sempre captam nuances de qualidade textual, fidelidade factual ou concisão.

5.5.3.2 Gerando e avaliando sumários

A geração e avaliação dos sumários pode ser feita em poucas linhas, usando bibliotecas como `transformers`, `evaluate` e `bert_score`. Veja exemplos para dois modelos, o PTT5-Summ e o Qwen3-8B:

```
# Sumarização com PTT5-Summ
from transformers import T5Tokenizer, T5ForConditionalGeneration
import torch

# Carrega componentes do modelo pré-treinado para português
tokenizer = T5Tokenizer.from_pretrained('unicamp-dl/ptt5-base-portuguese-vocab')
```

¹³Jupyter notebook para Sumarização Abstrativa com Transformers disponível em: <https://colab.research.google.com/drive/1cWhEWssFanREV94R2fc6N0IOlgkgeATY?usp=sharing>.



```

model = T5ForConditionalGeneration.from_pretrained('recogna-nlp/ptt5-base-summ')

# Prepara o texto para entrada no modelo
input_ids = tokenizer.encode(texto_fonte, return_tensors='pt')

# Gera o resumo com limite de tokens
summary_ids = model.generate(input_ids, max_length=256)

# Converte tokens de volta para texto e remove caracteres especiais
resumo = tokenizer.decode(summary_ids[0], skip_special_tokens=True)

```

Explicação passo a passo do que fizemos no código acima:

1° **Inicialização:** Carregamos o tokenizador e modelo T5 especializado em sumarização em português.

2° **Tokenização:** Convertemos o texto-fonte em representação numérica (`input_ids`).

3° **Geração:** O método `generate` cria o resumo usando arquitetura *encoder-decoder*.

4° **Decodificação:** Transformamos os *tokens* de saída em texto legível.

Agora, vamos ao código do LLM (Qwen3-8B):

```

# Sumarização com um LLM (Qwen3-8B)
from transformers import AutoTokenizer, AutoModelForCausalLM

# Carrega os componentes do LLM
tokenizer = AutoTokenizer.from_pretrained("Qwen/Qwen3-8B")
model = AutoModelForCausalLM.from_pretrained("Qwen/Qwen3-8B")

# Formata o prompt instrucional
prompt = f"Resuma o seguinte texto:\n{texto_fonte}"

# Processa entrada e gera saída
inputs = tokenizer(prompt, return_tensors="pt")
output = model.generate(**inputs, max_new_tokens=256)

# Decodifica a saída
resumo = tokenizer.decode(output[0], skip_special_tokens=True)

```

Vamos analisar passo a passo nosso código:

1° **Configuração:** Usamos o carregamento automático (`AutoTokenizer/AutoModel`) para carregar o *tokenizer* e o modelo de linguagem.

2° **Prompt:** Utilizamos instrução explícita para contextualizar a tarefa.

3° **Geração:** O modelo produz texto, *token* por *token*, em modo auto-regressivo.

Para comparar a qualidade dos resumos produzidos, utilizamos três tipos distintos de avaliação: (i) métricas automáticas baseadas em sobreposição, (ii) similaridade semântica e (iii) julgamento por um LLM. As avaliações também se tornam simples de realizar aproveitando as bibliotecas já existentes do Python.

Nossa primeira métrica será a ROUGE, ela tem como objetivo comparar a sobreposição de n-gramas entre o resumo gerado e um sumário de referência.

```

# Avaliação com ROUGE
import evaluate

# Carrega nosso método de avaliação
rouge = evaluate.load("rouge")

```




```
# Calcula sobreposição de unigramas, bigramas e subsequências mais longas entre resumo
# gerado e referência
results = rouge.compute(predictions=[resumo], references=[resumo_referencia])

# Visualizando o resultado com unigramas e bigramas
print("ROUGE-1:", results["rouge1"], "ROUGE-2:", results["rouge2"], "ROUGE-L:", results["rougeL"])
```

Agora, iremos aplicar a métrica BERTScore que vai além da sobreposição textual, calculando similaridade semântica entre os textos com base em representações vetoriais derivadas de modelos BERT.

```
# Avaliação com BERTScore
from bert_score import score as bert_score

# Compara vetores de embeddings do BERT, capturando equivalência semântica mesmo com
# palavras diferentes.
P, R, F1 = bert_score([resumo], [resumo_referencia], lang="pt", rescale_with_baseline=True)

# Foco no F1, que combina precisão e cobertura semântica.
print("BERTScore (F1):", F1.mean().item())
```

O BERTScore é uma métrica de avaliação de geração de texto que supera limitações das abordagens tradicionais baseadas em sobreposição lexical. Diferentemente da ROUGE (que conta n-gramas idênticos), ela opera no nível semântico usando representações vetoriais de linguagem. Assim, essa métrica é especialmente útil quando o sumário válido é muito diferente em termos de palavras, mas semelhante em conteúdo.

Por fim, empregamos o modelo GPT-4o como avaliador automático. Esse tipo de avaliação, conhecida como *LLM-as-a-judge*, busca aproximar o julgamento humano ao usar um LLM para avaliar outros modelos.

```
# Avaliação com LLM-as-judge
from openai import OpenAI
client = OpenAI()

# Um prompt simples e estruturado foi fornecido ao GPT-4o, solicitando uma nota de 0 a 10
# com justificativa, com base em critérios como coerência, fidelidade factual, completude
# e fluência.
messages = [
    {"role": "system", "content": "Você é um avaliador imparcial de resumos."},
    {"role": "user", "content": f"""Texto original:
    {prompt_text}

    Resumo gerado:
    {summary}

    Dê uma nota de 0 a 10 para este resumo com base em coerência, \
    fidelidade, completude e fluência. Justifique brevemente."""}]

completion = client.chat.completions.create(
    model="gpt-4o",
    messages=messages
)

# A resposta inclui tanto a nota quanto uma explicação, o que permite compreender melhor
```



```
# os pontos fortes e fracos de cada sumário.  
print(completion.choices[0].message.content)
```

A Tabela 5.4 mostra os principais resultados dos experimentos, unindo as três perspectivas de avaliação para dois textos jornalísticos.

Tabela 5.4: Resultado das avaliações dos sumários abstrativos

Modelo	Doc	ROUGE-1	ROUGE-2	BERTScore	Nota GPT-4o	Comentário GPT-4o (resumido)
internlm-chatbode-7b	1	0.133	0.019	0.138	8	Boa coerência, mas poderia ser mais completo.
ptt5-summ	1	0.222	0.093	0.271	3	Resumo superficial e incompleto.
qwen3-8b	1	0.160	0.016	0.136	9	Coerente e fiel, falta mais detalhes.
sabia-7b	1	0.228	0.116	0.054	5	Algumas informações corretas, mas erros e omissões.
internlm-chatbode-7b	2	0.442	0.174	0.468	6	Nome incorreto do presidente do BC; faltam detalhes.
ptt5-summ	2	0.231	0.064	0.309	5	Visão geral, mas faltam detalhes importantes.
qwen3-8b	2	0.413	0.210	0.487	9	Coerente e fiel, faltam detalhes de contexto.
sabia-7b	2	0.265	0.053	0.274	6	Boa introdução, mas incompleto e sem conclusão.

A seguir são apresentados o segundo documento, seu sumário de referência e os sumários candidatos gerados por cada modelo. Os resultados para o primeiro documento estão no Apêndice Apêndice B.

Quadro 5.6: Sumários gerados por cada modelo

Modelo	Texto
Documento2	
Tendência dos juros no Brasil	“A tendência de queda da taxa de juros no Brasil é real, é visível”, disse Meirelles, que participou na capital americana de uma série de reuniões e encontros com banqueiros e investidores que aconteceram paralelamente às reuniões do Fundo Monetário Internacional (FMI) e do Banco Mundial (Bird) no fim de semana. Para o presidente do BC, a atual política econômica do governo e a manutenção da taxa de inflação dentro da meta são fatores que garantem queda na taxa de juros a longo prazo. ... O presidente do Banco Central disse ainda que a crise política brasileira não foi um assunto de interesse prioritário dos investidores que encontrou no fim de semana.

Sumário de referência	Henrique Meirelles afirmou que a tendência de queda dos juros no Brasil é real e sustentada por políticas econômicas e controle da inflação. Ele destacou a importância de uma visão de longo prazo e disse que o trabalho do Banco Central garantirá uma redução gradual dos juros. Meirelles também celebrou o reconhecimento internacional do BC com o prêmio da revista The Banker, reforçando a credibilidade da instituição.
PTT5-Summ:	Meirelles diz que queda de juros no Brasil é real e é visível. Presidente do BC participou de reunião com banqueiros e investidores no fim de semana.
Qwen3-8B:	O presidente do Banco Central do Brasil, Henrique Meirelles, afirmou que a tendência de queda da taxa de juros no país é real e é resultado da política econômica e do controle da inflação. Ele destacou a importância do trabalho do BC, o reconhecimento internacional pela revista The Banker e ressaltou a necessidade de manter uma visão de longo prazo.
InternLM-Chatbode-7B:	O presidente do Banco Central do Brasil, Ilan Goldfajn, afirmou que a tendência de queda da taxa de juros no país é real e visível. Ele atribuiu essa tendência à política econômica do governo e à manutenção da taxa de inflação dentro da meta. Goldfajn também destacou a importância de considerar o médio e longo prazo, e que o trabalho do Banco Central para conter a inflação dentro da meta levará a uma queda gradual da taxa de juros. O Banco Central do Brasil recebeu o prêmio “Banco Central do ano” pela revista The Banker, que reconheceu o trabalho do instituição. Goldfajn também afirmou que a crise política brasileira não foi um assunto de interesse prioritário dos investidores que encontrou no fim de semana.
Sabiá-7B:	A tendência de queda da taxa de juros no Brasil é real, é visível, disse Meirelles, que participou na capital americana de uma série de reuniões e encontros com banqueiros e investidores que aconteceram paralelamente às reuniões do Fundo Monetário Internacional (FMI) e do Banco Mundial (Bird) no fim de semana. Para o presidente do BC, a atual política econômica do governo e a manutenção da taxa de inflação dentro da meta são fatores que garantem queda na taxa de juros a longo prazo. “Mas é importante que nós não olhemos para isso apenas no curto prazo. Temos que olhar no médio e longo prazos”, disse Meirelles. Para ele, o trabalho que o Banco Central tem feito para conter a inflação dentro da meta vai gerar queda gradual da taxa de juros. BC do ano Neste domingo, Meirelles participou da cerimônia de entrega do (texto foi truncado por atingir o limite de <i>tokens</i> da resposta, ajustado em 256 <i>tokens</i>)

Observando os exemplos acima, fica claro que cada modelo tem seu “estilo” ao sumarizar. O PTT5-Summ, por exemplo, gera um texto sintético, buscando ser objetivo e breve, como foi treinado para fazer. Isso faz com que, ao ser avaliado por um LLM (neste caso, o GPT-4o), que recebe apenas o texto-fonte e espera uma cobertura completa das informações, sua nota seja mais baixa. Da mesma forma, métricas como ROUGE e BERTScore penalizam sumários muito curtos porque eles deixam de cobrir n-gramas e conteúdos do sumário de referência.

O Qwen3-8B e o InternLM-Chatbode-7B produziram textos mais próximos em extensão e detalhamento do sumário de referência, sendo melhor avaliados tanto por métricas automáticas quanto pelo GPT-4o – ainda que, no caso do InternLM, tenha ocorrido um erro factual ao trocar o nome do presidente do Banco Central. O nome citado corresponde ao presidente do Banco Central no período de 2016 à 2019, o que sugere uma possível



alucinação causada por viés adquirido nos dados de treinamento, provavelmente mais concentrados nesse período. O mesmo modelo também cometeu um erro de concordância no trecho “do instituição”. O Sabiá-7B, por sua vez, apenas copia o texto, atingindo de forma que acaba rapidamente o limite de *tokens* e terminando o texto abruptamente. Embora tenha recebido as piores avaliações nesse caso, é importante perceber que, apesar de realmente ele ter obtido as piores avaliações para este texto, a essa completa falta de sumarização não foi bem contemplada por nenhuma das medidas. Nesse contexto, a avaliação com LLMs pode ser uma alternativa mais adequada, ao incorporar aspectos como concisão como fator de avaliação, a partir de ajustes no *prompt*.

5.6 A Sumarização automática e a língua portuguesa

Ao analisar as publicações registradas em diferentes repositórios, percebe-se que a SA aplicada ao português brasileiro acompanhou as tendências metodológicas e de aplicações da literatura internacional.

Com relação a alguns **recursos** desenvolvidos, destacamos o TeMário (Pardo; Rino, 2003) por ser um *corpus* de textos jornalísticos acompanhados de seus respectivos sumários, e que inspirou metodologicamente outros *corpora* que foram elaborados para fins de sumarização, como o CSTNews e o RecognaSumm (Paiola et al., 2024). Sendo o CSTNews tido como padrão ouro por apresentar diferentes camadas de anotação linguística, além de sumários humanos que servem de referência para o processo automatizado, enquanto o RecognaSumm se destaca por possuir cerca de 135 mil amostras, abrindo a possibilidade de ajuste-fino de modelos de linguagem. Para apoiar pesquisas de SA de opiniões baseadas em aspectos, há o *corpus* OpiSums-PT¹⁴ (López et al., 2015), que contém vários resumos extrativos e abstrativos de opiniões escritas em português brasileiro, nos quais cada resumo é derivado da análise de 10 opiniões sobre diferentes produtos.

Já com relação a algumas ferramentas, sublinhamos a importância da ferramenta CST-Tool (Aleixo; Pardo, 2008), que utiliza o modelo discursivo *Cross-document Structure Theory* (Radev, 2000)¹⁵ para solucionar desafios linguísticos da sumarização multidocumento. Outra ferramenta tida como precursora por muitos estudiosos da área é o Gistsumm (Balgage Filho et al., 2007; Pardo, 2002), que gera sumários extrativos a partir da identificação do tópico principal dos textos, utilizando métodos superficiais (como palavras-chave).

Cabe destacar que nos últimos anos, com o avanço de modelos de língua e abordagens mais robustas em PLN, a SA passou a apresentar sumários potencialmente com mais qualidade linguística. Nesse sentido, foram encontrados trabalhos que utilizam LLMs (Barros, 2022; Paiola, 2022), apresentando respostas às questões ora não bem resolvidas por abordagens utilizadas no início da SA. Além disso, observaram-se estudos que aplicam a SA em outros domínios para além do jornalístico, como o jurídico (Feltrin et al., 2023) e códigos de programação (Pontes et al., 2022). Aplicações baseadas na dependência de língua e/ou domínio seriam bastante custosas do ponto de vista do PLN, o que pode, em partes, ser resolvido com os LLMs.

Muitos outros trabalhos foram e vêm sendo desenvolvidos em SA, já que se trata de uma área ainda em expansão. Nesse sentido, indicamos que conheçam os trabalhos capitaneados pelo Núcleo Interinstitucional de Linguística Computacional (NILC)¹⁶, que deu suporte

¹⁴<https://sites.icmc.usp.br/taspardo/sucinto/files/OpiSums-PT.zip>

¹⁵Para saber mais detalhes sobre o modelo discursivo CST, confira o Capítulo [Modelos discursivos](#).

¹⁶<https://sites.google.com/view/nilc-usp/>



a muitas pesquisas¹⁷ de diferentes níveis de formação (graduação e pós-graduação). Tais trabalhos utilizaram distintas abordagens que vigoravam como o estado da arte, como SA mono e multidocumento, mono e multilíngue, ora com abordagem mais linguística, ora com abordagem mais estatística. Outra indicação nossa é o Sistema de Informação e Banco de Dados, da Universidade Federal de Campina Grande (SINBAD), que também abriga pesquisas recentes sobre SA. Por fim, outro grupo de destaque é o Grupo de pesquisa em sistemas inteligentes (GSI), da Universidade Tecnológica Federal do Paraná, que tem feito estudos em SA multidocumento e com modelos computacionais mais recentes, como o BERT.

5.7 Considerações finais

Neste capítulo, apresentamos um esboço sobre uma das áreas que tem evoluído no PLN, a SA. Como visto, o processo de automatização de sumários (extrativos ou abstrativos) não é uma tarefa simples. As dificuldades tendem a ser decorrentes de informações faltantes nos textos. Essa falta de dados exige ter-se algum conhecimento prévio do que é informado, de modo que se possa fazer associação com outros fatos relevantes, o que tende a gerar um melhor entendimento do sumário.

Ainda, apresentamos duas abordagens distintas para sumarização automática: a extrativa e a abstrativa. Inicialmente, detalhamos dois algoritmos voltados à sumarização extrativa. O primeiro, de abordagem mais simples, baseia-se na frequência de palavras para selecionar sentenças representativas do texto. O segundo utiliza o modelo BERT e aproveita o poder computacional de redes neurais profundas para realizar tarefas de forma contextualizada e precisa.

Em seguida, apresentamos a sumarização abstrativa, que avança em direção ao uso de modelos generativos capazes de produzir sentenças novas, com maior fluidez e coesão. Porém, essa flexibilidade aumenta os riscos de alucinação e de omissão de informações essenciais. Por isso, a avaliação de sumários abstrativos deve ir além de métricas tradicionais, como ROUGE, e considerar critérios como fidelidade semântica, completude e fluência.

Uma das tarefas que não pode ser negligenciada no processo de sumarização é a avaliação do produto gerado. Assim, pontuamos sobre os métodos de avaliação, com destaque para a ROUGE, além de uma abordagem mais linguística segundo a proposta da TAC/DUC. Cabe ressaltar que aqui também residem outros desafios à SA, já que aplicações de sumarização estão sendo implementadas em conteúdo gerado pelo usuário da Web (como avaliações de produtos e publicações em redes sociais). Nesse caso, concepções linguísticas sobre gênero e suporte textual deverão ser levadas em consideração para que os sumários gerados possam ser informativos, coerentes e coesos, para além de uma perspectiva normativista da língua.

Uma das motivações mais utilizadas para as pesquisas iniciais em SA era a falta de tempo que os usuários (especialmente na internet) tinham frente à quantidade de conteúdo que era constantemente disponibilizado. Atualmente, essa motivação continua viável, mas deve ser acolhida no interior da seguinte reflexão: material gerado por “usuários-autores” pode conter ainda mais enviesamento e incorrer em informações distorcidas ou falsas, podendo resultar em desinformação. Nesse caso, vislumbramos que a SA não poderá se afastar de pesquisas que abordem a língua de uma perspectiva funcional-descritiva, mas que também

¹⁷Indicamos acesso à página do projeto Sucinto, que abrigou diferentes pesquisas na área de SA ligadas ao NILC. Disponível em: <https://sites.icmc.usp.br/taspardo/sucinto/index.html>.



dialoguem com soluções de identificação de conteúdo falso e desinformação.

Diante dessas ponderações, podemos concordar que não esgotamos o tópico de sumarização apenas neste capítulo, especialmente em um contexto em que há uma tendência nas pesquisas em PLN em utilizarem LLMs. Em decorrência da constante evolução da qualidade dos sumários, talvez torne-se necessário a criação de métodos mais apurados que permitam identificar diferenças de estilos de escrita e a identificação de conteúdo relevante. Esse processo de evolução das tecnologias permite imaginar os avanços que podem ocorrer por conta de trabalhos com redes neurais generativas. Fato é que todo esse processo demandará melhorias ou desenvolvimento de novos algoritmos, em especial dos voltados para a língua portuguesa.



Capítulo 6

Complexidade Textual e suas Tarefas Relacionadas

Sidney Evaldo Leal

Felipe Ribas Serras

Marcelo Finger

Sandra Maria Aluísio

Publicado em: 13/03/2024

Atualizado em: 16/04/2026

 <https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/>

6.1 Introdução

Ratificando o redigido na subdivisão da obra acima epigrafada, os originadores da composição a seguir com afincos intentam discorrer sobre o significativo tópico da complexidade textual.

Humm... vamos recomeçar...

Neste capítulo, vamos tentar explicar o tema complexidade textual.

Como acabamos de ver, existem várias formas de dizer a mesma coisa, com graus de complexidade bem diferentes. O tema complexidade textual é largamente tratado em estudos do discurso, na educação, na psicolinguística, na linguística cognitiva, na fonoaudiologia, e no processamento de linguagens naturais (PLN). Aqui apresentaremos conceitos e soluções do ponto de vista do PLN.

Mas antes de tudo é importante dizer que a complexidade é sempre relativa, toda vez que falarmos e ouvirmos falar de complexidade, precisamos perguntar: complexo para quem?

Nesse ponto Dubay (2007) resume de forma certa a definição: “Inteligibilidade é a facilidade de leitura de um texto criada pela escolha de conteúdo, estilo, estruturação e organização que atende ao conhecimento prévio, habilidade de leitura, interesse e motivação da audiência.”

E já que acima aparece o termo **inteligibilidade**, vamos conversar um pouco sobre isso. Inteligibilidade textual vem da tradução do inglês *text readability* e às vezes também é traduzida como leiturabilidade; ambos os termos são bem representativos. Nós optamos por usar inteligibilidade pela relação com entender e dominar a língua (ser letrado) versus a habilidade de decodificar o sistema de escrita (ser alfabetizado). O mais importante é evitar o termo legibilidade, pois este está ligado com o que torna um texto fácil de ser lido (e não necessariamente entendido) como, por exemplo o tamanho e tipo da fonte, cor, estruturação em itens, etc. A inteligibilidade é inversamente correlacionada com a complexidade, isto é, quanto mais complexo um texto for, menos inteligível ele será, para o público alvo do texto.



6.2 Complexo para quem?

A complexidade só existe a partir do ponto de vista específico de quem está lendo, não é possível estudá-la sem o sujeito envolvido no processo da leitura. Mesmo em níveis de letramento próximos, pessoas diferentes podem achar o mesmo texto complexo e simples. Isso varia de acordo com o conhecimento de mundo adquirido e armazenado, a experiência, a habilidade de leitura e o grau de interesse no texto (Dubay, 2007). O Indicador de Alfabetismo Funcional (INAF)¹ é um ótimo retrato geral dos potenciais leitores adultos do Brasil (IPM, 2018). O levantamento é feito em média a cada dois anos, desde 2001 e classifica a população nos seguintes níveis de letramento:

- **Analfabeto:** Não consegue ler;
- **Rudimentar:** Localiza informações explícitas e literais;
- **Elementar:** Realiza pequenas inferências em textos de tamanho médio;
- **Intermediário:** Consegue interpretar textos e confrontar a moral da história com sua própria opinião ou senso comum;
- **Proficiente:** Interpreta e elabora textos de maior complexidade sem dificuldades.

É importante frisar que essa classificação em cinco níveis pode ser considerada relativamente arbitrária e agrupa internamente diversos níveis de letramento. O próprio INAF até 2011 utilizava apenas quatro níveis (Analfabeto, Rudimentar, Básico e Pleno). Mais um nível foi adicionado ao identificar que, após as ações do governo de combate ao analfabetismo, a maioria das pessoas subiu para o nível básico, porém os níveis superiores permaneceram estáveis. De 2001 a 2018, o nível analfabeto caiu de 12% para 8%, enquanto o nível proficiente se manteve em 12%.

Um ponto de atenção é que apesar de ter alfabetismo explícito no nome, o INAF avalia o nível de letramento (ou literacia) da população. Alfabetização está relacionada ao processo mecânico de reconhecer os grafemas, ligando-os aos fonemas, enquanto letramento é o uso social desse processo. Conforme Soares (1996): “Letramento é o resultado da ação de ensinar ou de aprender a ler e escrever: o estado ou a condição que adquire um grupo social ou um indivíduo como consequência de ter-se apropriado da escrita.”

Outro conceito interessante ligado à habilidade de leitura versus motivação é o estado de fluxo (Csikszentmihalyi, 2008). Aplicando ao contexto da leitura, se o texto for demasiado simples e a habilidade do leitor for alta, a experiência vai se tornar enfadonha. Por outro lado, se a habilidade do leitor for pequena demais para o nível de complexidade ou desafio apresentado, o esforço exigido vai ser bastante desmotivador. O estado de fluxo seria o casamento do nível de dificuldade adequado para o nível de proficiência do leitor.

Por que não ensinar a ler em vez de simplificar um texto?

Esta é uma crítica recorrente e muito importante, logo é bom abordá-la aqui. Concordamos plenamente que é sempre melhor ensinar a ler do que simplificar um dado texto original para que ele seja acessível a uma pessoa com dificuldade de entendimento. Dito isso, são citadas a seguir duas grandes exceções para se utilizar a simplificação:

¹<https://alfabetismofuncional.org.br/>

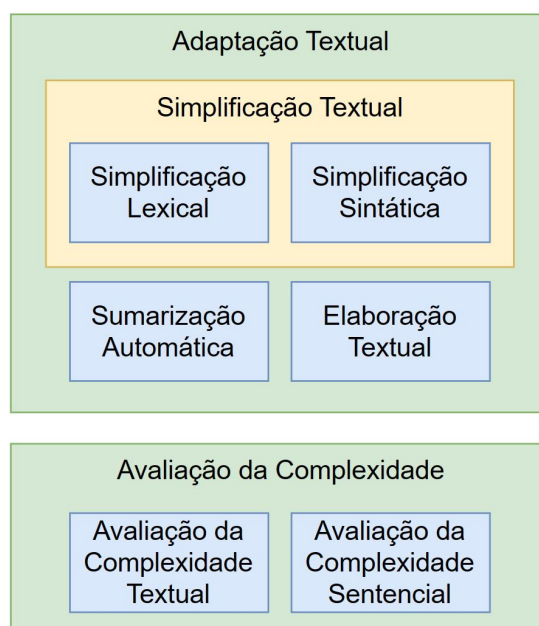


1. **Tempo versus Acesso.** Ensinar a ler exige tempo, enquanto simplificar pode permitir o acesso à informação no momento presente. Isso é uma verdade para a população adulta que possui dificuldades na leitura e, por diversos motivos, menos tempo para investir na própria educação. Além disso, para obter um resultado abrangente o suficiente, o investimento necessário na educação precisa partir do governo. Iniciativas isoladas conseguem bons resultados, mas quantitativamente o acesso à informação é maior simplificando os conteúdos publicados.
2. **Nível ideal de complexidade.** Conforme mencionado no tópico anterior, para um estudante em processo de aprendizagem, ser exposto a um texto demasiadamente difícil pode trazer mais prejuízos do que benefícios. A evolução do processo de ensino-aprendizagem pode ser muito mais eficiente se os textos fornecidos aos estudantes apresentarem um nível de desafio adequado.

6.3 Tarefas de PLN relacionadas à complexidade textual

As principais tarefas de PLN diretamente relacionadas com a complexidade textual estão representadas na Figura 6.1 e resumidas a seguir.

Figura 6.1: Principais tarefas de PLN relacionadas à complexidade textual



6.3.1 Adaptação textual

A Adaptação Textual é uma área de pesquisa de grande importância dentro da área de PLN, geralmente conectada com práticas educacionais, mas também com aplicações bem diversas como, por exemplo, auxiliar na recuperação de informações biomédicas (Jonnalagadda; Gonzalez, 2010). Ela permite alterar o conteúdo de um texto sem mudar seu significado, na maior parte das vezes. Possui duas grandes abordagens: Simplificação e

Elaboração Textual (Aluísio; Gasperin, 2010; Burstein, 2009; Hartmann; Aluísio, 2020; Mayer, 1980).

6.3.2 Simplificação textual

A Simplificação Textual consiste no processo de reduzir a complexidade de um texto, enquanto se preserva o conteúdo informativo e significado, tornando o texto mais fácil de ser compreendido por leitores humanos ou ser processado por programas (Carroll et al., 1998; Max, 2006; Shardlow, 2014; Siddharthan, 2006).

Os primeiros avanços na área de simplificação textual automática surgiram com a ideia de dividir sentenças longas em sentenças menores para melhorar os resultados dos analisadores sintáticos (Chandrasekar et al., 1996). Desde os trabalhos iniciais em simplificação textual, a área prosperou, avançando as pesquisas em cenários de aplicação, línguas e métodos.

Tipos de simplificação textual

Arfé et al. (2018) definem o objetivo da simplificação textual como a adaptação da complexidade do texto (ou *readability*, em inglês) para as habilidades de um determinado grupo de leitores e, desta forma, *readability measures* (ou medidas de complexidade textual) foram desenvolvidas para alinhar/escolher textos para leitores, pois essas medidas podem prever o quão difícil um texto será para seus leitores.

Desta relação entre *readability* e simplificação textual, surgem as abordagens profunda e superficial para *readability*, culminando nas abordagens cognitiva, temática (ou topical) e linguística para simplificação textual, explicadas abaixo.

Segundo as autoras acima, a abordagem superficial para complexidade textual se baseia no tamanho das palavras e sua frequência e no tamanho das orações para prever a complexidade literal dos textos, ou seja, a compreensão do significado estrito de uma única proposição. Enquanto que a abordagem profunda, baseada em *features* como a presença e densidade de marcadores discursivos e correferência no texto, consegue prever coerência e compreensão no nível inferencial, isto é, a integração entre segmentos de um texto e entre o texto e o conhecimento prévio do leitor.

Abaixo são definidos os três tipos de complexidade – a cognitiva, a linguística (envolvendo os níveis lexical e sintático) e a temática – que levam a três abordagens para simplificação textual, de mesmo nome.

A complexidade cognitiva está relacionada com a capacidade limitada de um leitor de identificar e compreender a estrutura global e local de um texto. A estrutura global é responsável por organizar as informações (ou tópicos) de um texto. As estruturas de textos informativos (jornalísticos, por exemplo) são mais variadas do que textos narrativos, podendo ser uma da seguinte lista: descrição, sequência, comparação e contraste, problema-solução e causa-efeito e, inclusive, aparecer de forma não-exclusiva, dificultando o seu reconhecimento. Essa dificuldade pode impedir um leitor de responder o que é dito no texto e de fazer um resumo dele, por exemplo.

A outra dificuldade se dá no processamento local de um texto, realizado pelo leitor, para conectar sentenças e identificar as suas relações (de contraste, exemplificação, causa, resultado, finalidade, dentre outras). As soluções para essas duas dificuldades são dois conjuntos de simplificações, chamadas de cognitiva no nível global e no nível local. No nível local são usadas para:



1. Aumentar a coesão via conectivos para explicitamente mostrar a relação entre sentenças;
2. Utilizar correferência para conectar as ideias.

No nível global, temos simplificações para:

1. Facilitar a retenção de novo conhecimento aprendido do texto via organização do conteúdo textual, ajudando o leitor a identificar a estrutura do discurso pelo uso de marcadores discursivos (linguísticos e tipográficos);
2. Organizar fatos e ideias presentes no texto pelo uso de subtítulos/seções que resumem o conteúdo dos parágrafos.

A complexidade temática/conceitual está associada à falta de conhecimento de mundo necessário para entender alguns temas. Existem projetos como o Newsela² (Xu et al., 2015) que realizam a simplificação conceitual, simplificando os conceitos expressos no texto. Por exemplo, o projeto Newsela inclui elaborações no texto para tornar certos conceitos mais explícitos ou redundâncias para enfatizar partes importantes do texto. Além disso, as operações reduzem e omitem informações que não são adequadas para determinado público-alvo.

Quanto às simplificações linguísticas, temos a lexical e a sintática. A complexidade lexical está relacionada ao desconhecimento do significado de palavras e expressões. A complexidade sintática está relacionada à capacidade ou não de processar alguns tipos de estrutura de sentenças. Na área de PLN, as simplificações linguísticas foram mais exploradas e muitos métodos criados, para várias línguas. Elas são detalhadas nas próximas seções.

6.3.3 Simplificação lexical

A Simplificação Lexical é uma forma de simplificação por meio da substituição de palavras raras ou complexas por hipônimos, hiperônimos ou sinônimos, equivalentes e mais simples, deixando a leitura com compreensão mais fácil para pessoas com baixo letramento (Alúcio; Gasperin, 2010), falantes não nativos de uma dada língua (Paetzold; Specia, 2016c), disléxicos e afásicos (Carroll et al., 1998), dentre outros (Boito, 2014). Um exemplo de sentença simplificada lexicalmente pode ser visto no Quadro 6.1.

A simplificação lexical geralmente é realizada com o apoio de dicionários compilados e recursos como WordNet (Fellbaum, 1998; Miller, 1995), de grandes *corpora* como a Simple Wikipedia (em uma abordagem de *ensembles*), e também de outras abordagens mais recentes baseadas em *word embeddings* (Paetzold; Specia, 2016a) e redes neurais para *ranking* (Paetzold; Specia, 2017).

Quadro 6.1: Exemplo de sentença simplificada lexicalmente

Original	Se acentuada e prolongada, a hipertermia pode causar a morte do animal.

²<https://newsela.com/>



Simplificada	Se acentuada e prolongada, a febre pode causar a morte do animal

6.3.4 Simplificação sintática

A análise sintática é o estudo da disposição das palavras em uma oração e é dividida em funções sintáticas (sujeito e predicado) e constituintes (sintagmas nominais, verbais, preposicionais, adjetivas e adverbiais) (Candido Junior, 2013).

A Simplificação Sintática consiste em dividir orações longas (como exemplificado no Quadro 6.2) ou alterar a estrutura sintática das orações, eliminando fenômenos sintáticos considerados complexos para a inteligibilidade e compreensão de uma classe de leitores. Ainda segundo Candido Junior (2013), alguns exemplos comuns de fenômenos sintáticos são: reordenação de componentes de uma oração para facilitar a compreensão da informação principal veiculada, mudança de voz passiva para ativa, resolução anafórica de pronomes relativos, reordenação de orações e divisão de orações.

Quadro 6.2: Exemplo de sentença simplificada sintaticamente

Original	O uso de forragem conservada, cujas formas mais comuns são: ensilagem e fenação, é uma solução para alimentar o rebanho.
Simplificada	O uso de forragem conservada é uma solução para alimentar o rebanho. As formas mais comuns para conservar forragens são: ensilagem e fenação.

Embora haja um compromisso entre simplificação sintática e aumento do texto – é natural que quebrar uma oração longa em várias torne o texto mais longo, pois sujeitos devem ser adicionados, para vários públicos essa é uma adaptação necessária para permitir o entendimento do texto.

A principal ferramenta de simplificação para o português brasileiro foi desenvolvida durante o projeto PorSimples (Aluísio; Gasperin, 2010), e é chamada Simplifica (Candido Junior et al., 2009; Scarton et al., 2010). Ela apoia autores na redação de textos mais simples, auxiliando tanto na simplificação lexical, que foi baseada em listas de palavras simples, quanto na sintática, realizada via regras baseadas no parser Palavras (Bick, 2000). Atualmente, não está disponível no site do NILC, mas sua interface pode ser vista no relatório do projeto³.

Inicialmente, a tarefa era solucionada por meio de regras fixas programadas, porém a abordagem mais recente utiliza modelos de redes neurais recorrentes inspiradas na tarefa de Tradução Automática (*Machine Translation*), nos quais o texto original é “traduzido” em sua versão simplificada dentro da própria língua, utilizando o conhecimento adquirido com treinamento em grandes *corpora* (Scarton; Specia, 2018).

6.3.5 Sumarização automática

A Sumarização Automática pode ser definida como a diminuição da extensão dos textos mantendo os conteúdos principais. Ela tem um papel muito importante na simplificação de textos, principalmente para os níveis mais baixos de letramento, nos quais o tamanho

³https://fapesp.br/publicacoes/microsoft/microsoft_aluisio.pdf



do texto já é um fator desestimulante para a leitura. Diversos métodos de sumarização, na abordagem extrativa (na qual o sumário é composto de orações retiradas do texto original, sem alterações), foram avaliados no projeto PorSimples (Margarido et al., 2008) e foi escolhido o método Extração de Palavras-Chave por frequências de Radicais (EPC-R) para ser usado na ferramenta Facilita, desenvolvida no mesmo projeto (Watanabe et al., 2009a, 2009b).

Para mais informações sobre a tarefa de Sumarização, confira o Capítulo 5.

6.3.6 Elaboração textual

A Elaboração Textual visa melhorar a compreensão de um texto e ampliar ou explorar o vocabulário do leitor, adicionando informações como: sinônimos ou antônimos ao lado de palavras ou expressões complexas, definição de conceitos ou ainda tornar explícitas as conexões entre as ideias (Aluísio; Gasperin, 2010).

A elaboração lexical, em contraste com a simplificação lexical, não substitui as palavras e sim adiciona uma ou uma lista de palavras para explicar o significado de uma palavra complexa. Também pode inserir uma definição curta conforme exemplificado no Quadro 6.3. Trata-se de uma abordagem adequada em situações como o aprendizado de crianças e de uma segunda língua, pois explica e enriquece o vocabulário do estudante (Hartmann; Aluísio, 2020).

Quadro 6.3: Exemplo de sentença simplificada por elaboração textual

Original	A ensilagem é o processo de conservação do alimento por fermentação anaeróbia.
Simplificada	A ensilagem é o processo de conservação do alimento por fermentação anaeróbia (sem a presença de ar) .

6.4 Avaliação da Complexidade e suas Métricas

As formas de medir automaticamente a complexidade de textos ou sentenças representam por si só uma ampla área de pesquisa, como dito na introdução. A análise automatizada da complexidade de textos, também conhecida em inglês por ARA (*Automatic Readability Assessment*), tem um viés de aplicação prática, pois ajuda a indicar material de leitura adequado, por exemplo para uma dada série escolar, mas também pode contribuir para um melhor entendimento dos processos de leitura e compreensão em populações com processamento típico e atípico de linguagem.

Graesser et al. (2011) dividem as abordagens de predição e medição da complexidade (ou simplicidade) de textos em:

- **Tradicionais:** que usam uma única métrica ou a combinação linear de poucas métricas de dificuldade;
- **Modernas:** que analisam textos com múltiplas características em vários níveis linguísticos e cognitivos, e foram alavancadas por métodos de AM (aprendizado de máquina) nas últimas duas décadas.



Um exemplo da primeira abordagem é o Índice Flesch que será visto na Seção 6.4.1 e outro da segunda abordagem é o Coh-Metrix, apresentado na Seção 6.4.2.

Um dos grandes desafios para a aplicação dos métodos de AM em textos é a criação de *corpora* grandes e balanceados, anotados com as classes de interesse, por professores ou linguistas. O aprendizado do modelo usa a conversão dos textos em valores, geralmente numéricos, para serem usados nas fases de treinamento e avaliação dos métodos. Isso geralmente é obtido por meio da extração e seleção de métricas dos textos, em diversos níveis da língua, para em seguida utilizá-las como *features* nos métodos de aprendizado de máquina.

Há uma crítica frequente a essa abordagem de anotação da complexidade que usa preditores com base em *corpus* com julgamento de especialistas: o fato de a anotação não ser baseada no desempenho real da leitura de estudantes, por exemplo. Entretanto, se já há grande dificuldade em anotar um grande *corpus* com avaliação de professores, conseguir um *corpus* de alunos é mais ainda difícil e demorado (Vajjala; Meurers, 2016). O *corpus Touchstone Applied Science Associates* (TASA), na língua inglesa, é o único grande *corpus* disponível que atende essa crítica, no melhor do nosso conhecimento, por ser formado por 37.520 amostras de textos, com o tamanho de um parágrafo de tamanho médio de 288,6 palavras (desvio padrão de 25,4), cujas dificuldades foram avaliadas via tarefa de leitura de estudantes, sendo anotados também com a métrica DRP (*Degrees of Reading Power*)⁴ (Graesser et al., 2011). Entretanto, a possibilidade de usar rastreamento ocular para capturar o processo de leitura de estudantes é muito bem-vinda e foi explorada na pesquisa de Leal (2021).

Para facilitar a apresentação, são mostradas nas próximas quatro seções as principais fontes de métricas para a tarefa de predição da complexidade (textual e sentencial): fórmulas clássicas, linguísticas, psicolinguísticas e de rastreamento ocular. Dentro de cada seção são descritas as principais métricas citadas na literatura.

6.4.1 Fórmulas clássicas

As primeiras fórmulas para avaliação de inteligibilidade textual surgiram na década de 1920 nos Estados Unidos, e por volta de 1980 já existiam mais de duzentas fórmulas diferentes (Dubay, 2007).

Mesmo com o advento das abordagens mais modernas para resolver a tarefa, essas fórmulas continuam a ter grande importância para as tarefas de PLN, e são usadas isoladamente ou em conjunto com outras *features*. As principais para o escopo deste capítulo são detalhadas a seguir. É importante ressaltar que essas fórmulas foram pensadas para a língua inglesa, portanto não devem ser utilizadas diretamente no português, mas já existem adaptações de algumas delas, como no caso do índice Flesch, por exemplo.

Índice Flesch

A fórmula *Flesch Reading-Ease Score* (FRES), para ser usada com textos em inglês:

$$F = 206.835 - 1.015 \left(\frac{\text{total palavras}}{\text{total sentenças}} \right) - 84.6 \left(\frac{\text{total sílabas}}{\text{total palavras}} \right) \quad (6.1)$$

O valor da fórmula pode ser interpretado com a seguinte escala:

- 90-100: Muito simples,

⁴<http://textcomplexity.questarai.com/getdrp/>



- 80-89: Simples,
- 70-79: Relativamente simples,
- 60-69: Padrão,
- 50-59: Relativamente complexo,
- 30-49: Complexo,
- 0-29: Muito complexo.

É uma das mais antigas e utilizadas fórmulas de inteligibilidade e foi criada por Rudolph Flesch em 1948 (Dell’Orletta et al., 2011; Sjöholm, 2012). Foi adaptada para o português brasileiro em 1996 pelo NILC (Martins et al., 1996), adicionando 42 pontos a todos os escores da fórmula original em inglês:

$$F = 248.835 - 1.015 \left(\frac{\text{total palavras}}{\text{total sentenças}} \right) - 84.6 \left(\frac{\text{total sílabas}}{\text{total palavras}} \right) \quad (6.2)$$

Flesch-Kincaid grade level

A fórmula *Flesch-Kincaid Grade Level* apresenta como resultado um número que corresponde a uma série no sistema educacional americano, facilitando a avaliação do nível de complexidade de livros e textos. Pode ser interpretada como o número de anos de educação necessários para a leitura de um dado texto:

$$FK = 0.39 \left(\frac{\text{total palavras}}{\text{total sentenças}} \right) + 11.8 \left(\frac{\text{total sílabas}}{\text{total palavras}} \right) - 15.59 \quad (6.3)$$

Foi desenvolvida por J. Peter Kincaid em 1975 (Kincaid et al., 1975) a partir da anterior criada por Rudolph Flesch. É também uma função linear que utiliza a média de sílabas por palavras e média de palavras por sentença, estimando assim as complexidades lexical e sintática do texto (Dell’Orletta et al., 2011; Sjöholm, 2012).

Dale-Chall

Inspirada pela fórmula Flesch, a fórmula *Dale-Chall* acrescenta validação da dificuldade das palavras contra um dicionário com 3 mil palavras simples, sendo também considerada a média do tamanho das sentenças:

$$DC = 0.1579 \left(\frac{\text{total palavras difíceis}}{\text{total palavras}} \times 100 \right) + 0.0496 \left(\frac{\text{total palavras}}{\text{total sentenças}} \right) \quad (6.4)$$

Foi criada em 1948 e atualizada posteriormente em 1995 por Edgard Dale e Jeanne Chall (Chall; Dale, 1995; Dell’Orletta et al., 2011).

Gunning Fog Index

Gunning Fog Index ou simplesmente FOG Index foi criada em 1952 por Robert Gunning:

$$GF = 0.4 \left[\left(\frac{\text{total palavras}}{\text{total sentenças}} \right) + 100 \left(\frac{\text{total palavras complexas}}{\text{total palavras}} \right) \right] \quad (6.5)$$

Ao avaliar a dificuldade de inteligibilidade dos jornais por estudantes de graduação, ele escreveu que os textos estavam repletos de incertezas, névoa (*fog* em inglês) e complexidade



desnecessária (Dubay, 2014). Palavras complexas nesse contexto são as que possuem três ou mais sílabas.

Coleman-Liau

Baseada em caracteres em vez de sílabas por palavra, possibilita utilizações mais mecânicas em textos:

$$CLI = 0.0588 \times L - 0.296 \times S - 15.8 \quad (6.6)$$

Na fórmula acima, L é a média da quantidade de letras por 100 palavras e S é a média do número de sentenças por 100 palavras (Coleman; Liau, 1975).

Brunét

O Índice de Brunét é uma variação da TTR (*Type Token Ratio*), mas insensível ao tamanho do texto:

$$W = N^{V^{-0.165}} \quad (6.7)$$

Na fórmula acima N é o número de *tokens* e V é o total de palavras do vocabulário (ou *types*). Foi criado por Étienne Brunet em 1978 (Cunha, 2015; Thomas et al., 2005). Os valores típicos da métrica variam entre 10 e 20, sendo que uma fala mais rica produz valores menores.

Honoré

A Estatística de Honoré é outra variação da TTR, também insensível ao tamanho do texto:

$$R = \frac{100 \log N}{1 - \frac{V_1}{V}} \quad (6.8)$$

Na fórmula acima N é o número de *tokens* e V_1 é o número de palavras do vocabulário que aparecem uma única vez e V é o número de itens lexicais (ou *types*). Foi criada por A. Honoré em 1979 (Cunha, 2015; Thomas et al., 2005), sendo que valores altos da fórmula indicam um vocabulário rico.

6.4.1.1 ALT - Análise de Legibilidade Textual

Com a devida ressalva já feita na introdução sobre o termo legibilidade, o ALT⁵ (Moreno et al., 2022) é uma ferramenta recente bem interessante e publicamente disponível que traz a implementação das fórmulas vistas nesta seção com uma representação visual bem moderna. Na Figura 6.2 podemos ver o resultado da análise para um texto simples.

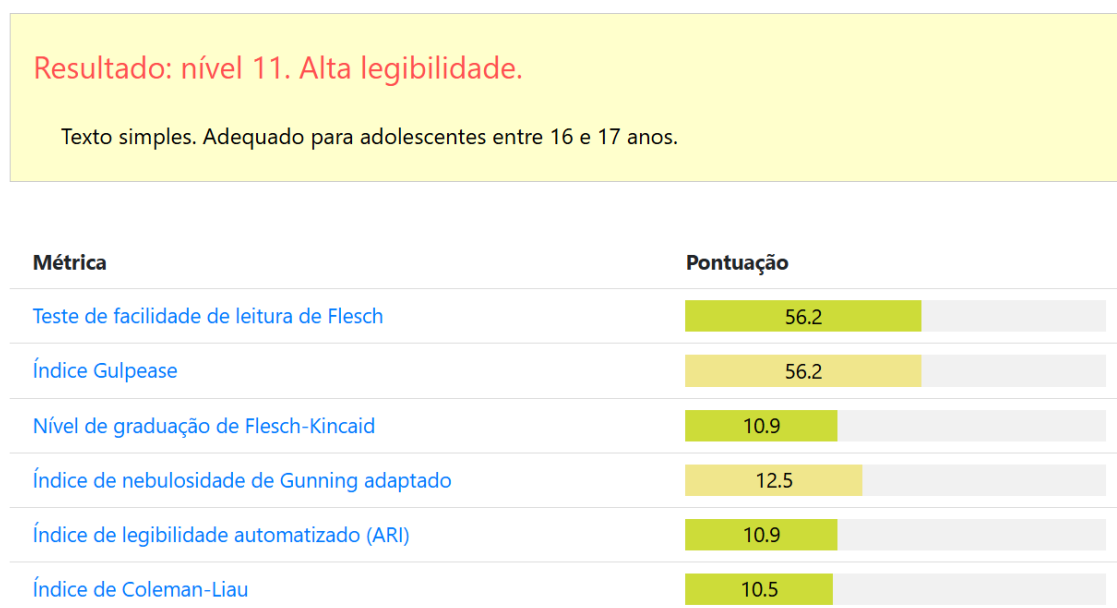
6.4.2 Métricas Linguísticas

As métricas linguísticas extraem características nos níveis lexical, morfossintático, sintático, semântico e discursivo da língua. Existem ferramentas com esse fim específico, que facilitam bastante o processo. Trazemos nas seções seguintes, um resumo de seis ferramentas, tanto para o inglês como para o português.

⁵<https://legibilidade.com/>



Figura 6.2: Tela de resultados da ferramenta ALT



Coh-Metrix

O Coh-Metrix⁶ (McNamara et al., 2014) é uma ferramenta desenvolvida para a língua inglesa, que extrai de um texto métricas de coesão e coerência, permitindo avaliar a complexidade da sua leitura.

Os autores definem coesão como a relação entre as características do texto que guiam o leitor para a representação mental do significado, e coerência é a representação mental que o leitor cria durante a leitura (Graesser et al., 2004).

A versão 3.0 do Coh-Metrix implementa 106 métricas para a língua inglesa, agrupadas nas 11 categorias: *Descriptive*, *Text Easability* *Principal Component Scores*, *Referential Cohesion*, *Latent Semantic Analysis** (LSA), *Lexical Diversity*, *Connectives*, *Situation Model*, *Syntactic Complexity*, *Syntactic Pattern Density*, *Word Information* e *Readability**.

A tela da ferramenta pode ser vista na Figura 6.3), que traz no lado direito os valores de diversas métricas do pequeno texto informado no lado esquerdo.

Em contraste com as fórmulas clássicas que analisam o texto apenas no nível das palavras e sentenças e geram um único valor para quantificar a complexidade do texto, o Coh-Metrix utiliza uma análise multinível, alinhada com teorias de compreensão textual (Graesser et al., 2011):

1. **Words:** Como o conhecimento do vocabulário de uma língua tem um grande impacto sobre o tempo de leitura e compreensão, Coh-Metrix tem uma grande quantidade de métricas relacionadas a palavras, incluindo: análise de categorias gramaticais ou *Part of Speech* (PoS), frequência, medidas psicolinguísticas como concretude, familiaridade, idade de aquisição, imageabilidade, categorias semânticas obtidas da WordNet de Princeton⁷;

⁶<http://cohmetrix.com/>

⁷<https://wordnet.princeton.edu/>

Figura 6.3: Tela do Coh-Metrix com um texto de exemplo

Created: September 1, 2012 **Coh-Metrix 3.0** Last updated: Aug. 16, 2017

Save Data

Enter your input

The methodology aims to develop print and electronic contents culturally contextualized, adapted and available according to the literacy level of farmers. This methodology adopts Human Computer Interaction and Natural Language Processing approaches, providing the lexical and syntactic simplification, using analogies and family vocabulary. Studies were carried out with experts, extension workers, students and farmers dedicated to milk production in order to verify the applicability of the methodology in a real scenario. Using this tool, enables the creation of content tailored for different levels of literacy. In doing so, farmers are able to understand the technical knowledge and consequently adopt the technologies offered and recommended to improve the quality and productivity of their respective production systems.

195M

Type text in the image

Submit

Number	Label	Label V2.x	Text	Full description
Descriptive				
1	DESPC	READNP	3	Paragraph count, number of paragraphs
2	DESSC	READNS	5	Sentence count, number of sentences
3	DESWC	READNW	115	Word count, number of words
4	DESPL	READAPL	1.667	Paragraph length, number of sentences in a paragraph, mean
5	DESPLd	n/a	0.577	Paragraph length, number of sentences in a paragraph, standard deviation
6	DESSL	READASL	23	Sentence length, number of words, mean
7	DESSLd	n/a	6.325	Sentence length, number of words, standard deviation
8	DESWLsy	READASW	2.191	Word length, number of syllables, mean
9	DESWLsyd	n/a	1.290	Word length, number of syllables, standard deviation
10	DESWLlt	n/a	6.113	Word length, number of letters, mean
11	DESWLltd	n/a	3.236	Word length, number of letters, standard deviation
Text Easability Principle Component Scores				
12	PCNARz	n/a	-1.291	Text Easability PC Narrativity, z score
13	PCNARp	n/a	9.850	Text Easability PC Narrativity, percentile
14	PCSYNZ	n/a	0.203	Text Easability PC Syntactic simplicity, z score
15	PCSYNp	n/a	57.930	Text Easability PC Syntactic simplicity, percentile
16	PCCNCz	n/a	-0.185	Text Easability PC Word concreteness, z score
17	PCCNCp	n/a	42.860	Text Easability PC Word concreteness, percentile
18	PCREFz	n/a	-0.584	Text Easability PC Referential cohesion, z score
19	PCREFp	n/a	28.100	Text Easability PC Referential cohesion, percentile
20	PCDCz	n/a	0.854	Text Easability PC Deep cohesion, z score
21	PCDCp	n/a	80.230	Text Easability PC Deep cohesion, percentile
22	PCVERBz	n/a	-2.346	Text Easability PC Verb cohesion, z score
23	PCVERBp	n/a	0.960	Text Easability PC Verb cohesion, percentile
24	PCCONNZ	n/a	-3.163	Text Easability PC Connectivity, z score
25	PCCONNp	n/a	0.080	Text Easability PC Connectivity, percentile
26	PCTEMPz	n/a	-0.702	Text Easability PC Temporality, z score
27	PCTEMPp	n/a	24.200	Text Easability PC Temporality, percentile
Referential Cohesion				
28	CRFNO1	CRFBN1um	0.5	Noun overlap, adjacent sentences, binary, mean
29	CRFAO1	CRFBA1um	0.5	Argument overlap, adjacent sentences, binary, mean
30	CRFSO1	CRFBS1um	0.5	Stem overlap, adjacent sentences, binary, mean
31	CRFNOa	CRFBNaum	0.600	Noun overlap, all sentences, binary, mean

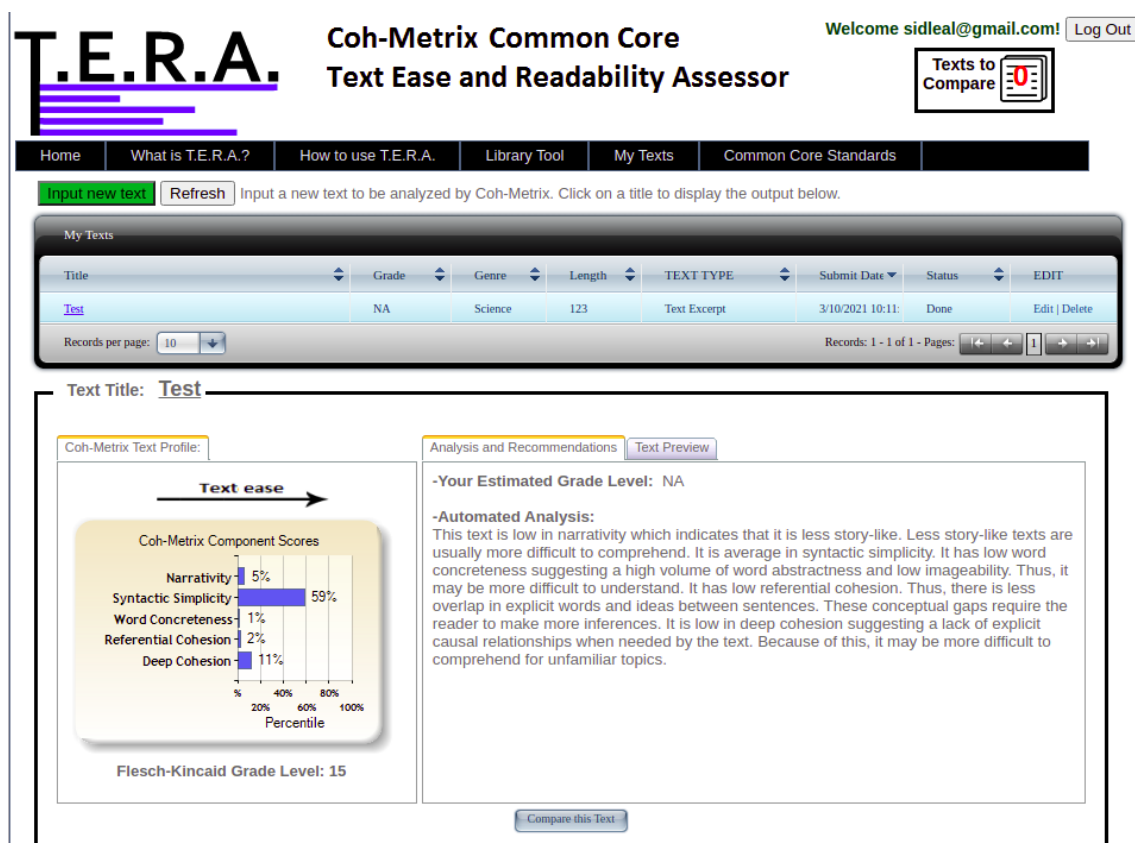
2. **Syntax:** Algumas sentenças do discurso oral são curtas, apresentam poucas orações relativas, poucas palavras nos sintagmas nominais e se apresentam na voz ativa, mas sentenças de textos escritos geralmente aparecem de forma oposta, demandando mais processamento da memória de trabalho. Coh-Metrix computa essas contagens e outras como similaridade de pares de sentenças adjacentes, que facilitam a leitura e compreensão;
3. **Textbase:** A base textual está relacionada com o significado em vez da análise de palavras e da sintaxe. A correferência é um mecanismo importante para conectar as proposições, as orações e sentenças na base textual, assim Coh-Metrix traz várias métricas para o cômputo da correferência como a sobreposição de palavras de conteúdo, de substantivos e de radicais (*content word overlap*, *noun overlap*, e *stem overlap*, respectivamente). A diversidade lexical está relacionada com a coesão porque um número elevado de palavras diferentes em um texto significa que as palavras novas precisam ser integradas no contexto do discurso. Coh-Metrix também computa várias métricas relacionadas com o modelo estatístico para cálculo de similaridade chamado *Latent Semantic Analysis* (LSA), pois ele ajuda a medir o conhecimento implícito do leitor em adição às palavras explícitas usadas no texto;
4. **Situation Model / Mental Model:** Textos narrativos incluem pessoas, objetos, ações, eventos, processos, planos e outros detalhes de uma história, já em textos informativos o modelo mental é diferente, pois devem ajudar a entender como modelos da física, biologia e outras ciências funcionam. Assim, há métricas para avaliar se há quebras no entendimento desses modelos mentais que emergem de um texto;

5. **Genre and Rhetorical Structure:** Exemplos de uma tipologia de gêneros são: narrativo, expositivo, persuasivo ou descritivo. Textos narrativos são mais fáceis de se ler, compreender e relembrar do que textos informativos. Coh-Metrix analisa se um texto pode ser classificado como narrativo ou informativo, via uma métrica chamada narratividade.

T.E.R.A

T.E.R.A.⁸ (*Text Ease and Readability Assessor*) é uma ferramenta construída pelos mesmos autores do Coh-Metrix, também para língua inglesa. Utiliza o Coh-Metrix para avaliar amostras dos textos, reduzindo as métricas em cinco fatores, levantados via PCA (*Principal Component Analysis*) ((Graesser et al., 2011; McNamara et al., 2013)): *Narrativity*, *Syntactic Simplicity*, *Word Concreteness*, *Referential Cohesion* (*Textbase*) e *Deep Cohesion* (*Situation Model*). Na Figura 6.4 é possível ver um exemplo da análise do texto com as cinco dimensões.

Figura 6.4: Tela de exemplo do Coh-Metrix-T.E.R.A.



Coh-Metrix-Port

O Coh-Metrix-Port⁹ (Scarton; Aluísio, 2010) é uma adaptação para o português brasileiro do Coh-Metrix, desenvolvida dentro do projeto PorSimples (Simplificação Textual do Por-

⁸<http://www.commoncoretera.com/>

⁹<http://fw.nilc.icmc.usp.br:22680/>



Figura 6.5: Tela de avaliação do Coh-Metrix-Port 2

The screenshot shows the Coh-Metrix-Port 2.0 web application. A modal window titled 'Submit a text' is open, allowing a user to input text for analysis. The form contains the following fields:

- Title:** Catálogo de Forrageiras Recomendadas pela Embrapa
- Author:** A. V. Pereira, S. C. Paciullo, C. A. M. Gomide, F. J. S. Léo
- Source:** <https://www.embrapa.br/busca-de-publicacoes/-/publicacao/1048272/catalogo-de-forrageiras-recomendadas-pe>
- Date:** 02/25/2018
- Genre:** Informativo
- Content:** A alimentação é o principal componente de custos na produção de leite, correspondendo a um percentual mínimo de 30% do custo total nos sistemas de produção estudados. Por outro lado, a maioria absoluta das propriedades leiteiras no Brasil está estruturada em torno de alimentação baseada em forrageiras. Para a Embrapa, estas duas características demonstram que é estratégico investir no melhoramento de forrageiras que apresentem maior produtividade, que sejam mais resistentes e adaptadas a cada um dos biomas específicos e que tenham manejo compatível com a realidade de cada processo produtivo. Reduzir custos de sistemas de produção e assegurar que sejam sustentáveis são, portanto, os objetivos perseguidos permanentemente pelo corpo técnico engajado na pesquisa de forrageiras da Embrapa. Para esse conjunto de pesquisadores e analistas é certo que os resultados de pesquisa, tecnologia e inovação

Buttons at the bottom of the form: Clear, Close, Submit text.

tuguês para Inclusão e Acessibilidade Digital), que teve como objetivo promover o acesso a textos da Web a pessoas com baixo letramento.

O Coh-Metrix-Port implementa 48 métricas específicas para o português brasileiro (Scarton et al., 2010), divididas nas categorias: contagens básicas, operadores lógicos, frequências, hiperônimos, tokens, constituintes, conectivos, ambiguidade, correferência e anáforas. A tela de cadastro dos textos, da versão 2.0¹⁰, pode ser vista na Figura 6.5.

Coh-Metrix-Dementia

O Coh-Metrix-Dementia¹¹ (Cunha, 2015) é uma adaptação do Coh-Metrix-Port para análise automática de distúrbios de linguagem nas demências (como Doença de Alzheimer) ou no Comprometimento Cognitivo Leve (CCL).

Ele adiciona 25 novas métricas às 48 do Coh-Metrix-Port, nas categorias: disfluências, análise de semântica latente, diversidade lexical, complexidade sintática e densidade semântica.

Disponibiliza no total 73 métricas para o português brasileiro. Sua tela principal pode ser vista na Figura 6.6.

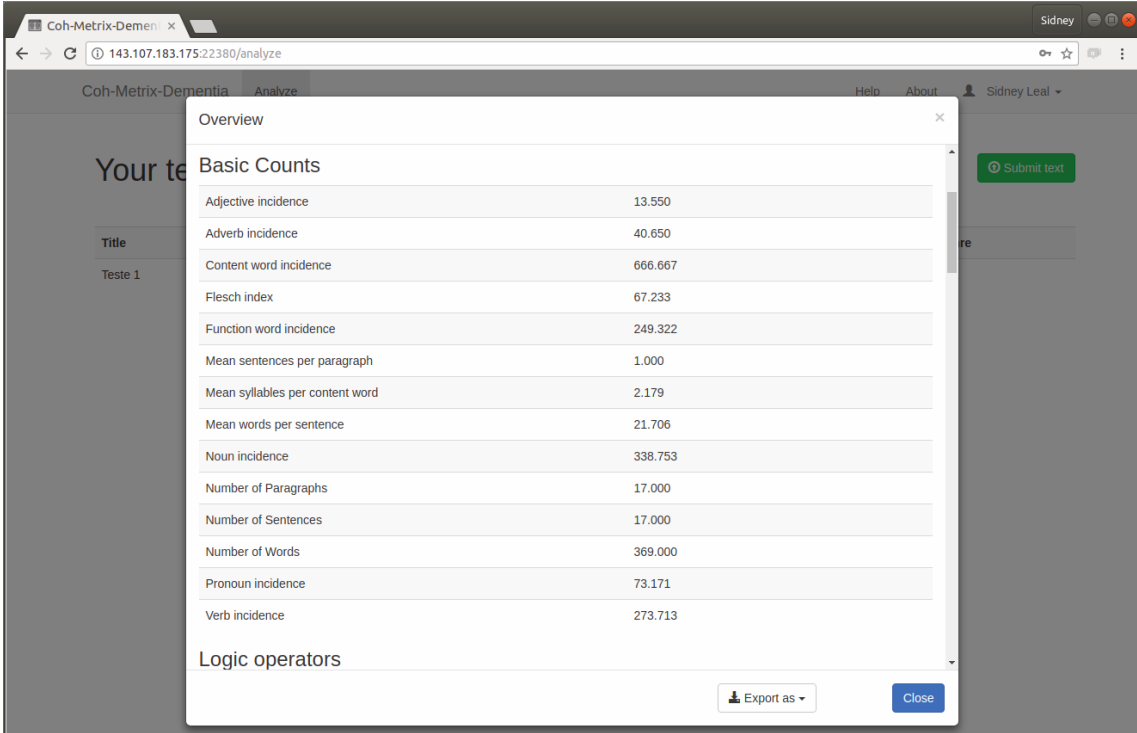
LIWC

LIWC (*Linguistic Inquiry and Word Count*) é uma ferramenta baseada em dicionários para análise dos vários componentes emocionais, cognitivos e linguísticos em amostras de textos (Pennebaker et al., 2015), com categorias como: estatísticas comuns do texto,

¹⁰Refeita por Cunha (2015)

¹¹<http://fw.nilc.icmc.usp.br:22380/>

Figura 6.6: Exemplo da tela de saída do Coh-Metrix-Dementia



The screenshot shows the Coh-Metrix-Dementia web application interface. A modal window titled 'Overview' is open, displaying a table of linguistic metrics. The table is organized into two sections: 'Basic Counts' and 'Logic operators'. The 'Basic Counts' section lists 15 metrics with their corresponding values. The 'Logic operators' section is currently empty. At the bottom of the modal, there is an 'Export as' button and a 'Close' button.

Basic Counts	
Adjective incidence	13.550
Adverb incidence	40.650
Content word incidence	666.667
Flesch index	67.233
Function word incidence	249.322
Mean sentences per paragraph	1.000
Mean syllables per content word	2.179
Mean words per sentence	21.706
Noun incidence	338.753
Number of Paragraphs	17.000
Number of Sentences	17.000
Number of Words	369.000
Pronoun incidence	73.171
Verb incidence	273.713
Logic operators	

dimensão linguística, processos psicológicos, relatividade, assuntos pessoais e miscelânea, totalizando aproximadamente 100 métricas (Cunha, 2015).

A sua primeira versão foi criada em 1993, a segunda em 2001, a terceira em 2007 e a última em 2015. O dicionário da versão inglesa conta com 6400 palavras, radicais e emoticons (Pennebaker et al., 2015).

A tradução e adaptação do dicionário para o português brasileiro foi realizada em uma colaboração entre NILC, Checon Pesquisa e Unisinos no período de 2010 a 2012 e está disponível no site do projeto PortLex¹².

AIC

Também criada dentro do contexto do PorSimples (Maziero et al., 2008), a ferramenta AIC (Análise Automática de Inteligibilidade de *Corpus*) traz 39 métricas, com o principal diferencial de utilizar o analisador sintático PALAVRAS (Bick, 2000) para o cálculo delas. Elas estão organizadas em seis classes: estatísticas do texto, voz passiva, características das orações, densidade, personalização e marcadores discursivos (Cunha, 2015; Reis, 2017). A tela de saída pode ser vista na Figura 6.7.

6.4.3 Métricas Psicolinguísticas

As palavras possuem algumas propriedades subjetivas estudadas pela psicolinguística como: imageabilidade, concretude, familiaridade e idade de aquisição (Santos et al., 2017a), detalhadas abaixo:

¹²<http://nilc.icmc.usp.br/portlex/index.php/pt/projetos/liwc>

Figura 6.7: Exemplo da tela de saída da AIC, atualmente não disponível no site do NILC

Tabela 1 - Estatísticas

N. de caracteres: **5216**
 N. médio de caracteres por palavra: **4.66965085049239**
 N. de palavras: **1117**
 N. médio de palavras por sentença: **16.9242424242424**
 N. de sentenças: **66**
 N. de palavras presentes no Dicionário da Biderman: **982 (87.9140555058192%)**

Tabela 2 - Voz Passiva

N. de sentenças na voz passiva: **4 (6.06060606060606%)**

Tabela 3 - Orações

N. de orações (cláusulas): **200 - Verbos (exceto auxiliares)**
 N. de sentenças que iniciam com conjunções subordinadas: **0 (0%)**
 N. de sentenças que iniciam com conjunções coordenadas: **2 (3.03030303030303%)**
 Conjunções que iniciam as cláusulas coordenadas: Mas, E,
 Sentenças com ...
 0 cláusula(s): **2 (3.03030303030303%)**
 1 cláusula(s): **11 (16.6666666666667%)**
 2 cláusula(s): **16 (24.2424242424242%)**
 3 cláusula(s): **15 (22.7272727272727%)**
 4 cláusula(s): **9 (13.6363636363636%)**
 5 cláusula(s): **9 (13.6363636363636%)**
 6 cláusula(s): **1 (1.51515151515152%)**
 7 cláusula(s): **1 (1.51515151515152%)**

Fonte: (Maziero et al., 2008)

- **Imageabilidade:** Envolve a facilidade e rapidez de evocar uma imagem mental da palavra.
- **Concretude:** É o grau com que uma palavra se refere a objetos, pessoas, lugares ou coisas que podem ser percebidas pelos sentidos, em contraste com os conceitos abstratos.
- **Familiaridade:** É o grau com que as pessoas conhecem e usam palavras no dia a dia.
- **Idade de Aquisição:** Estimativa da idade em que uma palavra foi aprendida, calculada via análise feita por adultos.

Essas propriedades têm um grande impacto na complexidade dos textos e sentenças, e trazem melhorias aos resultados de várias tarefas de PLN, como simplificação lexical e tarefas de classificação semântica quando utilizadas em conjunto com as demais métricas (Paetzold; Specia, 2016b).



Santos et al. (2017a) anotaram automaticamente essas métricas em um banco¹³ de 26.874 palavras do português brasileiro utilizando um método baseado em regressão e *Multi-View Learning*, com recursos fáceis de se obter em várias línguas.

6.4.4 NILC-Metrix

NILC-Metrix¹⁴ é um conjunto de 200 métricas desenvolvido no Centro Interinstitucional de Linguística Computacional – NILC¹⁵, do final de 2007 ao início de 2021 (Leal et al., 2023).

O principal objetivo dessas métricas é fornecer subsídios para avaliar a coesão, a coerência e a complexidade textual. O NILC-Metrix pode ajudar pesquisadores a investigar várias questões de pesquisa, por exemplo:

1. como as características do texto se correlacionam com a compreensão da leitura;
2. quais são as características mais desafiadoras de um determinado texto, ou seja, quais características tornam um texto ou corpus mais complexo;
3. quais textos possuem as características mais adequadas para desenvolver as habilidades dos alunos-alvo;
4. quais partes de um texto são desproporcionalmente complexas e devem ser simplificadas para atender a um determinado público.

Além de ser disponibilizado via interface web, o código fonte também foi disponibilizado sob licença AGPLv3¹⁶, facilitando a incorporação do código das métricas em aplicações. As métricas são agrupadas em 14 categorias, seguindo sua similaridade e fundamentação teórica. São elas:

1. Medidas Descritivas (10 métricas)
2. Simplicidade Textual (8 métricas)
3. Coesão Referencial (9 métricas)
4. Coesão Semântica via LSA (11 métricas)
5. Medidas Psicolinguísticas (24 métricas)
6. Diversidade Lexical (15 métricas)
7. Conectivos (12 métricas)
8. Léxico Temporal (12 métricas)
9. Complexidade Sintática (27 métricas)
10. Densidade de Padrões Sintáticos (4 métricas)

¹³A base está disponível em: <http://143.107.183.175:21380/portlex/index.php/en/component/content/article/2-uncategorised/23-psycholinguistic>

¹⁴<http://fw.nilc.icmc.usp.br:23380/nilcmetrix>

¹⁵<http://www.nilc.icmc.usp.br/>

¹⁶<https://github.com/nilc-nlp/nilcmetrix>



11. Informações Morfossintáticas de Palavras (11 métricas)
12. Informações Semânticas de Palavras (11 métricas)
13. Frequência de Palavras (10 métricas)
14. Índices de Leiturabilidade (5 métricas)

6.4.5 Rastreamento ocular

As métricas do rastreamento ocular trazem uma abordagem recente e diferente das métricas mostradas anteriormente. Sua contribuição é bastante relevante, uma vez que permitem uma aproximação da percepção mais realista da complexidade pelos leitores. Elas foram fundamentais no trabalho de avaliação da complexidade no nível sentencial para o português brasileiro (Leal et al., 2020).

Os movimentos dos olhos podem ser interpretados como uma janela para o processamento do cérebro, refletindo os tempos cognitivos envolvidos em determinada tarefa. Por exemplo, durante a leitura os movimentos dos olhos são controlados por uma complexa interação entre os fatores de baixo nível (por exemplo, o quanto o olho consegue ver e interpretar a cada fixação) e de alto nível (por exemplo, o processamento sintático) (Barrett et al., 2015).

Rayner (1998) divide a pesquisa sobre os movimentos dos olhos (ou rastreamento ocular) em três grandes eras. A primeira era vai desde as primeiras observações sobre os movimentos dos olhos durante a leitura em 1879 até os anos 1920. Algumas importantes descobertas foram feitas nessa era como, por exemplo, o fato de que não percebemos nenhuma informação durante o reposicionamento do olhar, denominado sacada ou *saccade*, em inglês.

A segunda era coincide com o movimento behaviorista na psicologia experimental, com os trabalhos com focos mais práticos – estudos dos movimentos dos olhos em si ou em aspectos superficiais da tarefa investigada – e menos concentrados na utilização dos movimentos para inferir o processamento cognitivo.

Figura 6.8: Rastreamento ocular de plataforma, com óculos simples, e com óculos especial de realidade virtual



Fonte: (Fove, 2018; Imotions, 2017)

A terceira era começa em meados dos anos 1970, com melhorias nos sistemas de rastreamento que permitiram medidas mais acuradas e simples de obter (Figura 6.8). Nessa era, juntamente com os avanços das teorias de processamento da linguagem, os movimentos dos olhos começaram a ser utilizados para exame crítico dos processos cognitivos durante a leitura.

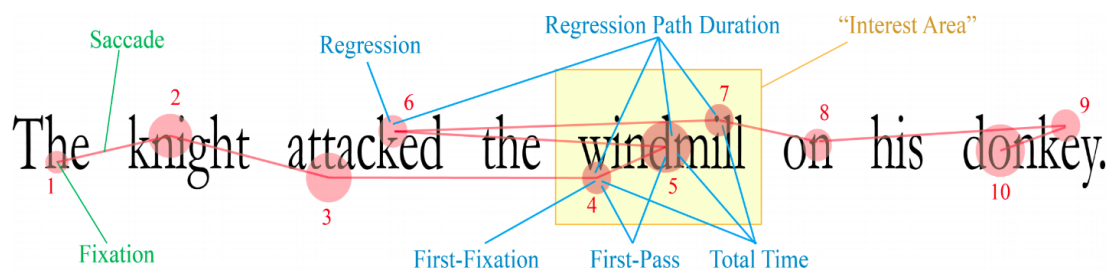


No português brasileiro, o rastreamento ocular já é utilizado há algum tempo na área da psicolinguística. Por exemplo, Maia et al. (2007) utilizaram para investigar o papel do processamento morfológico na identificação de palavras, Leitão et al. (2012) utilizaram na investigação do processamento anafórico e Teixeira et al. (2014) para evidenciar o custo de resolução de pronomes nulos e plenos.

As características básicas dos movimentos dos olhos são:

- **Sacadas (*Saccades* em inglês):** Os contínuos movimentos oculares, o reposicionamento do olhar (durante uma sacada nenhuma informação é percebida).
- **Fixações (*Fixations* em inglês):** Os tempos de fixação em um ponto de atenção entre as sacadas.

Figura 6.9: Principais métricas de rastreamento ocular



Fonte: (Zelenina, 2015)

A partir dessas duas características é possível medir diversas outras informações relevantes para o processo de leitura e interpretação de textos. As principais métricas obtidas pelo movimento dos olhos são exemplificadas na 8, com a simulação do caminho feito pelo olhar em dez fixações numeradas sequencialmente, e detalhadas a seguir:

- ***First fixation duration:*** Tempo da primeira fixação na palavra.
- ***First pass fixation duration:*** Quando uma palavra é longa, pode ser necessário um segundo ponto de fixação dentro da própria palavra. Essa métrica é a soma dos tempos das fixações na primeira passada pela palavra.
- ***Total fixation duration:*** Soma de todos os tempos de fixação na palavra.
- ***Average fixation duration:*** Tempo médio de fixação, quando se tem mais de um ponto por palavra ou média por sentença.
- ***Regression:*** Regressões no texto podem indicar necessidade de rever alguma informação para entendimento do ponto atual, por exemplo, para resolver uma coreferência. É uma métrica muito importante para medir a complexidade textual e sentencial.
- ***Regression path duration:*** Mede a extensão da regressão; quanto maior a regressão, maior o esforço despendido para a leitura, como resultado de um texto mais complexo.

- **Interest area:** Pontos de interesse, onde o leitor passou mais tempo fixando no texto. Calculado com a soma de todas as fixações.
- **Skipping rate:** Algumas palavras são naturalmente saltadas durante a leitura, como artigos e preposições. Não saltar essas palavras pode indicar um leitor com menor proficiência na leitura.
- **Number of fixations:** Quantidade de fixações na palavra; uma palavra simples só deve exigir uma única fixação.
- **Second pass fixation duration:** Tempo de fixação na segunda vez que o leitor retorna à palavra.
- **Spillover from previous word:** Nem sempre o processamento de uma palavra é completado antes que o olhar se mova para a próxima. Nesses casos ocorre o efeito de “transbordamento” do tempo para a palavra seguinte.

6.5 Recursos para Complexidade Textual e Sentencial no Português Brasileiro

Até onde foi possível verificar, existem poucos recursos disponíveis publicamente para as tarefas de complexidade no nível textual e sentencial para o português brasileiro. A seguir são apresentados resumos dos recursos e links para download, sempre que possível.

6.5.1 Recursos para Complexidade Textual

Nesta seção, consideramos os recursos com textos para a tarefa. O PorSimples traz textos simplificados em dois níveis, o RastrOS disponibiliza dados de rastreamento ocular em textos curtos e os *corpora* do projeto PorPopular são compilações representativas do português popular em jornais como o Diário Gaúcho¹⁷ e Massa¹⁸.

PorSimples

Corpus paralelo de textos originais e simplificados criado em 2009 no projeto PorSimples¹⁹ (Simplificação Textual do Português para Inclusão e Acessibilidade Digital) do NILC ((Casseli et al., 2009)).

Um editor²⁰ foi desenvolvido para a tarefa de anotação, cuja tela principal pode ser vista na Figura 6.10. No lado esquerdo, fica o texto original e à direita, sua versão simplificada. Os textos jornalísticos foram simplificados em dois níveis por especialistas linguistas:

- **Natural:** textos para os quais o anotador escolheu livremente as operações de simplificação, inclusive podendo escolher não simplificar uma sentença.
- **Forte:** Anotadores seguiram o manual de simplificação também desenvolvido no projeto.

¹⁷<https://diariogauchoclicrbs.com.br/>

¹⁸<https://jornalmassa.com.br/>

¹⁹<http://www.nilc.icmc.usp.br/nilc/index.php/porsimples>

²⁰Após mais de uma década, o editor continua disponível e funcionando, pode ser utilizado em: <http://fw.nilc.icmc.usp.br:23080/>



Figura 6.10: Exemplo da tela do editor de anotação do PorSimples



Fonte: (Caseli et al., 2009)

A primeira fase do *corpus* foi criada a partir de 104 textos do jornal Zero Hora. Na segunda fase, foram adicionados 50 textos do Caderno de Ciência do jornal Folha de São Paulo²¹, resultando em 154 trios alinhados, com o total de 462 textos e mais de 185 mil tokens.

Na Tabela 6.1 podem ser vistos os números de sentenças de cada nível, e para os tokens confira a Tabela 6.2. Uma característica importante do *corpus* PorSimples foi a anotação dos fenômenos linguísticos nas sentenças; uma extração deles pode ser vista para cada nível na Tabela 6.3. Com essa informação é possível verificar que alguns fenômenos adicionam mais complexidade que outros, por exemplo, as orações apositivas foram as que mais diminuíram em número durante o processo de simplificação, já as subordinadas contraintuitivamente aumentaram nos níveis mais simples.

Tabela 6.1: PorSimples - Estatísticas de Sentenças

	Total	Mínimo/Texto	Máximo/Texto	Média/Texto
Original	2.985	5	46	19
Natural	4.080	5	62	26
Forte	4.974	7	72	32

²¹<https://www1.folha.uol.com.br/fsp/>

Tabela 6.2: PorSimples - Estatísticas de Tokens

	Total	Mínimo/Sentença	Máximo/Sentença	Média/Sentença
Original	61.026	2	71	21
Natural	61.754	2	60	15
Forte	63.030	2	47	13

Tabela 6.3: PorSimples - Fenômenos Linguísticos

	Coordenadas	Subordinadas	Relativas	Passivas	Apositivas
Original	1.443	805	897	319	306
Natural	1.352	899	759	257	105
Forte	1.210	876	527	167	73

Um ponto muito importante é que esse *corpus* possui alinhamento no nível sentencial. Esse alinhamento foi utilizado para gerar o *corpus* PorSimplesSent, detalhado mais adiante.

Corpus de Complexidade Textual para Estágios Escolares do Sistema Educacional Brasileiro

Gazzola et al. (2019) compilaram um grande *corpus* com textos utilizados em diferentes etapas de ensino do Sistema Educacional Brasileiro. O *corpus* foi organizado nas quatro etapas utilizadas na Plataforma MEC de Recursos Educacionais Digitais (MEC-RED)²² para classificação nos estágios escolares:

- Ensino Fundamental I (primeiro ao quinto ano);
- Ensino Fundamental II (sexto ao nono ano);
- Ensino Médio;
- Ensino Superior.

O *corpus* está publicamente disponível²³ e inclui livros-texto, notícias da seção Para Seu Filho Ler (PSFL) do jornal Zero Hora (que apresenta algumas notícias sobre os mesmos tópicos do jornal, mas escritas para crianças de 8 a 11 anos de idade), Exames do SAEB, Livros Digitais do Wikilivros em Português e Exames do Enem dos anos 2015, 2016 e 2017.

Em números, o *corpus* disponibiliza 2.067 documentos (min = 300 palavras, max = 596 palavras, média = 448).

CorPop e PorPopular

O CorPop (*corpus* de referência do português popular escrito do Brasil) foi criado em 2018 por Pasqualini (2018) durante seu doutorado. Ele traz uma compilação bem avaliada de textos selecionados com base no nível de letramento médio dos leitores do país, das seguintes fontes:

²²<https://plataformaintegrada.mec.gov.br/>

²³<http://www.nilc.icmc.usp.br/nilc/index.php/tools-and-resources> e https://github.com/gazzola/corpus_readability_nlp_portuguese



1. Textos do jornalismo popular do Projeto PorPopular (jornal Diário Gaúcho) consumido maciçamente pelas classes C e D;
2. Textos e autores mais lidos pelos respondentes das últimas edições da pesquisa “Retratos da Leitura no Brasil”;
3. Coleção “É Só o Começo” (adaptação de clássicos da literatura brasileira para leitores com baixo letramento, realizada por linguistas);
4. Textos do jornal Boca de Rua²⁴, produzido por pessoas em situação de rua, com baixa escolaridade e baixo letramento;
5. Textos do Diário da Causa Operária²⁵, imprensa operária brasileira produzida também por pessoas dentro da faixa média de letramento do país.

O *corpus* possui 684 mil tokens. Está parcialmente disponível publicamente²⁶ (ferramentas e listas de palavras).

Além do CorPop, o projeto PorPopular²⁷ (Finatto, 2012) disponibiliza um *corpus* com amostras do jornal Diário Gaúcho e outro com amostras do jornal Massa.

MedSimples

A MedSimples²⁸ é uma ferramenta que auxilia na simplificação de textos sobre temas de saúde (Paraguassu et al., 2020). Inicialmente se baseou em terminologias e palavras potencialmente difíceis extraídas de um *corpus* sobre a doença de Parkinson, depois foram adicionados um *corpus* sobre oncologia e também a base do CorPop (Villar; Finatto, 2023).

Na ferramenta foi trabalhado o conceito de acessibilidade textual e terminológica, explorado no *e-book*²⁹ gratuito e publicamente disponível na internet (Finatto; Paraguassu, 2022).

RastrOS

RastrOS³⁰ é um *corpus* de 50 textos curtos com dados de rastreamento oculares de estudantes universitários durante a leitura silenciosa de parágrafos em português brasileiro. Ele foi criado para a evolução da tarefa de avaliação da complexidade sentencial (Leal et al., 2020, 2022; Vieira, 2020).

Além dos dados de rastreamento ocular de 37 participantes, ele também disponibiliza normas de previsibilidade semântica, obtidas por meio da aplicação de teste cloze com 393 participantes. O RastrOS pode ser obtido integralmente no repositório OSF (*Open Science Framework*)³¹.

²⁴<https://jornalbocaderua.wordpress.com/>

²⁵<https://causaoperaria.org.br/>

²⁶<http://www.ufrgs.br/textecc/porlexbras/corpop/index.php>

²⁷<https://www.ufrgs.br/textecc/porlexbras/porpopular/>

²⁸<https://www.ufrgs.br/textecc/acessibilidade/page/cartilha/>

²⁹https://repositorio.ufu.br/bitstream/123456789/35193/1/eClasse_Acessibilidade_Textual.pdf

³⁰<http://www.nilc.icmc.usp.br/nilc/index.php/rastros>

³¹<https://osf.io/9jxg3/>



6.5.2 Recursos para Complexidade Sentencial

Apresentamos a seguir os recursos disponíveis em português brasileiro para as tarefas que envolvem o trabalho com a complexidade no nível das sentenças.

PorSimplesSENT

O PorSimplesSENT é o primeiro *corpus* para o português brasileiro voltado para as tarefas de complexidade no nível sentencial. Foi compilado por Leal et al. (2018) a partir dos alinhamentos disponibilizados pelo *corpus* PorSimples.

O *corpus* está disponível publicamente³² com diversos agrupamentos, de acordo com as decisões de alinhamento das sentenças que sofreram operação de divisão. Os grupos mais importantes são:

- PSS1 - Todas as divisões: repete a sentença original para cada sentença resultante da divisão;
- PSS2 - Maior divisão: utiliza somente a maior sentença resultante da divisão no alinhamento;
- PSS3 - Somente sentenças que não sofreram operação de divisão.

Como exemplo de aplicação, o agrupamento PSS2 foi utilizado por Leal et al. (2019) e Leal et al. (2020) na tarefa de avaliação da complexidade sentencial. As estatísticas do *corpus* estão listadas na Tabela 6.4.

Tabela 6.4: PorSimplesSENT - Estatísticas

Grupos	PSS1	PSS2	PSS3
Original-Natural	3.535	2.372	1.543
Natural-Forte	4.976	1.501	782
Original-Forte	2.105	1.095	275
TOTAL	10.616	4.968	2.600

6.6 Uso de Modelos de Língua para a Simplificação Textual

Nos últimos anos tivemos uma evolução interessante nas tarefas de PLN com o treinamento de grandes modelos de língua, nessa seção abordamos de forma bem preliminar a utilização desses modelos nas tarefas de avaliação e tratamento da complexidade.

Na primeira versão deste capítulo, fizemos um pequeno experimento utilizando um dos textos do *corpus* PorSimples. Solicitamos ao Bard da Google³³, ao Copilot da Microsoft³⁴ e ao ChatGPT 3.5 da OpenAi³⁵ para simplificar esse texto, e também para explicar as operações de simplificação que foram utilizadas. Um ano depois os modelos evoluíram bastante, repetimos o experimento com o Gemini 2.5 Pro³⁶ e com o ChatGPT 5 Pro³⁷.

³²<http://www.nilc.icmc.usp.br/nilc/index.php/tools-and-resources> e <https://github.com/sidleal/porsimplesent>

³³<https://bard.google.com/>

³⁴<https://www.microsoft.com/pt-br/microsoft-copilot>

³⁵<https://openai.com/blog/chatgpt>

³⁶<https://gemini.google.com/>

³⁷<https://chatgpt.com/>



Quadro 6.4: Exemplos de simplificação utilizando modelos de língua

Original	O ano era 1978. As salas de cinema de todo o mundo exibiam uma produção do diretor Joe Dante em que um cardume de piranhas escapava de um laboratório militar e atacava participantes de um festival aquático. Quase 30 anos depois, banhistas assustados estão se afastando do principal balneário de Uruguaiana, na Fronteira Oeste. Mais de 20 pessoas foram mordidas por palometas (<i>Serrasalmus spilopleura</i>, espécie de piranha) que vivem nas águas da barragem Sanchuri, na margem da Br-472, a 40 quilômetros da cidade. - Os ataques se tornaram mais frequentes. Por isso, aconselhamos mais cautela- diz o subprefeito Nei Pinto. As mordidas em pés e canelas de banhistas não são novidade. O que chamou a atenção das autoridades foi o aumento no número de ataques em relação aos outros anos.
Bard (Gemini 1.0)	No passado, um filme de piranhas aterrorizou as pessoas. Hoje, banhistas de Uruguaiana estão com medo de serem mordidos por palometas. Elas vivem em uma barragem a 40 quilômetros da cidade. Os ataques estão acontecendo com mais frequência. Por isso, devemos ter cuidado - diz o subprefeito Nei Pinto.
ChatGPT (GPT3.5)	Em 1978, um filme sobre piranhas assustou as pessoas em todo o mundo. Agora, em Uruguaiana, muitas pessoas estão sendo mordidas por palometas, que são como piranhas. Essas palometas vivem em um lugar chamado barragem Sanchuri, perto da estrada BR-472. Os ataques estão acontecendo com mais frequência do que antes, então é importante ter cuidado ao nadar lá.
Copilot (GPT4)	Em um filme, peixes que mordem pessoas fogem de um lugar onde os cientistas faziam testes. Muitos anos depois, pessoas que nadavam em um rio foram mordidas por peixes parecidos com os do filme. Os peixes se chamam palometas e vivem em um rio perto de uma cidade. - As pessoas precisam tomar cuidado porque os peixes estão mordendo mais - fala o chefe da cidade. Os peixes já mordiam as pessoas antes, mas agora estão mordendo mais ainda.
ChatGPT 5 Pro	Há muitos anos, um filme mostrou piranhas atacando pessoas. Agora, algo parecido está acontecendo de verdade. Num lugar de banho perto de Uruguaiana, no Rio Grande do Sul, mais de 20 pessoas foram mordidas por palometas. Palometas são peixes parecidos com piranhas. Esses peixes vivem na barragem Sanchuri e mordem, principalmente, os pés e as pernas de quem entra na água. A prefeitura pediu que todos tomem cuidado, porque os ataques ficaram mais frequentes.
Gemini 2.5 Pro	Em uma cidade chamada Uruguaiana, as pessoas que vão nadar em um rio estão com medo. É que um tipo de piranha, chamada palometa, está mordendo os banhistas. Mais de 20 pessoas já foram mordidas nos pés e nas pernas. Uma autoridade da cidade disse que os ataques estão acontecendo mais vezes e pediu para todo mundo tomar mais cuidado.

Os resultados completos estão disponibilizados no [github llm-simplification](https://github.com/llm-simplification)



experiment³⁸.

O *prompt* fornecido aos modelos foi: **“Simplifique o seguinte texto para que um aluno do quarto ano do ensino fundamental consiga compreender, após a simplificação, forneça passo a passo os detalhes das mudanças e motivos para fazer as adaptações no texto:”**, seguido do texto. O texto de teste precisou ser truncado no final do parágrafo anterior antes de completar 2000 caracteres, que é o limite do *prompt* nas interfaces testadas.

Além das simplificações e explicações dos modelos, deixamos também no github as duas simplificações feitas por humanos do PorSimples, para comparação. Reproduzimos no Quadro 6.4 os três primeiros parágrafos do texto original e a simplificação resultante para fins ilustrativos.

De forma geral todos os modelos testados tiveram resultados satisfatórios na tarefa de simplificação proposta, porém o Copilot optou por não situar o texto geograficamente e temporalmente (Em 1978/No passado), eliminando as entidades nomeadas da cidade onde aconteceu o incidente (Uruguaiana), por exemplo. Esta decisão deixou o resultado um pouco superficial.

O Bard e o ChatGPT 3.5 fizeram duas simplificações diferentes, mas ambas com boa qualidade. De forma interessante, optaram por manter entidades nomeadas diferentes, o Bard manteve o nome do subprefeito, enquanto o ChatGPT manteve o nome da barragem e da estrada.

Nas explicações das operações de simplificação utilizadas (que podem ser conferidas integralmente no link do github disponibilizado), o Bard explicou parágrafo por parágrafo, enquanto o ChatGPT focou nas operações utilizadas. Aqui cabe destacar que o Bard teve uma “alucinação” no momento da explicação e inventou uma operação que não aconteceu realmente.

O ChatGPT 3.5 explicou o termo palometas, usando uma exemplificação: “palometas, que são como piranhas”, incluindo porém uma interpretação discutível, pois o texto original diz que palometas são uma espécie (no sentido biológico) de piranhas.

O GPT 5 Pro mostrou grandes avanços na simplificação, mas ainda mantém a interpretação que “palometas são peixes parecidos com piranhas”. Outro ponto interessante dessa simplificação foi a inclusão de um glossário ao final (que pode ser visto no exemplo completo no github).

O Gemini 2.5 Pro manteve a tendência de resumir bastante, mas acertou que a palometa é um tipo de piranha. O resumo removeu a maioria das entidades nomeadas, mas justificou bem as mudanças.

Com este pequeno experimento, fomos surpreendidos positivamente pelos resultados dos grandes modelos de língua na tarefa de simplificação textual. Porém também ficou claro que ainda é necessária uma revisão humana desses resultados, pelo menos por enquanto.

6.7 A complexidade como parâmetro de descrição da linguagem

Até o presente momento, discutimos diversas aplicações de métricas de complexidade textual, seja como meios de classificar a adequação de um texto a um determinado público (como na fórmula *Flesch-Kincaid grade level*), seja como uma etapa no processo de adequá-lo ao público-alvo (como na ferramenta *Simplifica*), seja como forma de diagnóstico de

³⁸<https://github.com/sidleal/llm-simplification-experiment>



distúrbios ou fatores que afetam a pessoa que produz a linguagem (como no Coh-Metrix-Dementia).

Um aspecto comum a todos esses casos é que o objeto central da nossa exploração é o falante – o indivíduo que utiliza a linguagem natural – e como uma métrica de complexidade pode ser um elemento na construção de tecnologias da linguagem que permitam auxiliar o indivíduo ou acessar informações sobre ele.

Entretanto, elementos externos à pessoa também influenciam a maneira como ela se expressa, como:

- a própria estrutura da comunicação humana: o fato de a pessoa querer ser entendida, de que ela tem que pensar e raciocinar enquanto está falando ou escrevendo e de que suas ideias podem mudar no processo;
- a língua utilizada: quais estruturas aquela língua oferece para construir a fala ou sua escrita, quais regras e restrições a língua impõe, e que outras línguas poderiam não impor;
- a situação em que a pessoa se encontra: as regras e convenções que definem como ela deve adaptar o uso da linguagem a contextos específicos.

Todos estes elementos representam níveis em que a linguagem humana é capaz de variar e sobre os quais agem diferentes forças ou restrições que influenciam a maneira como essa variação ocorre. Além disso, estas variações causam modificações palpáveis na complexidade do texto.

Por isso, podemos pensar na complexidade como um fator ou parâmetro da variação da linguagem que, se examinado, nos permitiria entender melhor o próprio fenômeno da variação linguística.

Ao observarmos, pela lente da complexidade, as dimensões em que a linguagem não varia, podemos deduzir coisas sobre a estrutura básica e comum da linguagem humana, subjacente às múltiplas formas em que essa habilidade se realiza, revelando leis e restrições que agem sobre o fenômeno da comunicação humana como um todo.

Assim, as mesmas métricas de quantificação da complexidade textual que nos permitem desenvolver tantas tecnologias importantes em benefício dos seres humanos, podem, também, nos permitir entender melhor o espaço sobre o qual essas tecnologias são desenvolvidas: a própria linguagem humana. Essa forma de pensar está intimamente ligada a duas áreas das ciências da linguagem: a **tipologia** e a **sociolinguística**.

A **tipologia** é a área do conhecimento que se debruça sobre a comparação estrutural entre as diferentes línguas e como a variação da competência da linguagem humana se dá no nível interlinguístico. A tipologia objetiva mapear e entender quais são os universais da linguagem, comuns a todas as línguas humanas, e modelar quais forças definem como e quais variações são possíveis sobre essa estrutura subjacente (Croft, 2002).

Ela ganhou sua forma moderna, principalmente, após o trabalho do linguista estadunidense Joseph H. Greenberg (Greenberg, 1966), no qual, ao comparar a estrutura gramatical de dezenas de línguas, ele mapeou um conjunto de universais linguísticos, comuns a todas – ou à maioria – das línguas da sua amostra. A tipologia está intimamente relacionada a uma visão funcionalista ou teleológica da linguagem, que a descreve a partir da sua função comunicativa. Mais recentemente, inclusive, muitos passaram a considerar a tipologia não como uma área da linguística, mas como uma abordagem à linguística, na sua versão fundida ao pensamento funcionalista, constituindo a chamada abordagem tipológico-funcional à linguagem (Croft, 2002).



A **sociolinguística**, por sua vez, é o ramo das ciências da linguagem interessado em como a linguagem varia e se modifica em diferentes contextos e situações sociais, indo além de como a gramática oficial da língua é definida, e buscando descrever como os diferentes grupos de uma sociedade realizam a língua em diferentes contextos, caracterizando as diversas forças que determinam e influenciam essa variação (Wolfram, 2006).

O **estudo dos registros**, em especial, é uma linha fortemente conectada à sociolinguística, que busca descrever como a linguagem varia entre registros – definidos aqui como variações da língua em função das características e parâmetros da situação em que ela é produzida (Biber, 2006).

Se a tipologia se debruça sobre a variação linguística no nível interlinguístico, a sociolinguística foca na variação no nível intralinguístico. Como veremos, entretanto, a complexidade parece ser um fenômeno transversal a esses níveis, embaralhando o limite entre essas duas áreas.

6.7.1 Questões sobre a complexidade como parâmetro

Um dos pioneiros do uso da complexidade linguística como parâmetro de variação, no sentido moderno, foi o linguista estado-unidense Charles F. Hockett, que, em (Hockett, 1958), apresentou duas conjecturas, nas quais a complexidade desempenha um papel central como parâmetro de variação: a hipótese da equicomplexidade e a hipótese do *trade-off* morfossintático:

- **Hipótese da Equicomplexidade:** todas as línguas humanas possuem a mesma capacidade comunicativa e a mesma complexidade geral, mesmo usando de mecanismos distintos para transmitir informações e construir estruturas complexas.
- **Hipótese do *trade-off* Morfossintático:** quanto mais complexa uma língua é morfologicamente – i.e. quanto mais informação ela transmite através de estruturas internas à palavra – menos complexa ela será sintaticamente – i.e. menos informação ela transmitirá através de estruturas externas às palavras.

Na visão de Hockett, a hipótese do *trade-off* morfossintático seria o principal mecanismo através do qual as línguas humanas manteriam a equicomplexidade.

Essas conjecturas exerceram um papel essencial no seu tempo, por virem como resposta a concepções prévias da linguagem bastante problemáticas, que caracterizavam as línguas indo-europeias como o pináculo de um dito processo evolutivo de complexidade, em detrimento das manifestações linguísticas de outros povos.

No mesmo compasso, várias outras conjecturas acerca da complexidade linguística, em especial da complexidade textual, foram surgindo após as ideias originais de Hockett, entre elas, podemos mencionar³⁹:

- **Hipótese da conexão demográfica:** a complexidade média da linguagem produzida por uma determinada população refletiria a distribuição de falantes primários e secundários da língua naquela população.
- **Hipótese da simplificação mediante contato:** manifestações linguísticas que surgem do contato entre populações que falam línguas diferentes apresentariam menor complexidade geral, justamente como forma de facilitar a comunicação entre essas populações (McWhorter, 2001).

³⁹Veja (Ehret et al., 2023) para uma discussão mais aprofundada sobre proposições em aberto.

- **Hipótese dos fatores geográficos e filogenéticos:** línguas com relações de parentesco, ou línguas geograficamente próximas, com interação frequente, apresentariam semelhanças nos seus perfis de complexidade – i.e. na maneira como distribuem sua complexidade em diferentes níveis e estratégias.

Entretanto, a avaliação objetiva dessas conjecturas seria muito difícil sem métricas que permitissem de alguma forma quantificar a complexidade de uma língua ou do produto de uma língua, como uma fala, um texto ou uma sentença. Daí, vários pesquisadores se debruçaram sobre como definir e mensurar a complexidade linguística.

Trabalhos como os de Nichols (1998), McWhorter (2001) e Dahl (2004) foram pioneiros na definição de tais métricas, entretanto suas abordagens usualmente exigiam a anotação manual por especialistas, a partir de grandes amostras textuais ou extensas gramáticas de cada língua.

Contudo, fatores como (i) o desenvolvimento da linguística computacional e do PLN, (ii) o avanço do poder de processamento computacional e (iii) a democratização da internet e o consequente aumento do acesso a textos de mais línguas através de meios digitais, mudaram o cenário de pesquisa, levando à proposição de métricas de complexidade computacionais.

Essas novas métricas trouxeram a vantagem de poderem ser computadas de forma mais rápida e sem tanta dependência da anotação de especialistas, além de permitirem comparação em grande escala de diversas línguas. Com esses avanços, aumentaram as nossas possibilidades científicas do uso da complexidade como um parâmetro de estudo da linguagem, em especial para a análise e verificação das hipóteses como as que listamos nesta seção.

6.7.2 Métricas de Complexidade Baseadas em Compressibilidade

Atualmente, contamos com diversas métricas computacionais de complexidade de linguagem. Muitas das medidas apresentadas anteriormente neste capítulo servem também ao propósito do estudo da complexidade como parâmetro linguístico, da forma que o apresentamos nesta seção.

Entretanto, ainda estaria fora do nosso alcance apresentar aqui todas as métricas existentes e os tantos experimentos e resultados obtidos através delas. Para dar um vislumbre do tipo de resultados e conclusões que podemos obter nessas investigações, apresentaremos brevemente alguns resultados obtidos a partir das métricas baseadas em compressibilidade, uma família de métricas de complexidade textual que tem apresentado resultados bastante interessantes e prototípicos, apesar de serem apenas uma parcela muito pequena do conjunto de métricas que podemos considerar.

As métricas de complexidade baseadas em compressibilidade são um conjunto de métricas de complexidade textual, desenvolvidas em (Juola, 1998), (Juola, 2008) e (Ehret; Szmrecsanyi, 2016). Essas métricas consideram que a complexidade de um texto é a quantidade de informação nele contida. A complexidade em um nível linguístico – como morfológico ou sintático – é, então, a parte da informação total do texto sendo transmitida através das estratégias daquele nível específico.

Para poder acessar a quantidade de informação contida no texto e em cada um dos níveis de interesse, essas métricas utilizam algoritmos de compressão de dados simples, como o *gzip*. Esses algoritmos são capazes de perceber padrões de recorrência dentro de um conjunto de dados, reduzindo-o a uma quantidade de informação mínima, a partir da qual o arquivo original pode ser reconstruído. As métricas baseadas em compressibilidade



tomam, então, o tamanho de um texto comprimido como uma aproximação da quantidade de informação – e, portanto, da complexidade – nele contida.

Para aproximar a quantidade de informação em níveis específicos, realizam-se perturbações automáticas e aleatórias no conteúdo do texto, que visam afetar os padrões e estratégias através dos quais a informação é transmitida no nível de interesse, medindo o impacto gerado na quantidade de informação total do texto.

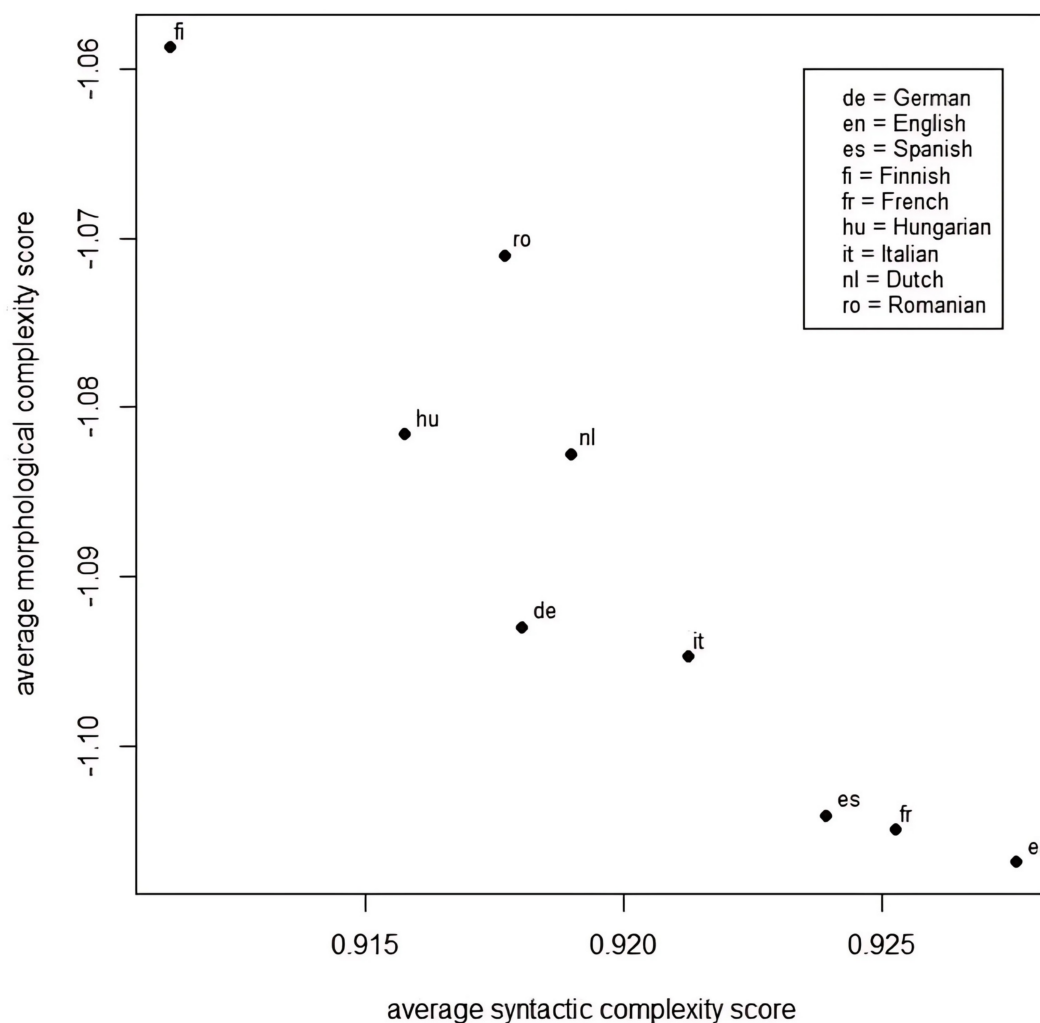
Quando aplicadas ao mesmo texto traduzido em diferentes línguas – visando verificar a hipótese do *trade-off* morfossintático – essas métricas nos apresentam resultados bastante interessantes. As Figuras 6.11 e 6.12 mostram uma renderização do *trade-off* em dois experimentos distintos. O primeiro, realizado em Juola (2008), mostra uma troca clara entre as complexidades sintática e morfológica, como previsto por Hockett, para algumas línguas europeias.

Já o gráfico da Figura 6.12, extraído de (Serras et al., 2024), mostra o mesmo *trade-off* para um conjunto mais amplo, composto majoritariamente de línguas nativas da América do Sul, com algumas línguas europeias de referência. Aqui, também é perceptível uma tendência inversa entre as complexidades sintática e morfológica, entretanto a dispersão é muito maior. Talvez isso possa ser explicado pela grande diversidade filogenética desse conjunto – de fato, é possível perceber que muitos dos dados parecem estar agrupados por semelhança filogenética – o que pode ser lido como um indício da hipótese dos fatores geográficos e filogenéticos, apresentada anteriormente.

Um outro resultado interessante é que, quando as mesmas métricas são computadas sobre conjuntos de textos originários da mesma língua, mas de diferentes registros dentro dessa língua, i.e., sobre textos gerados em situações específicas distintas, observamos o mesmo padrão do *trade-off* morfossintático. Isso fica claro no gráfico da Figura 6.13, extraído de (Ehret, 2021). Nesse gráfico, vemos o *trade-off* morfossintático para vários registros ou gêneros distintos da língua inglesa. Também é possível ver uma separação clara entre registros mais e menos oralizados, no que tange ao seu perfil de complexidade.

Apesar de breve, essa comparação de resultados não só nos mostra a robustez dessas métricas, mas a capacidade que a quantificação da complexidade nos traz, ao permitir examinar hipóteses até então de difícil verificação e conectar fenômenos em escopos bastante distintos. Essas conexões talvez permitam que, nos próximos anos, construamos modelos cada vez mais profundos da linguagem e de como ela funciona em diferentes níveis. Talvez, esses novos modelos possam, inclusive, acarretar futuras melhorias nas tecnologias relacionadas à complexidade textual, apresentadas nas seções anteriores deste capítulo.

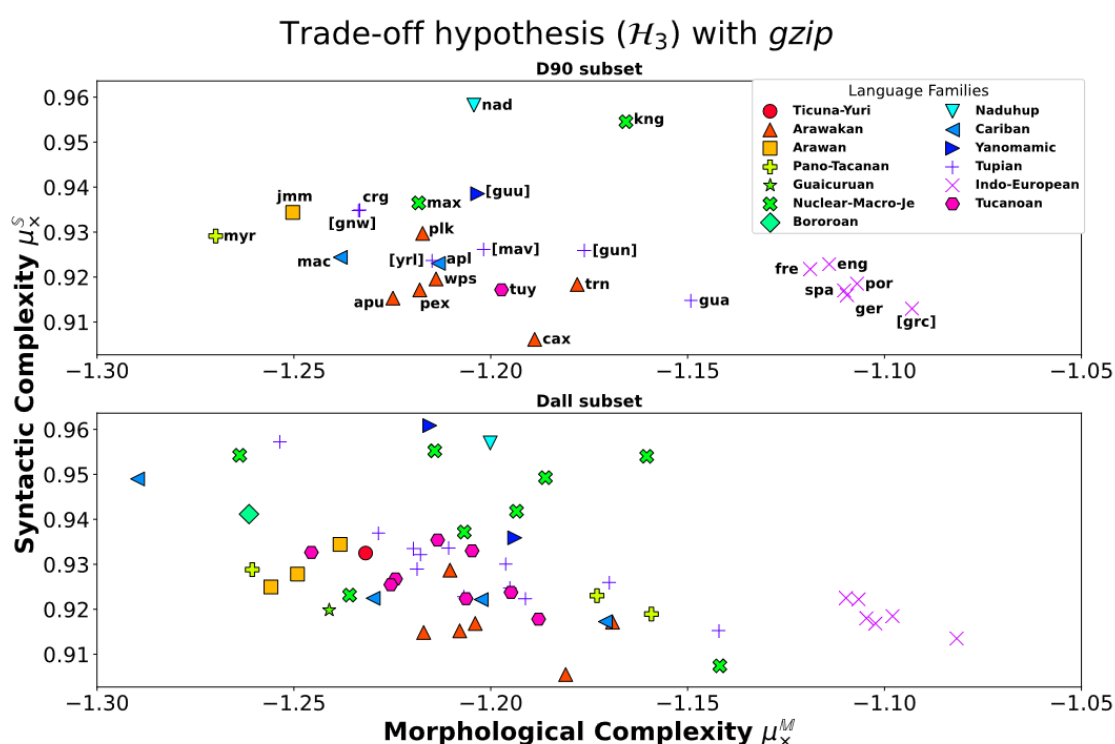
Figura 6.11: Gráfico apresentando o *trade-off* entre complexidade morfológica e sintática, usando métricas de complexidade de linguagem baseadas em compressibilidade, computadas sobre um conjunto de traduções paralelas de textos bíblicos em várias línguas europeias.



Fonte: (Juola, 2008)

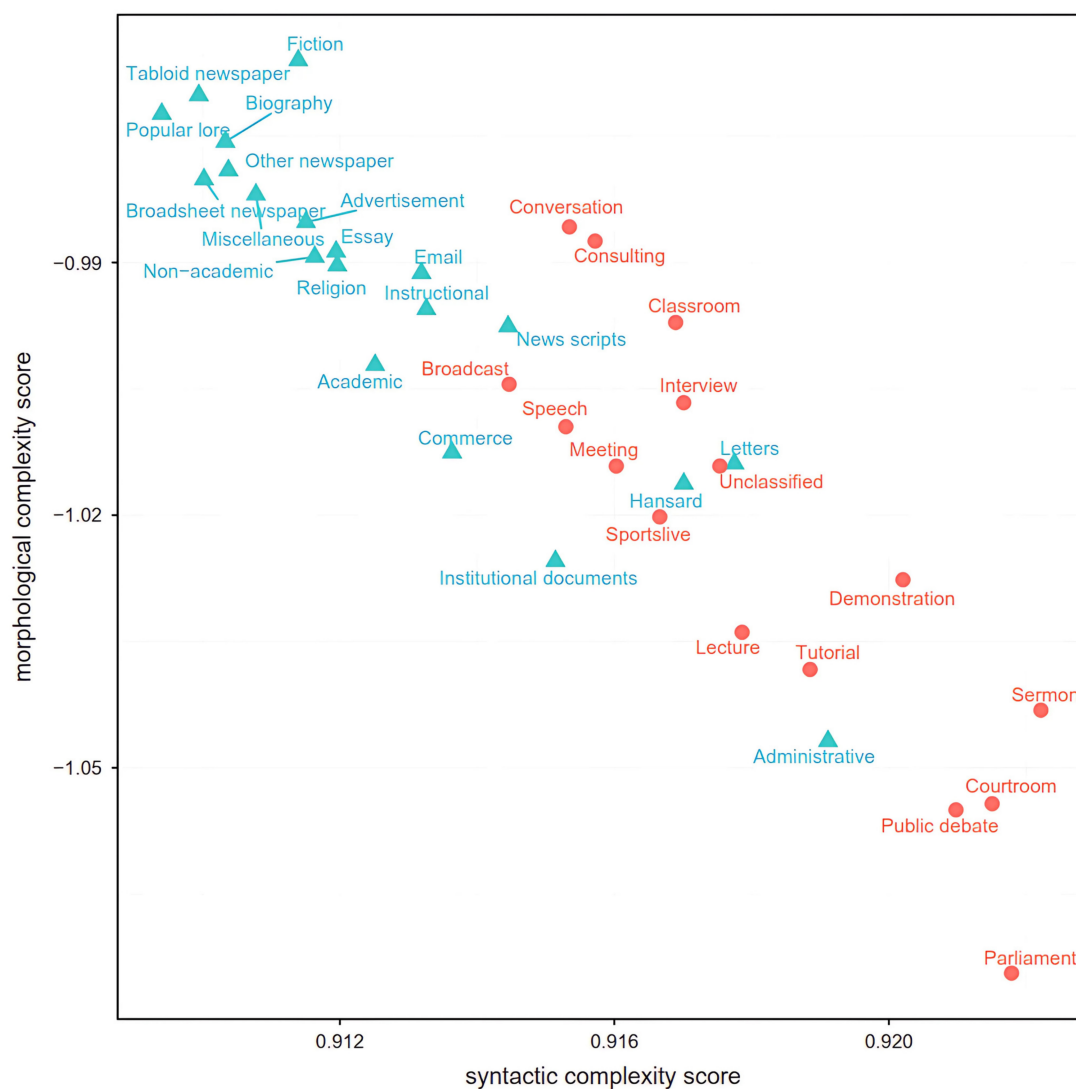


Figura 6.12: Gráfico apresentando o *trade-off* entre complexidade morfológica e sintática, usando métricas de complexidade de linguagem baseadas em compressibilidade, computadas sobre um conjunto de traduções paralelas de textos bíblicos em diversas línguas, principalmente nativas da América do sul. Os dois painéis representam técnicas distintas de seleção dos dados.



Fonte: (Serras et al., 2024)

Figura 6.13: Gráfico apresentando o *trade-off* entre complexidade morfológica e sintática, usando métricas de complexidade de linguagem baseadas em compressibilidade, computadas sobre um conjunto de textos de vários registros ou gêneros distintos da língua inglesa. Os pontos triangulares azuis indicam registros/gêneros escritos e os pontos circulares alaranjados indicam registros/gêneros falados.



Fonte: (Ehret, 2021)



6.8 Considerações finais

Neste capítulo, apresentamos uma visão geral sobre o tema complexidade textual, começando pelas tarefas relacionadas, passando pelas diversas métricas disponíveis para avaliação automática, pelos recursos disponíveis atualmente para a avaliação das tarefas no português brasileiro e apresentamos também como grandes modelos de linguagem podem ser utilizados para a tarefa de simplificação textual. Por fim, apresentamos como a medição da complexidade textual pode ser usada como parâmetro na construção de modelos sobre a linguagem humana.

Como os grandes modelos de linguagem impactaram em várias tarefas do PLN nos anos recentes, apresentamos também como eles podem ser utilizados para a tarefa de simplificação textual.

Esperamos que esse capítulo sobre a complexidade textual e suas tarefas relacionadas desperte a curiosidade de mais alunos e pesquisadores e contribua de forma positiva para o desenvolvimento da grande área de PLN. Esperamos também que no futuro sejam criadas ferramentas incríveis para adequação dos textos para os leitores, mantendo sempre o nível de desafio adequado para o entendimento e motivação.



Capítulo 7

Correção Automática de Redação

Amanda Pontes Rassi
Priscilla de Abreu Lopes

Publicado em: 26/09/2023
Atualizado em: 16/04/2026

 <https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/>

7.1 Introdução

A Correção Automática de Redação (CAR) é uma das diversas aplicações do Processamento de Linguagem Natural (PLN). De forma geral, ela pode ser definida como “o processo de avaliação e atribuição de nota em textos escritos em prosa, via programas computacionais” (Shermis; Burstein, 2013) ¹.

Embora a correção manual de redações seja uma prática bastante antiga – os primeiros sistemas automáticos para essa finalidade surgiram na década de 60, para a língua inglesa – só mais recentemente foram adaptados para o português.

Em inglês, as áreas de *Automated Essay Scoring* (AES) e *Automated Essay Evaluation* (AEE) surgem como distintas, porém complementares e, às vezes, com alguma intersecção. A AES tem como principal objetivo automatizar a atribuição de nota, enquanto a AEE amplia esse foco, incluindo também a geração de feedbacks para o aluno, contribuindo de forma mais direta com o processo de aprendizagem da escrita.

A AES costuma ser traduzida para o português como **Avaliação Automática de Redação** (AAR) (Bittencourt Jr., 2020; Da Silva Jr., 2021; Lima et al., 2023), enquanto a AEE está associada ao termo **Correção Automática de Redação** (CAR), apesar de se tratar de um falso cognato. Neste capítulo, adotamos o termo CAR para nos referirmos a uma solução mais abrangente, que contempla tanto a avaliação quanto o feedback, isto é, que reúne as funções de AES e AEE. Para que seja considerada como solução completa de CAR, a aplicação deve contemplar pelo menos três etapas básicas:

- i) a detecção de desvios no texto;
- ii) a atribuição da nota, seja ela global ou por critério; e
- iii) um feedback para o aluno.

Cada uma dessas etapas pode ser vista como uma aplicação independente no PLN. Por exemplo, existem várias ferramentas de auxílio à escrita, bem como corretores ortográficos e gramaticais, que executam exclusivamente a tarefa de identificação de desvios no texto; e isso constitui uma aplicação em si. Da mesma forma, o fornecimento de sugestões de melhoria se aproxima de aplicações de geração de linguagem natural (ou *Natural Language Generation*).

¹Tradução nossa. Do original: “the process of evaluating and scoring written prose via computer programs”.



Ainda que essas funcionalidades possam ser oferecidas de forma independente, consideramos que uma solução pedagógica robusta de CAR deve integrar as três etapas mencionadas, as quais serão exploradas com mais profundidade ao longo deste capítulo.

Antes de abordar cada uma das etapas, porém, faremos uma breve explicação sobre o objeto de estudo da CAR, que é a redação escolar, definindo e exemplificando os principais gêneros e tipos textuais, os critérios avaliados e alguns modelos brasileiros de correção de redação.

7.1.1 O que é uma redação escolar?

A redação escolar é considerada um gênero textual, mas também pode ser distribuída em vários tipos e gêneros textuais. As redações, ou textos² de redação escolar, são geralmente utilizadas para avaliar as habilidades de escrita, interpretação, argumentação e criatividade dos alunos, bem como para desenvolver o pensamento crítico e a capacidade de expressão escrita. As redações podem abordar temas diversos, desde assuntos cotidianos até questões mais complexas e abstratas, e são uma forma importante de avaliar o progresso dos alunos ao longo do tempo.

Para fins didáticos e de correção de redação, é importante salientar a diferença entre tipo textual e gênero textual, já que as redações devem atender a um tipo específico e a algum gênero específico, a depender da proposta de redação. Por exemplo, a redação do Enem³ é sempre do tipo argumentativo e do gênero dissertação-argumentativa.

Os **tipos textuais** (ou “modos textuais”, para Marcuschi (2008, p. 154)) se referem à forma como o texto é organizado, ou seja, a sequência linguística e os aspectos lexicais, sintáticos, tempos verbais, relações lógicas que são mobilizados para constituir o texto. Existe um conjunto bastante limitado de tipos textuais, o qual abrange: narração, argumentação, descrição, exposição e injunção. Ressaltamos que um texto raramente apresenta apenas características de um mesmo tipo. Deste modo, classificamos um texto como sendo de um determinado tipo quando há predominância de elementos que o caracterizam.

Já os **gêneros textuais** são formas de comunicação que se desenvolvem em diferentes contextos sociais e culturais e se caracterizam pelo seu propósito ou objetivo comunicativo. Em outras palavras, cada gênero tem uma finalidade específica e é utilizado em determinadas situações comunicativas, dependendo de fatores sociais, culturais, dos falantes, da relação entre eles, do contexto, da finalidade da comunicação, dentre vários outros. Dependendo da situação comunicativa, cada gênero pode exigir um registro ou vocabulário específico, a norma culta ou coloquial, na modalidade escrita ou oral da língua.

Por esse motivo, os gêneros são mais fluidos, podendo surgir, modificar-se, mesclar com outros, desaparecer e reaparecer com outra roupagem em outro contexto ou época. São exemplos de gêneros textuais: bula de remédio, carta pessoal, diálogo informal, e-mail, edital de concurso, inquérito policial, piada, receita culinária, reportagem, resenha, sermão etc.

As redações podem apresentar diversos formatos e objetivos, dependendo do nível de

²Existe uma longa discussão conceitual e técnica sobre a definição do termo “texto” em Linguística Textual. Para o propósito deste capítulo, adotaremos como conceito de “texto” um conjunto de palavras e frases organizadas de forma coerente e coesa, com o objetivo de transmitir uma mensagem ou ideia. Em outras palavras, o texto é uma unidade de linguagem que tem um sentido completo e pode ser compreendido em um contexto específico.

³O Exame Nacional do Ensino Médio (Enem) é uma prova do Governo Federal que avalia o desempenho escolar dos alunos ao término do Ensino Médio. Essa prova avalia várias áreas do conhecimento e também a produção de uma redação.

ensino e do tema proposto pelo professor ou pela instituição de ensino. Dentre os tipos e gêneros textuais mais comuns associados à redação escolar, convém mencionar:

- **Dissertação**, em que o autor apresenta e disserta sobre um determinado tema, apresentando informações e argumentos relacionados ao assunto;
- **Narração**, em que o autor conta uma história, relatando fatos e acontecimentos em alguma ordem que pode ser cronológica ou não;
- **Carta**, em que o autor se dirige a um destinatário específico, fazendo requisições, solicitações ou expressando suas opiniões e sentimentos;
- **Artigo de opinião**, em que o autor defende um ponto de vista sobre um tema específico, utilizando argumentos e evidências para sustentá-lo;
- **Resenha**, em que o autor faz uma análise crítica de um texto, obra ou produto.

Esses são apenas alguns dos tipos e gêneros de redação escolar mais comuns. Os demais incluem a descrição, a exposição, a crônica, o relatório, o conto, a fábula, entre outros.

7.1.2 O que é avaliado?

Vários aspectos do texto são avaliados em uma correção de redação, tais como o uso da norma padrão da língua portuguesa, a adequação ao tema e ao gênero, questões relacionadas à coesão, à coerência, à progressão textual etc. Cada modelo de correção organiza e nomeia seus critérios de avaliação de formas distintas, mas, basicamente, todos eles analisam:

Língua portuguesa Avalia a linguagem usada para expressar o conteúdo, verificando se há desvios ortográficos e/ou gramaticais, se a norma (culto ou coloquial) está de acordo com o tipo de texto exigido, se há problemas de estrutura sintática nas frases, orações e períodos, se o vocabulário foi usado adequadamente etc. Outros nomes para esse critério incluem “Escrita”, “Modalidade escrita”, “Norma culta”, “Norma padrão”, “Correção gramatical e adequação vocabular” ou “Expressão (modalidade)”.

Tema Esse aspecto avalia a adequação da redação em relação à temática proposta, verificando se a abordagem do tema foi completa, se tangenciou ou fugiu do tema proposto, se o abordou de forma superficial ou profunda etc. Também é chamado de “Abordagem temática”, “Desenvolvimento do tema”, “Proposta temática” ou “Progressão temática”.

Gênero Esse critério considera a adequação da redação em relação ao tipo textual e ao gênero textual exigidos na proposta. Também pode ser chamado de “Gênero textual”, “Adequação ao tipo textual”, “Organização do texto dissertativo-argumentativo” ou “Estrutura (gênero/tipo de texto)”.

Coerência Neste quesito, avalia-se a coerência entre as ideias, a ordem dos argumentos, a profundidade da argumentação, a clareza e autoria das ideias desenvolvidas, assim como verifica-se se há contradições no texto, se as informações são vagas e/ou muito generalistas, se falta informação, dentre outros. Também pode ser chamado de “Progressão textual”, “Defesa do ponto de vista”, “Coerência dos argumentos”, “Estrutura (coerência)”, “Indícios de autoria” e outros termos.



Coesão Avalia o uso correto ou incorreto, presença ou ausência, pertinência ou não de operadores coesivos, tais como conjunções, preposições, pronomes e expressões discursivas. O critério é também chamado de “Coesão e articulação”, “Articulação das partes do texto”, “Expressão (coesão)”, “Conexão entre os parágrafos”, “Uso de operadores argumentativos”, “Recursos coesivos”, dentre outros termos.

Além desses aspectos que são comuns a todos os modelos de correção, alguns professores, instituições de ensino e vestibulares também podem avaliar a “Leitura”, ou seja, o uso e interpretação dos textos motivadores ou da coletânea que embasa a proposta de redação, e também a presença e adequação da “Proposta de intervenção”, que é um critério exclusivo do Enem.

7.1.3 Alguns modelos brasileiros de correção

O principal modelo de correção de redação, no Brasil, é o Enem, responsável pela avaliação anual de milhões de alunos⁴. Mas também existem outros modelos de correção relacionados a vestibulares e universidades específicas, tais como Fuvest, Unesp, Unicamp, FGV e outros igualmente relevantes. Apesar de haver critérios gerais que são avaliados por todos eles, cada um tem autonomia para definir sua grade específica, os pesos de cada critério e sua própria forma de avaliação.

O Quadro 7.1 apresenta quatro modelos brasileiros de correção relacionados a vestibulares, indicando o gênero textual exigido, seus critérios de avaliação e faixas de nota possíveis.

Quadro 7.1: Modelos de correção de vestibulares.

Modelo	Tipo/Gênero	Critérios avaliativos	Nota mín. critério	Nota máx. critério	Nota global
Enem	Dissertação-argumentativa	Língua Portuguesa (Competência 1)	0	200	1000
		Abordagem temática e adequação ao tipo textual (Competência 2)	0	200	
		Progressão textual e defesa do ponto de vista (Competência 3)	0	200	
		Coesão e articulação (Competência 4)	0	200	
		Proposta de intervenção (Competência 5)	0	200	
Fuvest	Dissertação ou narrativa	Desenvolvimento do tema uso da coletânea e autoria	0	15	50
		Compreensão e atendimento da proposta quanto ao gênero e tipo de texto	0	10	

⁴No último Enem 2025, foram mais de 4,8 milhões de inscrições confirmadas: <https://www.gov.br/mec/pt-br/assuntos/noticias/2025/julho/enem-2025-mais-de-4-8-milhoes-de-inscritos-confirmados>



Unesp	Dissertação-argumentativa	Recursos linguísticos (coesão e coerência) e progressão textual	0	15	28
		Convenções da escrita e adequação vocabular	0	10	
		Tema	–	–	
		Estrutura (gênero/tipo de texto e coerência)	–	–	
Unicamp	variados	Expressão (modalidade e coesão)	–	–	12
		Proposta temática	0	2	
		Gênero	0	3	
		Leitura	0	3	
		Articulação escrita	1	4	

No modelo de correção do **Enem**⁵, o aluno é avaliado quanto à produção de um texto do tipo dissertativo-argumentativo para um tema específico, que muda todo ano. A avaliação é dividida em 5 competências (critérios avaliativos), cada uma no intervalo de notas de 0 a 200. A soma direta das notas das competências leva à nota total, que fica no intervalo de 0 a 1000. Considerando os critérios básicos descritos na Seção 7.1.2, vale dizer que o Enem os divide da seguinte forma: (i) Língua Portuguesa, (ii) Tema e Gênero, (iii) Coerência e (iv) Coesão. Além desses, o modelo ainda avalia um quinto critério, que é a presença e adequação da “Proposta de Intervenção”, que consiste na sugestão de ação ou medida interventiva para solucionar ou minimizar o problema associado ao tema proposto.

O vestibular da **Fuvest**⁶ também exige um texto do gênero dissertativo-argumentativo para um tema específico que muda todo ano. A partir do vestibular de 2026, o candidato também tem mais uma opção de gênero (narrativo), além da dissertação, ambos sobre o mesmo tema. O modelo de avaliação até 2024 agrupava os critérios básicos em 3 (Tema/Gênero, Coerência/Coesão e Língua Portuguesa). A partir do vestibular 2026, essa grade está dividida em 4 critérios: (i) Desenvolvimento do tema, uso da coletânea e autoria, que vale de 0 a 15 pontos; (ii) Compreensão e atendimento da proposta quanto ao gênero e tipo de texto, que pode somar até 10 pontos; (iii) Recursos linguísticos (coesão e coerência) e progressão textual, também de 0 a 15 pontos; e (iv) Convenções da escrita e adequação vocabular, com pontuação máxima igual a 10. Com a soma de pontos dos 4 critérios, o aluno pode tirar de 0 a 50.

Já no modelo de correção da **Unesp**⁷, os textos, que devem seguir o gênero dissertação-argumentativa, também são avaliados em três eixos, agrupados da seguinte forma: (i) Tema, (ii) Gênero/Coerência e (iii) Língua Portuguesa/Coesão. A pontuação individual ou peso por critério não é divulgado no material do candidato, mas é definido que a pontuação final fica entre 0 e 28 pontos.

A redação da **Unicamp**⁸, a cada ano, varia a exigência dos tipos e gêneros textuais⁹,

⁵Cartilha do participante do Enem 2022: https://download.inep.gov.br/download/enem/cartilha_do_participante_enem_2022.pdf.

⁶Manual do candidato do vestibular Fuvest 2023: https://www.fuvest.br/wp-content/uploads/fuvest2023_manual_candidato_retificado_29112022.pdf.

⁷Manual do candidato do vestibular Unesp 2023: <https://documento.vunesp.com.br/documento/stream/MzQxOTk5NA%3d%3d>.

⁸Manual do ingresso (https://www.comvest.unicamp.br/wp-content/uploads/2023/02/Manual_do_Ingresso_2023_Atualizado.pdf) e grade de redação (<https://www.comvest.unicamp.br/vestibular-2023/grade-da-redacao/>) do vestibular 2023 da Unicamp.

⁹Além da dissertação, outros gêneros textuais já exigidos pela Unicamp são: síntese e carta-convite (2015),



geralmente oferecendo duas alternativas das quais o candidato deve escolher uma para execução. A Unicamp agrupa os critérios básicos da seguinte forma: (i) Tema, (ii) Gênero e (iii) Língua Portuguesa/Coerência/Coesão. Além desses três eixos, também avalia a “Leitura”, que corresponde à leitura e interpretação crítica dos textos fornecidos na proposta, sem contudo copiá-los ou parafraseá-los. Na avaliação, cada critério possui pesos diferentes: Tema varia entre 0 e 2 pontos, Gênero entre 0 e 3, Leitura entre 0 e 3 e Língua Portuguesa/Coerência/Coesão varia entre 1 e 4 pontos. A soma dos pontos de cada critério leva à nota final, cujo valor máximo é de 12 pontos.

Fora as diferenças já apontadas, convém ressaltar que todos os modelos penalizam o aluno (zerando a redação) no caso de falhas graves. No entanto, cada modelo define um conjunto específico de falhas graves, que podem ser: fuga ao tema, fuga ao gênero, assinatura na prova, desenho ou sinal gráfico, redação em língua estrangeira, caligrafia ilegível, recado para o corretor, parte desconectada, texto insuficiente, dentre outras situações¹⁰.

Quanto ao título da redação, o Enem e a Unesp não exigem, mas também não proíbem; simplesmente desconsideram para a avaliação da redação. Já a Fuvest não menciona a exigência de título no Manual do candidato do vestibular Fuvest 2024¹¹, mas coloca como instrução no Caderno de prova¹². Já a Unicamp pode ou não exigir título, a depender do gênero textual proposto.

Essa grande variedade de modelos de correção é considerada um dos grandes desafios para a CAR, não sendo recomendado treinar um modelo computacional que abarque todos os tipos de correção ou que misture redações dos vários tipos como dados de treinamento para algum modelo. A exigência por modelagens de nota específicas por modelo de correção não impede, no entanto, o reaproveitamento de parte das ferramentas de correção, como a detecção automática de desvios no texto, desde que modelos de correção distintos tenham diretrizes similares para esse tipo de tarefa.

Todas essas questões serão mais detalhadas ao longo deste capítulo, que está organizado da seguinte forma: a Seção 7.2 descreve como fazer a detecção de desvios em textos em português, demonstrando alguns tipos de desvios e os formalismos usados. A Seção 7.3 apresenta os principais trabalhos da literatura que realizam a atribuição da nota para redações em português. A Seção 7.4 demonstra as possibilidades de geração de um feedback para o aluno. Na Seção 7.5, discutimos as vantagens e desvantagens da correção manual e da correção automática, a fim de esclarecer ao leitor que ambas possuem potencialidades, mas também limitações. Por fim, nas Considerações finais (Seção 7.6), retomamos os pontos principais do capítulo.

7.2 Detecção de desvios no texto

Conforme apresentado na Introdução (Seção 7.1), consideramos que uma das etapas da Correção Automática de Redação (CAR) é a detecção ou identificação de desvios¹³. Essa

resenha e texto de divulgação científica (2016), carta argumentativa e texto de apresentação (2017), palestra e artigo de opinião (2018), abaixo-assinado e postagem em fórum (2019), roteiro de podcast e crônica (2020), discurso político e diário (2021), postagem para redes sociais e manifesto coletivo (2022), depoimento sigiloso e convocação (2023) e carta denúncia e discurso (2024).

¹⁰Para uma descrição completa e exemplos de todos os casos que zeram a redação em cada modelo de correção, sugere-se consultar os respectivos manuais do candidato ou cartilhas do participante.

¹¹https://www.fuvest.br/wp-content/uploads/fuvest2024_manual_candidato.pdf

¹²https://acervo.fuvest.br/fuvest/2020/fuv2020_2fase_dia_1.pdf

¹³Adotaremos o conceito de “desvio” como sinônimo de “erro”, mas evitaremos esse segundo termo para mitigar a carga negativa, os preconceitos e os julgamentos contidos na palavra “erro”.



etapa nem sempre é realizada nos trabalhos de Avaliação Automática de Redação, ou, por vezes, os desvios são contabilizados para o cálculo da nota, porém não são apresentados ao aluno.

A detecção desses desvios costuma ser feita por meio de três abordagens distintas: baseada em regras (**abordagem simbólica**), baseada em modelos estatísticos (**abordagem estatística**) e a baseada em LLMs (**abordagem gerativa**). Os sistemas baseados em regras são mais adequados para identificar desvios gramaticais, o que é mais comum de ser cometido por falantes nativos da própria língua, enquanto os estatísticos capturam melhor os desvios de uso, que são erros mais comuns por não-nativos¹⁴. A abordagem gerativa tem se tornado uma opção muito usual desde a popularização dos LLMs (Subseção **Modelos de Linguagem Neurais Modernos**) e suas constantes melhorias.

Embora a abordagem simbólica (baseada em regras) seja considerada obsoleta para tarefas mais complexas, ainda é a mais utilizada atualmente para detectar desvios na área de CAR. Para outros tipos de tarefas, modelos estatísticos e neurais performam melhor e são mais escaláveis do que modelos simbólicos. A abordagem gerativa pode ser interessante como ferramenta de correção do texto, porém, ao apontar desvios, as respostas estão sujeitas a alucinações (situação exemplificada no Capítulo **ChatGPT, MariTalk e outros agentes de conversação**). Apesar do desempenho de modelos estatísticos e LLMs, a abordagem simbólica permite, com alta precisão, mostrar ao aluno o trecho exato que contém o desvio, explicar por que está errado e ainda fazer sugestões de correção.

Para o português, existem recursos disponíveis, tais como o CoGroo¹⁵ e o Language-Tool¹⁶, que são repositórios contendo regras gramaticais para a língua portuguesa. Esses recursos têm versões livres, gratuitas e de código-aberto, com extensão para navegadores web e também acopláveis a editores de texto.

Também há plataformas de correção de redação que desenvolveram seu próprio conjunto de recursos linguísticos e regras gramaticais, o que é uma boa opção quando há um padrão muito claro e estruturado que se possa expressar com regras simbólicas ou expressões regulares, que é o caso dos desvios mais comuns em redações.

Na Seção 7.2.1 caracterizamos alguns dos tipos de desvios mais comuns em redações. Posteriormente, na 7.2.2, apresentamos duas alternativas de formalismo para a definição de regras de detecção de desvios.

7.2.1 Tipos de desvios

Existem diversos tipos de desvios que podem ser marcados em uma redação, como os ortográficos, os gramaticais (ou sintáticos), os de uso de vocabulário ou registro, os desvios no uso de recursos coesivos, dentre outros. Para cada modelo de correção de redação, é possível criar uma taxonomia própria de tipos de desvios que se pretende identificar em um texto.

Ressaltamos que a criação de recursos para esse tipo de tarefa é um processo difícil, moroso e custoso, que depende de especialistas. Deste modo, é importante estabelecer um planejamento criterioso caso seja necessário criar recursos próprios.

Nesta seção exploramos alguns tipos de desvios, por serem os mais comuns, mas é importante esclarecer que os tipos de desvios não se limitam aos indicados neste capítulo.

¹⁴Para uma explicação detalhada dos vários sistemas que usam cada uma das abordagens simbólica e estatística para detecção de desvios, ver Leacock et al. (2010) e Gamon et al. (2013).

¹⁵<https://cogroo.sourceforge.net/>

¹⁶<https://languagetool.org/pt-BR/>



Na Seção 7.2.1.1 descrevemos os desvios ortográficos, na Seção 7.2.1.2, os gramaticais e na Seção 7.2.1.3, outros tipos de desvios mais comuns em redações.

7.2.1.1 Desvios ortográficos

A grande maioria dos desvios ortográficos é facilmente detectável e tratável. O simples uso de um bom dicionário de língua portuguesa já indica quais palavras existem e quais não existem na língua. Portanto, identificar palavras com grafia desviante do léxico é uma tarefa relativamente simples.

O Unitex¹⁷, por exemplo, dispõe de três dicionários muito completos para o português: o **Delas** (com cerca de 75.000 formas canônicas simples), o **Delaf** (com cerca de 9.000.000 entradas de palavras flexionadas) e o **Delacf** (com cerca de 4.000 entradas de formas compostas). Esse recurso pode ser usado como uma primeira etapa de identificação de desvios ortográficos, a fim de identificar palavras que existem no léxico do português e palavras desviantes.

Outros desvios ortográficos se dividem em:

- problemas de falta de sinal gráfico ou uso indevido de acentuação (ex: “prática” x “pratica”);
- problemas de capitalização (uso de maiúscula onde deveria ser minúscula ou uso de minúscula onde deveria ser maiúscula);
- grafia incorreta de palavras homônimas ou parônimas (ex: “mas” x “mais” ou “há” x “a”);
- problemas de segmentação (uso de hífen, palavras juntas que deveriam ser separadas ou palavras separadas que deveriam ser escritas juntas); e
- desvios com relação à nova ortografia.

Para todos esses casos, a abordagem baseada em regras precisa identificar corretamente o contexto em que a palavra-alvo está inserida. O que torna essa tarefa complexa e nem sempre bem sucedida é que a identificação do contexto linguístico em uma abordagem simbólica muitas vezes depende de um bom *parser* e um bom *tagger*. Conforme foi apresentado em capítulos anteriores (Capítulo *Sequência de caracteres e palavras* e Capítulo *Ferramentas e recursos para o processamento sintático*), essas ferramentas nem sempre têm uma ótima performance em português.

7.2.1.2 Desvios gramaticais

Os desvios gramaticais, também chamados de desvios sintáticos, correspondem aos problemas de estrutura sintática, ou seja, nas relações entre as palavras, que podem estar no escopo de uma sentença, um sintagma, um grupo ou uma *string*. Por exemplo, na sentença “As menina dançam”, existe um desvio de concordância nominal entre “As” (plural) e “menina” (singular) e/ou um desvio de concordância verbal entre “menina” (singular) e “dançam” (plural).

Além dos desvios de concordância (que correspondem a cerca de 18,9%), também são comuns em redações escolares: os de vírgula e pontuação (44%), de formas verbais (6,8%), pronomes (5,8%), preposições (5,7%), crase (4,2%), segmentação (4,1%), regência (3,4%), conjunções (2,3%), determinantes (2%) e outros (2,3%)¹⁸.

¹⁷<http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/index.html>

¹⁸Esses percentuais foram calculados a partir dos números absolutos da Tabela 8 de Ramisch (2020, p. 76), que anotou os desvios sintáticos em uma amostra de 1.045 redações.



Apesar de os desvios de pontuação e vírgula serem os mais frequentes, são também os mais difíceis de serem tratados, pois em geral a vírgula separa constituintes sintagmáticos, o que exigiria uma análise sintática por constituintes, e esse tipo de *parser* é raro para o Português¹⁹.

As regras gramaticais que exigem um contexto linguístico local, por exemplo, para avaliação da crase ou concordância nominal, geralmente funcionam melhor, ao passo que as regras que dependem de um contexto linguístico maior, com macrorrelações de dependência (Capítulo [Sequência de caracteres e palavras](#)), ou quando um *token* está muito distante do outro, tendem a performar mal.

7.2.1.3 Outros tipos de desvios

Além dos desvios ortográficos e gramaticais, há outros problemas textuais que podem ser marcados automaticamente em redações, como:

- Desvios de vocabulário e registro – quando o texto apresenta escolhas lexicais inadequadas ao gênero ou à situação comunicativa. Por exemplo, em uma dissertação argumentativa, o uso excessivo de marcas de oralidade (“tipo”, “a gente”) ou de pronomes em primeira pessoa (“eu acho”, “na minha opinião”) pode ser considerado um desvio de registro.
- Desvios de coesão – referentes à falta de recursos coesivos ou ao emprego incorreto de conectivos, pronomes ou expressões responsáveis por articular as ideias do texto. Um exemplo recorrente é o uso de “contudo” no sentido de conclusão, ou o uso do pronome “onde” em contextos não locativos (“A época onde...”).
- Desvios de adequação ao gênero – quando o aluno não segue as convenções esperadas para o gênero textual proposto. Quando o aluno conta histórias ou traz longos trechos narrativos em dissertações argumentativas, por exemplo, pode configurar desvio.
- Desvios de progressão temática – quando há repetição excessiva de ideias, incoerências internas, rupturas bruscas na argumentação ou inserção de trechos desconectados do tema central, prejudicando a continuidade do texto.

Enfim, dependendo do modelo de correção adotado e da parametrização dos critérios de correção de cada banca, definem-se quais tipos de desvios devem ou não ser marcados no texto. A detecção desses desvios pode ser feita por meio de regras baseadas em listas lexicais, padrões sintáticos ou análise do contexto discursivo. Assim como nos casos ortográficos e gramaticais, as abordagens mais eficazes são aquelas que conseguem levar em conta o contexto em que a palavra ou expressão ocorre, reduzindo falsos positivos. Cabe destacar que, além da marcação de desvios, é igualmente relevante identificar e registrar aspectos positivos presentes na redação. A sinalização de usos adequados e escolhas bem-sucedidas contribui não apenas para reforçar práticas corretas, mas também para justificar a atribuição de notas mais altas nos critérios correspondentes.

7.2.2 Formalismos de regras

Há inúmeras maneiras de escrever regras de forma que o computador consiga lê-las e interpretá-las. Cada ferramenta pode criar seu próprio formalismo e mecanismo de infe-

¹⁹O *parser* PALAVRAS (Bick, 2000) dispõe de um módulo de análise sintática por constituintes.



rência, mas também há alguns disponíveis gratuitamente e que podem ser usados para um projeto inicial.

O LanguageTool²⁰ implementa um mecanismo de inferência para regras formalizadas em XML (*Extensible Markup Language*). Para o português, o software disponibiliza cerca de 2.880 regras abrangendo várias categorias linguísticas, tais como: gramática geral, ortografia, pontuação, capitalização, tipografia, estilo, redundância, palavra composta, semântica, repetição, linguagem informal, uso de pronomes, dentre outras. Essas regras podem ser consultadas via repositório Language Tool Community²¹.

Como exemplo, reproduzimos o formalismo de uma regra para identificar redundância quando se escreve “gelo gelado”, na Figura 7.1²².

Figura 7.1: Exemplo de formalismo de regra do LanguageTool

```

1 <rule id="GELO_GELADO" name="Gelo ">
2   <pattern>
3     <marker>
4       <token>gelo</token>
5       <token>gelado</token>
6     </marker>
7   </pattern>
8   <message>
9     Substitua «gelo gelado» por <suggestion>gelo</suggestion>.
10  </message>
11  <example correction="gelo">
12    Vendemos uma pedra de <marker>gelo gelado</marker>.
13  </example>
14 </rule>

```

Neste exemplo, consta o id e o nome da regra (linha 1), seguidos do padrão a ser buscado (linhas de 2 a 7), seguido da mensagem a ser mostrada (linhas 8 a 10), e de um exemplo de uso (linhas 11 a 13).

A complexidade das regras pode variar dependendo da complexidade do problema linguístico ou do padrão a ser buscado. No caso do código na Figura 7.1, o problema linguístico em questão – a redundância – é muito simples, e isso se reflete na simplicidade da regra, a qual procura basicamente dois *tokens*: o primeiro é “gelo”, imediatamente seguido do segundo, que é “gelado”. Por outro lado, problemas linguísticos mais complexos também exigem regras mais complexas que podem usar lemas, *tokens*, etiquetas morfológicas, morfossintáticas, expressões regulares, relações de dependência, entidades nomeadas, dentre outros.

Ressaltamos que a performance das regras do LanguageTool não é ótima, mas é um recurso útil para quem não quer começar essa tarefa do zero. Considerando que o software possui versão aberta²³, é possível corrigir e definir novas regras usando o mesmo formalismo e avaliá-las por meio da própria ferramenta.

Outra ferramenta que podemos indicar para esse tipo de tarefa é o módulo Python

²⁰<https://languagetool.org/pt-BR/>

²¹<https://community.languagetool.org/rule/list?lang=pt>

²²Fonte: regras para português no repositório languagetool (Github) (<https://github.com/languagetool-org/languagetool/blob/50c9a5eb145f6289762fc64a2b8773629ca085e1/languagetool-language-modules/pt/src/main/resources/org/languagetool/rules/pt/pt-BR/style.xml#L142-L150>).

²³<https://github.com/languagetool-org/languagetool>



spaCy²⁴, que implementa três mecanismos para identificação de padrões em textos que podem ser bastante úteis na tarefa de detecção de desvios. Esses mecanismos fazem parte do sub-módulo chamado *Rule-based matching*, que permite a busca por um *token* em determinado contexto (chamado *Matcher*), por uma frase ou sintagma (chamado *Phrase matcher*), ou ainda por relações de dependências entre elementos da sentença (chamado *Dependency matcher*). Eles podem ser usados separadamente ou combinados entre si para garantir melhor acurácia na busca por padrões linguísticos.

Na Figura 7.2, apresentamos código Python que utiliza a classe `Matcher` do spaCy para a definição da regra que identifica a redundância “gelo gelado”, que foi reescrita, e executa a busca por padrões em um texto.

Figura 7.2: Exemplo de formalismo da regra “gelo gelado” usando a biblioteca spaCy

```

1 import spacy
2 from spacy.matcher import Matcher
3
4 nlp = spacy.load('pt_core_news_sm')
5 matcher = Matcher(nlp.vocab)
6
7 rule_id = 'GELO_GELADO'
8
9 pattern = [
10     {'LOWER': 'gelo'},
11     {'LOWER': 'gelado'}
12 ]
13
14 message = 'Substitua «gelo gelado» por "gelo"'
15
16 matcher.add(
17     rule_id,
18     [pattern],
19     on_match=lambda matcher, doc, i, matches: print(message)
20 )
21
22 doc = nlp('Gelo gelado gela a garganta.')
23 matches = matcher(doc)

```

O exemplo inicia importando o módulo `spacy` e, especificamente, a classe `Matcher` (linhas 1 e 2). Em seguida um *pipeline* pré-treinado do spaCy para português é carregado (linha 4) e seu vocabulário é utilizado para inicializar uma instância da classe `Matcher` (linha 5). Em seguida, são definidos um identificador para a regra (linha 7), o padrão buscado de dois *tokens* (linhas 9 a 12), cada um representado por um dicionário, e a mensagem que deve ser impressa na tela caso o padrão seja identificado no texto (linha 14). Nas linhas de 16 a 20, a regra é adicionada à instância `matcher`, incluindo a definição de uma função para a impressão da mensagem na tela quando o padrão é encontrado (`on_match`). As linhas 22 a 23 definem um texto para teste de busca do padrão e a execução dessa busca.

O spaCy não conta com um repositório de regras pré-definidas para detecção de desvios. Contudo, por ser uma ferramenta de PLN, disponibiliza uma série de funcionalidades que podem contribuir para essa tarefa de maneira mais simples, i.e. sem a necessidade de

²⁴<https://spacy.io/>



alterar a implementação dos mecanismos de busca já disponíveis.

A detecção de desvios no texto é uma etapa importante em CAR, especialmente por indicar e colaborar para a aprendizagem da escrita. Os desvios encontrados podem, inclusive, ser utilizados na etapa de atribuição de nota à redação. Na Seção 7.3 apresentamos as principais abordagens e tendências nessa área, além de citar os principais trabalhos dedicados a redações em português.

7.3 Atribuição de nota

A atribuição de nota a uma redação pode ser feita de forma global, ou seja, uma nota única para a redação inteira, ou por meio de notas individuais para cada critério de avaliação. O principal desafio desse processo está na grande diversidade de matrizes de correção existentes (ver Seção 7.1.3) – que variam quanto ao modelo adotado, à escala e às faixas de pontuação, bem como à quantidade e à distribuição dos critérios avaliativos.

Até recentemente, as abordagens faziam uso de *corpus* rotulado, i.e., conjuntos de redações que já foram avaliadas manualmente e possuem indicação de nota e/ou adequação da redação em relação ao critério avaliado. Desse modo, as técnicas utilizadas para atribuição de nota se enquadram na área de aprendizado supervisionado por classificação ou regressão. Hoje, com a popularização dos LLMs, muitas empresas optam por essa abordagem gerativa, por ser mais rápida e barata, apesar de menos controlável.

O *Project Essay Grade* (PEG) (Ajay et al., 1973) foi uma das primeiras ferramentas estáveis para a atribuição de notas em redações com boa performance dentro do contexto aplicado: redações universitárias curtas em inglês. No entanto, a falta de acesso a computadores foi, por algum tempo, impedimento para o desenvolvimento de outras soluções. Na metade da década de 90, dados os avanços tecnológicos de hardware e software, a área de AES viu um reaquecimento e, desde então, surgiram novos trabalhos consistentemente, inclusive apoiados por abordagens que tiveram ascensão a partir da década de 2010, como *deep learning* e Transformers.

Como mencionado na Seção 7.1, é importante conhecer o contexto e modelo de correção para realizar a atribuição de nota de forma efetiva. A despeito disso, diferentes estratégias podem ser reaproveitadas e combinadas para a avaliação de redações de modelos de correção distintos. Na Seção 7.3.1 apresentamos uma visão geral de técnicas e estratégias para a atribuição de notas em redações. Dada a relevância do Enem para o contexto de redações em português, a Seção 7.3.2 traz trabalhos especificamente voltados para a automatização da avaliação em redações desse modelo de correção.

7.3.1 Como atribuir nota a redações?

A abordagem clássica para atribuição de notas envolve a extração de atributos (*features*) a partir do texto, que são utilizados para descrever redações de um conjunto de treinamento, além da transformação e seleção desses atributos, em um processo nomeado engenharia de atributos. Os dados extraídos servem de entrada para um algoritmo de aprendizado para a geração de modelos capazes de atribuir nota a novas redações a partir dos valores de seus atributos.

A primeira versão do PEG (Ajay et al., 1973) utilizava atributos baseados em contagens de diferentes elementos do texto, categorizadas em: (i) **simples** (e.g. número de adjetivos na redação): redações com mais adjetivos são avaliadas com notas maiores por humanos



(relação linear); (ii) **enganosamente simples** (e.g. número de palavras na redação): redações muito curtas são penalizadas, porém, conforme o tamanho da redação aumenta, esse atributo perde importância para atribuição de nota (relação logarítmica); e (iii) **s sofisticadas** (e.g. número de palavras que podem representar contextos maiores): o número de conectivos, por exemplo, pode indicar a complexidade de uma sentença.

Considerando a tarefa de atribuição de nota, é possível utilizar atributos que sejam independentes. Ferramentas como Coh-Metrix²⁵ e Linguistic Inquiry Word Count (LIWC)²⁶, são utilizadas em trabalhos como Ferreira et al. (2021) e Ferreira Mello et al. (2022) para a extração de informações linguísticas, como legibilidade e coesão.

Trabalhos que utilizam métricas independentes de conteúdo são capazes de representar critérios de avaliação como Coerência e Coesão. No entanto, critérios como Tema são melhor avaliados por atributos dependentes de conteúdo, como exemplo as matrizes de termos, descritas no Capítulo **Conjunto de dados, dataset e corpus**, e métricas calculadas a partir dessas matrizes, como a similaridade de cosseno utilizada entre tema e redação em Amorim; Veloso (2017). Em Louis; Higgins (2010) e Persing; Ng (2014), são propostos cálculos de atributos dependentes de conteúdo com base em recursos linguísticos pré-definidos e associados aos temas relacionados às redações utilizadas nos experimentos.

Ainda sobre a extração de atributos, vale mencionar o trabalho de Sousa et al. (2021) que, além de aspectos linguísticos, explora aspectos relacionados à construção da argumentação, por meio de mineração de argumentos. A combinação de diferentes estratégias para extração de atributos é bastante comum, conforme realizado por Amorim; Veloso (2017) que, além de aspectos linguísticos e associados ao tema, incluem métricas associadas ao correto uso da língua, calculadas com base em desvios identificados por ferramentas externas, como as mencionadas na Seção 7.2.

É importante ressaltar que a inclusão de atributos relacionados a critérios de avaliação específicos não é imprescindível para atribuição de nota global. No entanto, a partir do momento em que se propõe atribuir notas por critérios avaliativos, é interessante incluir atributos que representem cada critério, ou poderá haver discrepância significativa no resultado obtido entre critérios, como observado em alguns trabalhos (Amorim; Veloso, 2017; Fonseca et al., 2018).

Selecionado um conjunto de atributos e realizada a análise estatística dos dados, podemos seguir à etapa de treinamento de modelos. Não convém aqui sugerirmos esta ou aquela técnica ou algoritmo, uma vez que conjuntos de dados distintos podem apresentar resultados também diferentes para os mesmos algoritmos (Ferreira Mello et al., 2022; Ferreira et al., 2021; Fonseca et al., 2018; Marinho et al., 2022a). Ao treinar modelos para atribuição de nota, assim como modelos com outros objetivos, é fundamental definir mais de um algoritmo e configurações para, então, realizar uma comparação estatística entre os resultados obtidos.

Entre os trabalhos que utilizam a abordagem de extração de atributos, há modelos de classificação e regressão treinados com diversos algoritmos, como: regressão linear (Fonseca et al., 2018), *Support Vector Machines* (SVM) (Haendchen Filho et al., 2018, 2019), *Gradient Boosting* (Fonseca et al., 2018; Marinho et al., 2022a). A comparação entre os modelos se dá, principalmente, pela avaliação dos valores obtidos para métricas

²⁵Coh-Metrix é uma ferramenta computacional que calcula métricas e índices para aspectos linguísticos e discursivos em um texto e que será melhor explorada neste capítulo na Seção 7.4.1. Disponível em: <http://cohmetrix.memphis.edu/cohmetrixhome>.

²⁶LIWC é uma ferramenta computacional, que realiza análise de textos baseada em métricas. Disponível em: <https://www.liwc.app/>.



como precisão, revocação, medida-F, RMSE e Kappa de Cohen.

Embora seja possível obter resultados satisfatórios pela engenharia de atributos e treinamento de modelos por algoritmos clássicos de aprendizado de máquina, é notável o esforço humano necessário para o processo de extração e seleção de atributos, considerando que muitos dos conjuntos de atributos são compostos por algumas centenas de métricas. Com isso em vista, surgem trabalhos que utilizam outras técnicas para a representação de textos e algoritmos de redes neurais profundas para a tarefa de atribuição de nota.

Alikaniotis et al. (2016) propõem uma técnica de *word embeddings* treinada com base em notas de redações, que é utilizada com redes neurais LSTM. O trabalho relata melhores resultados obtidos em comparação com outras abordagens.

Em Fonseca et al. (2018), as *word embeddings* GloVe são combinadas com redes LSTM bidirecionais e os resultados são comparados, também, com uma abordagem que utiliza engenharia de atributos. Os autores relatam que, embora a técnica de redes neurais tenha gerado bons resultados, o modelo gerado a partir de atributos se mostrou superior em diferentes aspectos.

Mayfield; Black (2020) realizam *fine-tuning* de modelos pré-treinados (BERT e variações) para a atribuição automática de notas. Apesar de relatar resultados até 5% melhores do que modelos baseados em n-gramas, os autores discutem sobre o tempo de treinamento deste tipo de modelo, que é cerca de 100 vezes mais demorado do que outras abordagens, e sobre o impacto que isso pode ter em fluxos mais dinâmicos de trabalho.

Bittencourt Jr. (2020) define 14 técnicas baseadas em combinações de diferentes representações de palavras e arquiteturas de redes neurais profundas para a execução da tarefa de atribuição de nota a redações. Os experimentos são realizados com um conjunto composto por redações de 18 temas, sendo que cada técnica é utilizada para o treinamento de um modelo por tema (18 modelos por técnica). Também é proposta uma abordagem para treinamento de modelo multi-tema, ou seja, um modelo único para a atribuição de notas para redações de mais de um tema.

O trabalho de Marinho et al. (2022a) compara 3 tipos de abordagens: (i) engenharia de atributos com algoritmo de regressão, (ii) *doc embeddings* com algoritmo de regressão e (iii) *word embeddings* com LSTM. As abordagens (i) e (iii) apresentaram melhores resultados para critérios de avaliação distintos, sendo a abordagem (iii) eleita pelos autores como a melhor. Os resultados da abordagem (iii) ainda foram comparados com resultados de Amorim et al. (2018) e Fonseca et al. (2018), sendo relatado melhor desempenho desta abordagem na atribuição de nota por critério de avaliação.

Nos últimos anos, os LLMs também têm sido explorados para a atribuição automática de notas a redações. Esses modelos podem ser aplicados tanto em cenários de *fine-tuning* supervisionado (Yang et al., 2020b), a partir de bases de redações previamente anotadas, quanto em cenários de *few-shot* ou *zero-shot learning* (Shibata; Miyamura, 2025), em que o próprio modelo, orientado por instruções, realiza a atribuição de notas sem necessidade de treinamento extensivo (Pack et al., 2024). Estratégias híbridas também têm sido propostas, como a incorporação de *shallow features* ou de comparações par-a-par no processo de atribuição de nota, com ganhos de desempenho em relação a métodos tradicionais (Faseeh et al., 2024; Shibata; Miyamura, 2025).

A principal vantagem dessa abordagem é a redução do esforço manual na engenharia de atributos e a possibilidade de maior generalização a diferentes critérios avaliativos, gêneros textuais e modelos de correção. Além disso, *frameworks* inovadores como o Rank-Then-Score (Cai et al., 2025) e abordagens colaborativas entre humanos e LLMs (Xiao et al., 2025) têm demonstrado resultados promissores, combinando consistência estatística e



explicabilidade. Por outro lado, os LLMs ainda apresentam desafios importantes, como custo computacional elevado, necessidade de dados para ajuste fino, bem como riscos de variação e falta de interpretabilidade. Apesar dessas limitações, os LLMs podem se consolidar como alternativa barata para avaliação automática de redações em larga escala, inclusive em cenários multilíngues ou em línguas com menos recursos computacionais, como é o caso do português.

7.3.2 Atribuição de nota para redações do Enem

Especificamente para o português, a maior parte dos trabalhos relacionados à atribuição de notas (e CAR) utiliza *corpora* compostos por redações do modelo de correção do Enem como base de treinamento. Dada a importância e dimensão do exame no Brasil, há interesse particular em encontrar soluções para a atribuição de nota exclusivamente para esse modelo de correção.

Como descrito na Seção 7.1.3, o Enem exige a produção de um texto do gênero dissertativo-argumentativo sobre um tema específico que é avaliado em 5 critérios, também chamados de competências: (1) Língua portuguesa, (2) Abordagem temática e adequação ao tipo textual, (3) Progressão textual e defesa do ponto de vista, (4) Coesão e articulação e (5) Proposta de intervenção. Os trabalhos que treinam modelos de atribuição de nota para o Enem predizem uma nota global, porém alguns tentam também aperfeiçoar por competência.

Barbosa de Lima et al. (2023) fazem uma revisão sistemática da literatura envolvendo avaliação automática de redações do Enem. Entre as conclusões apresentadas, destacamos o foco dos trabalhos selecionados no uso de atributos extraídos do texto em vez do uso de modelos pré-treinados baseados em Deep Learning, o baixo detalhamento dos feedbacks gerados e a escassez de análise do impacto prático em aplicações no mundo real. Alguns dos trabalhos apontados por Barbosa de Lima et al. (2023) são detalhados a seguir, por apresentarem relevância teórica ou técnica.

Em Amorim; Veloso (2017), Fonseca et al. (2018), Marinho et al. (2022a) e Bittencourt Jr. (2020), o foco está na atribuição de notas para cada uma das competências, o que pode ser feito com base em modelos treinados para cada competência ou um modelo único que prediz as notas para cada uma delas. Já em Haendchen Filho et al. (2018), é explorada a atribuição de notas para a competência 2, especificamente.

Ao realizar estudo sobre a predição de notas para cada uma das competências do Enem, Haendchen Filho et al. (2019) notaram o significativo desbalanceamento do conjunto de redações e tornaram esse o foco de seu trabalho, a fim de analisar o impacto e tratamento de conjuntos de dados desbalanceados na tarefa de atribuição de nota.

Alguns trabalhos que utilizam redações em português não têm como foco direto a atribuição de nota, mas a proposta de técnicas mensuráveis relacionadas a critérios cuja avaliação pode ser mais complexa. Como exemplo, citam-se as contribuições de Ferreira et al. (2021), Sousa et al. (2021) e Ferreira Mello et al. (2022) para a avaliação das competências 3 e 4.

É notável que, no momento da escrita deste capítulo, não pudemos encontrar nenhum trabalho em que se dê atenção em particular para a melhoria de atribuição de nota na competência 5 do Enem.

Vale ressaltar que os conjuntos de dados utilizados pelos referidos trabalhos não são muito representativos, possuindo até alguns milhares de redações de uma baixa diversidade de temas. O maior conjunto relatado possui 56.644 redações, sem indicação de número



de temas (Fonseca et al., 2018). O conjunto de redações com maior número de temas relatado, que também é o segundo em número de redações, conta com 27.184 redações distribuídas entre 18 temas, sendo que o número de redações por tema varia entre 3.070 e 710 (Bittencourt Jr., 2020). Além disso, ambos os maiores conjuntos foram fornecidos por empresas privadas e, portanto, não são públicos.

O tamanho e distribuição do conjunto de dados são considerados obstáculos para o treinamento de modelos de atribuição de notas, especialmente quando utilizadas técnicas de *deep learning*. Mesmo com o uso de abordagens híbridas ou baseadas em redes neurais profundas, a comprovação e generalização de resultados é um desafio. No entanto, há uma iniciativa para a criação de um conjunto público de redações do modelo Enem para utilização em trabalhos de CAR: até agosto de 2022 era composto por 6.579 redações pré-processadas e divididas em 151 temas (Marinho et al., 2022b) ²⁷.

Para o português brasileiro e para o contexto da avaliação do ENEM, não foram encontradas abordagens baseadas em LLMs voltadas para a correção ou avaliação de redações. Identificamos apenas o trabalho de Locatelli et al. (2025), que exploraram o comportamento de LLMs (GPT-3.5, GPT-4 e MariTalk) no contexto do ENEM, mas no sentido de comparar redações geradas pelos modelos com aquelas escritas por alunos. Portanto, não tiveram o intuito de avaliar automaticamente os textos.

Também não encontramos nenhum trabalho de PLN que tenha relatado a atribuição de notas em redações de outros modelos de correção, como Fuvest, Unicamp, FGV ou outros.

Enfim, acreditamos que ainda há espaço para trabalhos quanto à tarefa de atribuição de notas em redações em português. Contudo, para atingir a meta de soluções completas de correção de redação, apenas a nota é insuficiente do ponto de vista do processo de ensino e aprendizagem. Para suprir essa lacuna, a Seção 7.4 discute a terceira tarefa de CAR, referente ao provimento de feedback relacionado ao texto.

7.4 Feedback para o aluno

Conforme apresentado na Introdução (Seção 7.1), a última etapa da Correção Automática de Redação (CAR) é o fornecimento de um feedback para o aluno. Até pouco tempo atrás, a correção automática produzia basicamente uma nota como resultado da avaliação da redação. Mas isso já não era mais suficiente e foi surgindo a necessidade de explicar ou justificar essa nota. De acordo com Shermis; Burstein (2013), os primeiros trabalhos se limitavam a dar feedbacks sobre as características e propriedades linguísticas do texto. Pesquisas mais recentes vêm focando em aspectos mais complexos e profundos da língua, que vão além da superficialidade do texto ²⁸.

Em uma **correção manual**, esse feedback é feito pelo próprio corretor da redação, na forma de comentário livre, em linguagem natural, sem seguir nenhum tipo de padronização, podendo tecer críticas, fazer sugestões, elencar pontos fortes e pontos a melhorar, abordar questões gerais ou específicas da redação, enfim, de formas bastante variadas.

Já em uma **correção automática**, as plataformas que dão algum tipo de feedback sobre a correção o fazem de forma sistematizada. Porém, são raras as empresas que fornecem esse tipo de devolutiva ao aluno. Lima et al. (2023) fizeram uma revisão sistemática da

²⁷<https://github.com/lplnufpi/essay-br>

²⁸Vale ressaltar que os feedbacks baseados em características e propriedades linguísticas do texto ainda são os mais praticados hoje pelas plataformas brasileiras ou que processam o português, então focaremos nessa abordagem ao longo desta seção.



literatura sobre CAR e uma das lacunas que identificaram nos trabalhos para o português é o baixo detalhamento nos feedbacks retornados pelos modelos de avaliação.

Na prática, os corretores automáticos costumam apontar apenas estatísticas básicas do texto, tais como quantidade de conectivos (conjunções), variação lexical (taxa de *types* por *tokens*), quantidade de palavras de conteúdo (substantivos, adjetivos, verbos e alguns advérbios), tamanho médio das palavras, frases e parágrafos, dentre outros, o que geralmente não tem utilidade pedagógica para o aluno. A Seção 7.4.1 apresenta como essas informações são calculadas e exibidas.

Algumas plataformas de CAR também disponibilizam para o aluno sistemas ou *bots* baseados em assistentes de escrita ou ferramentas computacionais de auxílio à escrita. Na Seção 7.4.2 apresentamos como esses recursos e ferramentas são utilizadas em sistemas de CAR.

Mais recentemente, com o surgimento e popularização de LLMs, algumas empresas também já começaram a fornecer feedbacks gerados automaticamente por esses modelos gerativos. Também é possível gerar automaticamente as devolutivas a partir de elementos encontrados ou não encontrados no texto, instanciando palavras ou trechos do texto da redação. Mas isso só é possível se for usada uma abordagem simbólica. Nesse sentido, o feedback pode conter críticas referenciando os desvios apresentados na Seção 7.2 e/ou elogios aos pontos fortes, como será apresentado na Seção 7.4.3.

7.4.1 Estatísticas básicas do texto

Algumas plataformas e empresas privadas que oferecem serviço de CAR apresentam para o aluno contagens básicas do texto, tais como a quantidade de palavras, caracteres, sentenças, parágrafos e até a quantidade de palavras por classe gramatical (verbos, substantivos, adjetivos, preposições, conjunções etc.). Outras oferecem um pouco mais de informação baseada em estatísticas simples, como a proporção de palavras únicas (*types*) em relação à quantidade total de palavras no texto (*tokens*), alguma medida de similaridade entre as sentenças, desvio padrão dos parágrafos, dentre outras.

Um dos recursos disponíveis para recuperar essas informações é o NILC-Metrix (Leal et al., 2023), uma versão brasileira do Coh-Metrix. O NILC-Metrix²⁹ é a atualização mais recente do Coh-Metrix-Port (Scarton; Aluísio, 2010), contendo 200 métricas³⁰ distribuídas nas 14 categorias apresentadas no Quadro 7.2, as quais avaliam a coerência, a coesão, a inteligibilidade, a complexidade e outros aspectos:

Quadro 7.2: Categorias de métricas disponíveis no NILC-Metrix.

Categoria	Qtde.	Descrição das métricas
Medidas Descritivas	10	Quantificam sílabas por palavra, palavras por sentença, sentenças por parágrafo, assim como quantidades absolutas de palavras, sentenças e parágrafos.
Simplicidade Textual	8	Analisa proporção de palavras fáceis e difíceis em relação ao total de palavras, bem como sentenças longas e curtas.
Coesão Referencial	9	Avaliam quantidades de palavras de conteúdo, radicais de conteúdo, médias de referentes, de candidatos a referentes e proporção de pronomes anafóricos.

²⁹<http://fw.nilc.icmc.usp.br:23380/metrixdoc>

³⁰Definição, explicação e exemplos das métricas podem ser conferidos na Documentação do NILC-Metrix (<http://fw.nilc.icmc.usp.br:23380/metrixdoc>).



Coesão Semântica	11	Calculam similaridade, entropia, desvio padrão e taxa de <i>givenness</i> , ou seja, quantidade de informação dada e nova baseada em LSA (Capítulo Representação vetorial e semântica distribucional).
Medidas Psicolinguísticas	24	Verificam 4 critérios: concretude, imageabilidade, familiaridade e idade de aquisição das palavras.
Diversidade Lexical	15	Calculam proporções e desvio padrão de <i>types</i> e <i>tokens</i> , tanto gerais quanto por PoS e por categorias de palavras lexicais e funcionais.
Conectivos	12	Verificam a proporção de vários tipos de conectivos (aditivos, causais, lógicos) e operadores lógicos (positivos e negativos) em relação ao total de palavras do texto.
Léxico Temporal	12	Consideram verbos flexionados nos diferentes tempos e modos verbais, bem como as formas regular e irregular de particípio.
Complexidade Sintática	27	São bastante diversas, considerando desde a quantidade de orações por sentença, coordenação e subordinação, até aspectos como voz ativa e passiva, apostro, distância na árvore de dependência, fórmulas de Frazier e Yngve, dentre vários outros.
Densidade de Padrões Sintáticos	4	Calculam proporção de verbos no gerúndio e também tamanhos de sintagmas nominais.
Informações Morfossintáticas de Palavras	42	Verificam quantidade média, proporção ou desvio padrão das palavras por PoS.
Informações Semânticas de Palavras	11	Incluem dados sobre substantivos abstratos, palavras polissêmicas, hiperônimos, nomes próprios e polaridade (positiva e negativa) das palavras.
Frequência de Palavras	10	Calculam frequências de vários tipos de palavra (de conteúdo, e.g.) com base na curva de Zipf e considerando diferentes <i>corpora</i> .
Índices de Leiturabilidade	5	Incluem índices e fórmulas já consolidados em PLN, tais como Fórmula Dale Chall, Índice de Brunet, Flesch, Gunning Fog e Estatística de Honoré.

Os cálculos dessas métricas geralmente resultam em um valor numérico, o qual não se faz útil para o aluno. Porém, há diferentes maneiras de devolver ao aluno um feedback textual com a interpretação de algumas dessas métricas. Por exemplo, se considerarmos os valores de 4 métricas de Simplicidade Textual, referentes a tamanho de sentença (a saber: *long_sentence_ratio*³¹, *medium_long_sentence_ratio*³², *medium_short_sentence_ratio*³³, *short_sentence_ratio*³⁴), é possível criar um resultado interpretável para dizer ao aluno que ele constrói sentenças muito longas e isso pode prejudicar a compreensão das ideias do texto.

Tanto as estatísticas básicas quanto as métricas do NILC-Metrix podem ser utilizadas não apenas para devolver feedbacks aos alunos, mas também como atributos para calcular a nota da redação ou de alguns aspectos da redação, conforme apresentado na Seção 7.3.

7.4.2 Assistentes de escrita e ferramentas de auxílio à escrita

Para prover uma devolutiva ao aluno, também é possível recorrer a sistemas prontos de PLN, como os assistentes virtuais, assistentes de escrita ou ferramentas de auxílio à escrita. Essas soluções podem ser entendidas como aplicações finais, mas, na área de CAR, elas são usadas como recursos ou ferramentas intermediárias para subsidiar a solução completa de CAR.

Essas ferramentas são capazes de gerar, melhorar, reformular e personalizar qualquer tipo de conteúdo textual, incluindo redações. Algumas delas funcionam de forma síncrona *real-time*, fazendo correções e dando sugestões à medida que o texto está sendo escrito,

³¹Proporção de sentenças muito longas em relação a todas as sentenças do texto.

³²Proporção de sentenças longas em relação a todas as sentenças do texto.

³³Proporção de sentenças médias em relação a todas as sentenças do texto.

³⁴Proporção de sentenças curtas em relação a todas as sentenças do texto.



enquanto outras funcionam a posteriori, ou seja, depois que o aluno submete sua redação à plataforma de correção, ele recebe uma devolutiva com críticas e/ou elogios.

Para a língua inglesa, há inúmeros assistentes de escrita e muitos deles conhecidos no Brasil porque as pessoas usam o inglês para escrever, por exemplo, artigos científicos. Um dos mais populares é o Grammarly³⁵, mas também há outros bastante usados, como Linguix³⁶, Ginger³⁷, Reverso³⁸, Writer³⁹, Hemingway App⁴⁰ e outros.

Para o português, também existem vários softwares comerciais, sendo a maioria paga. As ferramentas de auxílio à escrita, ao lado dos simplificadores textuais e dos sumarizadores automáticos, podem contribuir com a área de CAR, pois fornecem:

- **Correção ortográfica e gramatical:** Os sistemas podem usar regras ou modelos de linguagem treinados em um grande volume de textos em português para identificar erros ortográficos e gramaticais comuns.
- **Análise de contexto:** As ferramentas não apenas verificam palavras isoladas, mas também consideram o contexto da frase em que uma palavra está inserida. Isso ajuda a evitar falsos positivos e permite que o sistema forneça sugestões de correção mais precisas.
- **Sugestões de melhoria:** Quando uma palavra é identificada como incorreta ou quando uma construção gramatical suspeita é detectada, o assistente de escrita oferece sugestões para corrigir o problema. Essas sugestões podem incluir substituições de palavras, ajustes na estrutura da frase ou correções de pontuação.
- **Deteção de estilo:** Além de corrigir erros básicos, um assistente de escrita também pode oferecer sugestões para melhorar o estilo de escrita. Isso inclui alertas sobre uso excessivo de palavras, repetições, uso inadequado de voz passiva, entre outros aspectos.
- **Feedback de clareza:** As ferramentas também podem avaliar a clareza do texto, identificando frases longas e complexas que podem ser difíceis de entender, podendo sugerir dividir essas frases ou reformulá-las para tornar o conteúdo mais acessível.
- **Verificação de plágio:** Algumas soluções comerciais oferecem uma funcionalidade adicional para verificar a originalidade do texto, identificando trechos que possam ser semelhantes a outras fontes online. Isso é especialmente útil para evitar acidentalmente usar conteúdo plagiado.
- **Aprendizado contínuo:** Assim como outras ferramentas de PLN, os assistentes de escrita também continuam aprendendo e melhorando com o tempo. Eles são atualizados com novos dados e feedbacks dos usuários, o que ajuda a aprimorar seus modelos e a abordagem dos problemas linguísticos.
- **Extensões e integrações:** Muitos deles oferecem extensões para navegadores, complementos para processadores de texto e aplicativos móveis, o que permite aos usuários verificar seu conteúdo em tempo real enquanto escrevem em várias plataformas.
- **Personalização:** Em alguns desses sistemas, o usuário pode personalizar as configurações com base em suas preferências de estilo e escrita. Isso permite adaptar as sugestões e correções de acordo com o contexto e o público-alvo.

³⁵<https://www.grammarly.com/>

³⁶<https://linguix.com/>

³⁷<https://www.gingersoftware.com/>

³⁸<https://www.reverso.net/tradu%C3%A7%C3%A3o-texto>

³⁹<https://writer.com/grammar-checker/>

⁴⁰<https://hemingwayapp.com/>



Conforme dito anteriormente, as melhores ferramentas de auxílio à escrita que existem hoje para o português são soluções comerciais de empresas privadas. Existem também alguns sistemas desenvolvidos a partir de pesquisas acadêmicas e científicas, mas nenhuma focada em redação. Por exemplo, o SciPo⁴¹ (Feltrim et al., 2003), que é um sistema de auxílio à escrita de resumos acadêmicos em português, especialmente para teses e dissertações da área da Ciência da Computação. Outro exemplo é o WRITEME⁴² (Leite et al., 2020), que é ferramenta de auxílio à escrita de READMEs que usa dados abertos dos repositórios mais populares do GitHub para gerar recomendações de seções, mas também não é focada em redação.

7.4.3 Identificação de pontos fortes e elogiáveis

Na Seção 7.2 falamos da detecção de pontos fracos e desvios no texto. Por outro lado, também é importante detectar pontos fortes e elogiáveis e demonstrá-los ao aluno para que ele continue usando a mesma estratégia nos próximos textos.

Esses pontos fortes podem ser identificados por meio de regras formais, mas também é possível usar diferentes estratégias para cada aspecto da avaliação.

Tendo identificado todos ou alguns aspectos (positivos ou negativos) do texto, é possível retornar essas informações ao aluno na forma de feedbacks construtivos para auxiliá-lo a se tornar um escritor mais habilidoso e confiante.

Avaliação de coesão Na Seção 7.2.1.3, falamos brevemente de como identificar usos corretos de recursos coesivos usando regras em contextos linguísticos menores, como dentro de uma sentença.

Também é possível criar regras formais que percorrem todo o texto procurando as ocorrências de conectivos, avaliar a sua distribuição ao longo do texto, calcular a variabilidade e diversificação deles e até procurar conectivos em pontos específicos da redação, como no início da conclusão, por exemplo.

Com o objetivo de fornecer um feedback baseado na avaliação da coesão do texto, uma solução simples é usar um *tagger* que identifique palavras etiquetadas como conjunções, preposições e advérbios, ou usando listas e léxicos específicos. A outra solução, que é um pouco mais rebuscada, é recorrer às métricas do NILC-Metrix que incidem sobre a coesão textual.

Identificação de repertórios Para avaliar a abordagem temática, referente à competência 2 do Enem, podemos elogiar a presença (ou criticar a ausência) de repertórios socioculturais, que são informações, fatos, citações, definições ou termos de alguma área do conhecimento, ou ainda experiências pessoais que, de alguma forma, contribuem como argumento para defender um ponto de vista.

Pelo Manual de leitura do Enem⁴³, os repertórios socioculturais podem ser legitimados (com citação da fonte) ou não legitimados (sem citação da fonte), ter uso produtivo (pertinente à discussão em mais de um momento do texto) ou não, pertencente ao tema ou não e ainda devem ser penalizados se forem exclusivamente baseados nos textos motivadores. Identificar automaticamente todos esses tipos e usos (corretos ou não) dos repertórios não é uma tarefa simples. Porém isso pode ser

⁴¹<https://escritacientifica.sc.usp.br/scipo/>

⁴²<https://repositorio.ufpe.br/handle/123456789/50043>

⁴³https://download.inep.gov.br/educacao_basica/enem/downloads/2020/Competencia_2.pdf



feito usando modelos de extração de entidades nomeadas (Capítulo 3), buscando, por exemplo, as citações de filósofos, sociólogos e outros estudiosos, ou buscando as menções a livros, filmes, séries, dentre outras entidades que funcionem como repertórios legitimados.

Avaliação da Progressão textual Para a avaliação da progressão textual, é possível treinar e usar modelos de tópico, a exemplo do *Hidden Topic Markov Models* (HTMM) (Gruber et al., 2007), que classificam as sentenças de um texto por tópicos ou assuntos, o que nos permite avaliar a progressão, a continuidade, a retomada e até a circularidade entre os assuntos, a partir da distribuição dos tópicos em um texto.

Blei; Moreno (2001) apresentam resultados dessa abordagem de segmentar um texto não estruturado em tópicos, testando em notícias do New York Times. Os autores propuseram uma combinação do tradicional modelo oculto de Markov (*Hidden Markov Model* – HMM) com o modelo de semântica latente de Hofmann (Hofmann, 1999), resultando em um novo método probabilístico que segmenta um texto em tópicos. Essa abordagem pode ser muito útil para avaliar o encadeamento das ideias de um texto, principalmente no caso de redação do Enem, pois o gênero dissertativo-argumentativo costuma seguir um padrão bem definido de: (i) apresentação do tema, introdução ao ponto de vista a ser defendido e breve menção aos argumentos a serem utilizados (no primeiro parágrafo); (ii) exposição do primeiro argumento (no segundo parágrafo); (iii) exposição do segundo argumento (no terceiro parágrafo); (iv) proposta de intervenção para solucionar o problema discorrido e retomada da tese na forma de conclusão (no quarto parágrafo).

Identificação de proposta de intervenção No Enem, para que uma redação receba nota máxima na Competência 5, o aluno precisa criar uma proposta de intervenção que contenha pelo menos 5 elementos: o agente (quem?), a ação (o quê?), o modo ou meio (como?), a finalidade (para quê?) e o detalhamento de algum dos elementos anteriores.

Para a correta identificação desses elementos, pode-se usar modelos de extração de informação (Capítulo 3) ou recorrer a extração de entidades nomeadas ou recursos linguísticos como listas e léxicos específicos.

O Inep disponibiliza a Cartilha do participante⁴⁴ com instruções sobre agentes que devem ser considerados nulos, ações interventivas que devem ser consideradas nulas, propostas de intervenção negativas ou condicionais, dentre outras orientações que podem se transformar em atributos para modelos.

Não encontramos nenhum trabalho para o português que reporte bons resultados quanto à identificação da proposta de intervenção e que valha ser replicado. É um dos campos de CAR que merece ser mais explorado.

Ao longo desta Seção 7.4, discutimos algumas formas possíveis de devolver um feedback ao aluno, que podem ser: indicando números, percentuais e estatísticas básicas do texto, ou acoplando um assistente de escrita ao corretor automático para fazer isso em tempo real, ou ainda instanciando elementos da redação (recursos coesivos, repertórios, sequência de tópicos, elementos da proposta de intervenção) em uma mensagem gerada automaticamente. Mas as possibilidades não se limitam a essas indicadas neste capítulo. Para outras formas de geração de feedbacks em redações escolares, ver Gamon et al. (2013).

⁴⁴https://download.inep.gov.br/download/enem/cartilha_do_participante_enem_2022.pdf



Tendo em vista todo o conteúdo apresentado na Seção 7.2, Seção 7.3 e Seção 7.4, pode surgir o questionamento sobre o papel (ou até extinção) das correções manuais, dados os avanços em CAR. Para tanto, na Seção 7.5, propomos uma discussão sobre prós e contras de cada uma das abordagens de correção: a manual e a automática, apresentando alguns casos de sucesso e defendendo uma correção híbrida que se beneficie do potencial de cada abordagem.

7.5 Correção manual vs(?) correção automática

A correção automática de redações (CAR) divide opiniões entre alunos, escritores, professores de redação, bancas de avaliação em série, especialistas em Linguística Computacional, cientistas de dados e desenvolvedores de sistemas. Ainda existe muito preconceito quando se trata de correção automática de redação, mas já é consensual aceitar as vantagens dos corretores ortográficos e gramaticais quando embutidos em outras soluções, como no pacote Office, no Gdrive, em redes sociais ou nos teclados dos smartphones.

A discussão principal gira em torno de seus prós e contras, se a correção automática deve substituir ou complementar a correção humana, sobre questões éticas relacionadas à correção automática, sobre a subversão dos valores pedagógicos e educacionais da avaliação manual para uma avaliação automática de textos; enfim, para uma discussão mais filosófica e profunda sobre todos esses aspectos, ver Elliot; Klobucar (2013) e Hakuta (2013).

Nesta seção, abordaremos apenas questões práticas relacionadas à correção manual e à correção automática de redações para, ao final, defender uma correção híbrida, que utilize as principais potencialidades de cada tipo, reconhecendo-se também suas limitações.

7.5.1 Avanços dos últimos anos

Até as décadas de 80 e 90, as avaliações de redação no Brasil eram holísticas, ou seja, o avaliador do texto atribuía uma nota global (de 0 a 100, por exemplo) para a redação, sem seguir rigorosamente nenhum critério previamente estabelecido. Por volta dos anos 2000, essas avaliações passaram a ser analíticas, tendo que explicitar todos os critérios e todos os conceitos que deveriam ser avaliados. Ao mesmo tempo, as avaliações passaram a ser em duplas às cegas, ou seja, cada redação deveria ser avaliada por dois corretores independentes, o que exigia maior sistematicidade e coerência entre eles.

Nessa transição de avaliação holística para analítica, as grades de correção de redações se tornaram mais padronizadas. E sabe-se que tarefas mais padronizadas são melhor executadas por máquinas do que por humanos.

Mesmo com a tentativa (por vezes, falha) de padronização das grades, ainda se percebe a falta de objetividade na definição de critérios por parte de alguns modelos de correção. Quando a grade de correção é muito aberta ou não apresenta os critérios bem definidos para cada faixa de nota, aumentam as chances de haver divergência entre duas avaliações cegas. Por outro lado, quando os corretores humanos passam por treinamentos rigorosos, tal como é feito no Enem, isso pode reduzir o número de inconsistências nas avaliações, mas ainda assim não elimina as divergências, já que pessoas diferentes podem ter interpretações diferentes sobre a mesma instrução. Prova disso são os índices de redações do Enem que vão para uma terceira correção⁴⁵, nos casos de discrepância de 80 ou mais pontos em uma

⁴⁵Os índices de terceira correção variam a cada ano, pois dependem de vários aspectos, inclusive a mudança dos critérios do Inep para a terceira correção. A título de exemplificação, podemos citar o índice de 20,10% em 2012 disponível no Portal do MEC (<http://portal.mec.gov.br/component/tags/tag/corre>



competência ou de 100 ou mais pontos na nota final.

Posto isso, a correção automática no Brasil passou a ser considerada como uma alternativa à manual, já que esta última sempre foi passível de subjetividade e viés.

7.5.2 Vantagens da correção automática

Correções manuais estão sujeitas a subjetividade e viés, além do cansaço humano, a pressão por produtividade, a cobrança por eficiência, o desinteresse pela tarefa, dentre outros fatores que podem prejudicar a qualidade da avaliação ou comprometer sua validade.

Para além dessas questões de limitação humana, é necessário mencionar também o tempo e o custo da correção manual. De acordo com uma matéria veiculada no Portal G1⁴⁶ em 2016, os corretores humanos conseguem corrigir, em média, 74 redações por dia. Já Bittencourt Jr. (2020, p. 19) apresenta uma média de 12 minutos por correção, o que daria 40 redações por dia, considerando-se 8 horas de trabalho. E o custo de cada correção de redação do Enem para o Governo Federal era de R\$15,88 em 2015. No mesmo ano foram corrigidas 6.54 milhões de redações, perfazendo um custo aproximado de R\$104 milhões para o governo. Esse valor provavelmente está defasado, mas foi o último registro oficial encontrado.

Automatizar a correção de redações traz como vantagem a redução do custo de correção e elimina os fatores problemáticos relacionados ao trabalho humano.

Outro aspecto da correção que merece ser comparado é a confiança (ou *reliability*, em inglês). Os sistemas automáticos têm confiança de 100%, o que não pode ser afirmado para a correção manual. Isso significa que toda vez que a mesma redação passar pelo mesmo sistema de correção automática, receberá a mesma correção e a mesma nota. Isso parece óbvio, mas não é o que acontece na correção humana. Diferentes pessoas que corrigirem a mesma redação poderão naturalmente atribuir diferentes notas e/ou apontar diferentes aspectos a serem melhorados. O que também ocorre é que a mesma redação, quando corrigida pelo mesmo corretor humano em diferentes momentos, também pode receber avaliações muito diferentes, o que abre brecha para reclamações.

7.5.3 Vantagens da correção manual

Apesar de todos esses aspectos negativos em relação à correção manual, deve-se ressaltar o ponto forte desse tipo de correção, que é a possibilidade que o humano tem de observar todo e qualquer aspecto relacionado ao processo de construção de sentidos em um texto, o que a máquina não é capaz de fazer.

A produção textual é um processo sócio-cognitivo muito complexo que vai além da capacidade dos sistemas computacionais. A máquina não entende a redação, não interpreta o conteúdo veiculado pelo texto, mas apenas se comporta da forma como ela foi treinada para fazê-lo. Por mais que alguns modelos computacionais possam ser “interpretáveis”, é impossível identificar e definir todos os fatores sociais, psicológicos, cognitivos, emocionais

cao). Também se pode inferir o índice de 43,52% em 2014, a partir DE “Ao todo, foram corrigidos 6.193.565 textos. [...] foram encaminhadas 2.695.949 redações para um terceiro corretor.” disponível no Portal do MEC (<http://portal.mec.gov.br/component/tags/tag/espelho-da-redacao>). Ou uma estimativa de 29% em 2017 “O Inep estima que das 4,1 milhões de redações corrigidas, cerca de 1,2 milhão receberão a terceira correção.” disponível no Portal do MEC (<http://portal.mec.gov.br/component/tags/tag/correcao>).

⁴⁶<https://g1.globo.com/educacao/enem/2016/noticia/corretores-de-redacao-do-enem-avaliam-em-media-74-redacoes-por-dia.ghtml>



etc. que podem interferir tanto no processo de escrita por parte do aluno quanto no processo de correção por parte do avaliador.

Nesse sentido, considerando que a correção automática é limitada, é passível de erros e está mais voltada para a avaliação da forma do que do conteúdo, levanta-se o seguinte questionamento: O uso da correção automática não levaria o aluno a focar sua atenção apenas nos aspectos formais da escrita, excluindo os aspectos mais ricos da construção de sentidos no texto? Por trás desse questionamento, existe uma preocupação legítima de que o aluno não construa sua própria autonomia enquanto escritor, mas apenas seja “adestrado” a escrever de uma forma que o algoritmo lhe dê uma nota boa.

7.5.4 O exemplo da língua inglesa

Para a língua inglesa, algumas instituições educacionais (e.g. **ETS** – *Educational Testing Service*) utilizam modelos de AEE para auxiliar (e não substituir) a correção manual. Esses modelos computacionais são usados como uma segunda avaliação, complementar à avaliação humana. Por exemplo, a avaliação do **TOEFL** (*Test of English as a Foreign Language Internet-based Test*) é dupla, sendo uma feita por humanos e outra feita por sistemas automáticos. A nota final do aluno é dada pela média das duas avaliações. No caso de divergência entre as notas, a redação é enviada a terceiro corretor, semelhante ao que ocorre na avaliação do Enem. Mas, no caso do Brasil, tanto o primeiro quanto o segundo avaliador são humanos. Vale ressaltar que, segundo Bridgeman (2013, p. 227), “a experiência com o programa TOEFL iBT sugere que, quando há discrepância e a redação é enviada a um avaliador humano adicional, esse avaliador tende a concordar com a máquina mais do que com o outro humano”⁴⁷.

Um processo semelhante ocorre na avaliação do **GRE** (*Graduate Record Examination*), mas neste último caso a nota atribuída automaticamente é usada como se fosse uma validação para a avaliação humana. Em outras palavras, a correção automática é usada para monitorar a performance dos corretores humanos, a fim de identificar avaliadores desalinhados ou que precisam passar por novo treinamento.

A correção automática também pode auxiliar a correção manual no sentido de “nivelar” diferentes níveis de rigor. Sabe-se que diferentes avaliadores humanos podem ser sistematicamente mais rígidos ou mais permissivos em suas correções. Segundo Braun (1988, p. 1), “Quando o grau de leniência/severidade do avaliador pode ser atestado adequadamente, é possível calibrar estatisticamente os avaliadores e ajustar as pontuações corretamente [...] Essa calibração estatística parece ser uma abordagem econômica para aumentar a confiabilidade da nota quando comparada ao simples aumento do número de avaliadores por artigo.”⁴⁸.

Nesse sentido, os modelos de AEE podem auxiliar a calibrar essas diferenças de rigidez, atribuindo um peso maior às correções dos avaliadores mais permissivos e um peso menor às correções dos avaliadores mais rigorosos.

⁴⁷Tradução nossa. Do inglês: “*In fact, experience with the TOEFL iBT program suggests that when flagged discrepant scores are sent to an additional human rater, that rater tends to agree with the machine more often than she or he agrees with the other human score.*”

⁴⁸Tradução nossa. Do inglês: “*When rater leniency/severity can be adequately documented, it is possible to statistically calibrate raters and adjust scores accordingly [...] This statistical calibration appears to be a cost-effective approach to enhancing scoring reliability when compared to simply increasing the number of readings per paper.*”



7.5.5 O que defendemos

Levantamos todos esses questionamentos ao longo da Seção 7.5 a fim de tornar explícitas as potencialidades da área de CAR, mas, ao mesmo tempo, esclarecer ao leitor sobre suas limitações, da mesma forma que a correção humana também possui vantagens e desvantagens.

Tendo considerado os vários aspectos das duas abordagens, defendemos neste capítulo uma correção híbrida, semelhante ao que é praticado para a língua inglesa (Seção 7.5.4), que possa se beneficiar dos pontos positivos da correção automática, mas mantendo a correção manual para garantir a responsabilização do humano sobre a avaliação.

7.6 Considerações finais

Neste capítulo exploramos uma das várias aplicações do Processamento de Linguagem Natural (PLN), a chamada Correção Automática de Redação (CAR), a qual abarca duas áreas de PLN em inglês, representadas pelas siglas AES (*Automated Essay Scoring*) e AEE (*Automated Essay Evaluation*).

Ademais, defendemos uma abordagem holística para a CAR, abrangendo, no mínimo, três fases essenciais: (i) a detecção de desvios no texto, (ii) a atribuição de nota, e (iii) a geração de feedback construtivo para o aluno. Apesar de termos dividido essas etapas para fins didáticos, é crucial reconhecer sua interdependência no decorrer do processo. Por exemplo:

1. os desvios gramaticais, ortográficos, de vocabulário etc. podem ser usados como atributos para o cálculo da nota;
2. as métricas e estatísticas básicas do texto, além de serem usadas para criar os feedbacks, também podem servir como atributos para o cálculo da nota;
3. a nota atribuída pelo modelo pode restringir, limitar ou ajudar a selecionar o feedback mais apropriado a ser exibido para o aluno;
4. A caracterização dos desvios (por tipos e quantidades) também pode ser usada para a geração do feedback.

Assim, ainda que tenhamos delimitado didaticamente essas três etapas, é importante ressaltar que, no contexto das tarefas de CAR, tais fases são intrinsecamente entrelaçadas e interdependentes, colaborando harmoniosamente para aprimorar a avaliação da redação.

Embora existam numerosos estudos nesses campos para o inglês e outras línguas, a documentação relevante para o português ainda é escassa e a maioria dos trabalhos acadêmicos confiáveis foi conduzida em pequenas amostras de dados. O progresso mais notável para textos em português provém de empresas e plataformas privadas que oferecem serviços de CAR. No entanto, os métodos e resultados dessas empresas nem sempre são divulgados, e, mesmo se o fossem, seria difícil compará-los devido à falta de uniformidade entre as soluções apresentadas.

Nesse sentido, a área de CAR ainda apresenta um vasto campo de trabalho a ser explorado por novos pesquisadores. Para o português, ainda faltam bons *datasets* de redações, que contenham, além dos textos, as notas por competência, anotação e apontamentos feitos por humanos; também faltam ferramentas robustas de detecção de desvios e de auxílio à escrita, bem como bons *parsers* e *taggers*; e faltam trabalhos que reportem bons resultados, com engenharia de atributos, comparação da performance dos algoritmos utilizados e uma análise aprofundada dos resultados.



Capítulo 8

Detecção Automática de Notícias Falsas

Renato Moraes Silva

Roney Lira de Sales Santos

Thiago Alexandre Salgueiro Pardo

Publicado em: 20/11/2024

Atualizado em: 16/04/2026

 <https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/>

8.1 Introdução

Notícias falsas podem matar, podem destruir vidas e podem interferir nos rumos de povos e nações. Apesar das maravilhas que a tecnologia traz, a própria tecnologia (que inclui dispositivos e serviços, como smartphones, redes sociais e mensageiros instantâneos) tem permitido que notícias falsas se espalhem rapidamente, alcançando milhares de pessoas em pouco tempo, causando danos de todo tipo, afetando desde decisões políticas até a saúde física e mental da população. Grandes eventos e acontecimentos recentes, como eleições presidenciais e a pandemia de COVID-19, trazem muitas evidências do impacto dessas informações enganosas. Mais do que isso, a Inteligência Artificial, apesar dos avanços impressionantes, tem produzido sistemas que, como efeito colateral, podem criar desinformação. Assim, o problema das notícias falsas, embora antigo, ganhou uma nova dimensão com a tecnologia, tornando-se uma ameaça significativa para a sociedade.

Nesse contexto, a tarefa de detecção automática de notícias falsas tem grande apelo junto à sociedade (quer seja em nível individual, quer seja em nível governamental e de gestão), com potencial de trazer contribuições reais, com soluções aplicadas no dia a dia, podendo salvar vidas. Do ponto de vista científico, é também uma tarefa de grande interesse e com desafios instigantes, que incluem caracterizar o estilo de escrita de notícias falsas (e, complementarmente, das verdadeiras), explicar e mapear a distribuição dessas notícias e como elas afetam as várias camadas da sociedade, e como desenvolver sistemas computacionais que as identifiquem automaticamente.

Neste capítulo, apresenta-se o problema e relatam-se os esforços da área de Processamento de Linguagem Natural (PLN) para lidar com a questão, com foco especial na língua portuguesa do Brasil. Inicialmente, o problema é contextualizado e são apresentadas as bases teóricas para a compreensão do fenômeno das notícias falsas. Em seguida, são introduzidos os principais *corpora* existentes para o português, sobre os quais sistemas computacionais são desenvolvidos e avaliados. São então discutidos métodos e abordagens aplicados à identificação automática de notícias falsas. Finaliza-se com uma discussão sobre consequências sociais e considerações finais sobre este tópico tão relevante.



8.1.1 Conteúdo enganoso: definições

Conteúdo enganoso, desilusão, fraude e mentira são traduções aceitas para a expressão “*deception*”, comum na língua inglesa para se referir a uma informação transmitida para criar uma falsa impressão ou conclusão. As notícias falsas, também popularmente chamadas pelo termo em inglês, *fake news*, são provavelmente o tipo mais citado de conteúdo enganoso. Elas são propositalmente divulgadas para enganar, sendo prejudiciais, incluindo os casos de quando o conteúdo é desconectado de sua origem e contexto (Lazer et al., 2018; Rubin, 2014).

Além das notícias falsas, Rubin et al. (2015) incluem os trotes/embustes (*hoaxes*, em inglês) e o conteúdo humorístico (que inclui ironia, sarcasmo e sátira, projetados para serem percebidos pelo leitor/ouvinte) como tipos de conteúdo enganoso. Santos (2022) e Chen et al. (2015) também citam os casos das revisões falsas de produtos (em sites de comércio eletrônico, por exemplo, para promover ou prejudicar produtos ou empresas) e os *clickbait*s (conteúdos que chamam a atenção, visando que o usuário clique em algum link).

Outra classificação interessante refere-se aos termos *misinformation*, *disinformation* e *malinformation* (Wardle; Derakhshan, 2017):

- *Misinformation*: as informações são falsas, mas não são criadas com a intenção de causar dano ao leitor. Esta definição abrange os erros não intencionais, tais como legendas de fotos erradas, datas, estatísticas, erros de tradução e até situações em que ironias e sátiras são levadas a sério.
- *Disinformation*: as informações são falsas e são criadas deliberadamente de forma a prejudicar uma pessoa, grupo social, organização ou país. Aqui são incluídos conteúdos de áudio ou visual (vídeos) fabricados. As notícias falsas fazem parte desse grupo, normalmente traduzido como “desinformação”.
- *Malinformation*: informações baseadas na realidade, usadas para infligir danos a uma pessoa, organização ou país. Publicações de informações privadas (vazamentos), mudança deliberada de contexto, data ou tempo, assédio e discurso de ódio são exemplos desta definição.

Também se pode falar em meias verdades, quando fatos verdadeiros e falsos se misturam nas notícias em diferentes graus. Em geral, os fatos verdadeiros são utilizados de tal forma que os falsos sejam amparados, aumentando suas chances de serem percebidos como legítimos (Clem, 2017). Mais recentemente, entraram em cena as pós-verdades, quando os fatos objetivos em si são menos importantes do que o apelo às emoções que trazem e às crenças que evocam. Assim, os chamados “fatos alternativos” ganham relevância frente a “fatos reais” e sentimentos têm mais peso do que evidências (Saquete et al., 2019). O termo surgiu no âmbito político, mas logo foi popularizado em outros domínios.

Especialmente com os avanços da Inteligência Artificial generativa, outras modalidades de conteúdo enganoso ganharam força, envolvendo a produção de vídeos, imagens e áudios falsos ou manipulados digitalmente, contribuindo para a proliferação das chamadas *deepfakes*. Esse nome surgiu da junção dos termos *deep learning* e *fake*, sendo citado originalmente em 2017 por um usuário anônimo na rede social Reddit para se referir à substituição do rosto de uma pessoa por outra em um vídeo pornográfico (Rana et al.,



2022). Mais recentemente, tem-se falado em *deepfakes* textuais (Li et al., 2024) especificamente para se referir ao conteúdo escrito produzido por grandes modelos de língua (em inglês, *Large Language Models* ou LLMs).

Como se pode perceber, há todo um ecossistema em que as notícias falsas residem, sendo um ambiente em evolução, que muda conforme a sociedade, sua cultura e a tecnologia avançam. Conforme Saquete et al. (2019) discutem, esse ecossistema visa principalmente manipular a opinião pública, influenciar o comportamento das pessoas e, em última instância, trazer poder e ganhos financeiros aos envolvidos.

Esse capítulo foca especificamente nas notícias falsas em sua modalidade escrita, que têm sido o alvo de pesquisa da maior parte dos esforços atuais em PLN.

8.1.2 Notícias falsas no tempo e no espaço

Engana-se quem pensa que notícias falsas são um mal recente ou que se restringem a países com situações socioeconômicas mais frágeis.

A fabricação e a propagação de conteúdo enganoso são um fenômeno de longa data. Antigamente, as informações eram repassadas por meio de fofocas, no “boca a boca” e também nos tabloides (a chamada “imprensa marrom”), tanto para falar sobre outras pessoas de modo “inocente” como para, intencionalmente, prejudicar a reputação de outras pessoas ou empresas rivais. Até mesmo jornais e mídias televisivas podem, inadvertidamente, divulgar conteúdo enganoso, o que pode gerar severas consequências no mundo real. Por exemplo, em 1924, um documento forjado foi publicado por um jornal britânico renomado apenas quatro dias antes das eleições gerais, com o objetivo de desestabilizar o processo eleitoral¹. Outro exemplo ocorreu depois de um acidente na Inglaterra, em 1989, quando muitas pessoas morreram devido à superlotação e falta de segurança em um estádio. Os jornais da época relataram que, entre as pessoas que morreram, algumas estavam alcoolizadas e brigaram com policiais que estavam tentando resgatar os feridos. Anos depois, provou-se que essas afirmações eram falsas².

Hoje as redes sociais e os aplicativos de mensagens instantâneas permitem que tais conteúdos enganosos alcancem uma audiência nunca imaginada antes da era da web, não discriminando países. Há casos famosos de eventos afetados por notícias falsas em todo o mundo, como a eleição americana de 2016, a saída do Reino Unido da União Europeia (que ficou conhecida pelo termo “*Brexit*”) em 2020, a pandemia de COVID-19 entre 2020 e 2023, ou a recomendação de tratamento médico para a epidemia recorrente de dengue no Brasil. O Quadro 8.1 mostra um exemplo típico desse último caso, que transcreve uma fala divulgada por vídeo sobre tratamento para dengue, amplamente desmentida por diversas agências de checagem de fatos³.

Quadro 8.1: Exemplo de notícia falsa sobre tratamento para dengue.

Esprema 200 ml do limão taiti. Dá mais ou menos um copo de limão. Beba todo o copo de uma vez só. Basta apenas uma dose. A carga viral vai acabar, bem como todos os sintomas, nas próximas horas. Nos casos de dengue, em que as plaquetas começam a baixar, o que pode levar à dengue hemorrágica, é preciso tomar os 200 ml de limão para erradicar a carga viral

¹<https://www.theguardian.com/politics/1999/feb/04/uk.politicalnews6>

²<https://www.theguardian.com/football/2016/apr/26/hillsborough-disaster-deadly-mistakes-and-lies-that-lasted-decades>

³Veja, por exemplo, <https://g1.globo.com/fato-ou-fake/noticia/2024/02/14/e-fake-que-suco-de-limao-e-caldo-de-cana-curam-dengue.ghml>



primeiro e logo depois fazer o seguinte procedimento: tomar de hora em hora 50 ml de caldo de cana. Pode tomar esse volume de caldo de cana por umas seis horas consecutivas. Esses dois procedimentos vão eliminar a carga viral e aumentar muito rapidamente as plaquetas no sangue.

Atualmente, tais desafios têm ganhado proporções ainda maiores. Há contas forjadas e robôs nas redes sociais que replicam grandes volumes desse tipo de conteúdo (para fins diversos). O advento da Inteligência Artificial generativa, apesar de sua utilidade, também traz preocupações, pois há muita desinformação produzida (dadas as limitações ainda existentes nas técnicas empregadas e seu possível mal uso), trazendo ao cenário não apenas notícias falsas produzidas por humanos, mas também conteúdo enganoso gerado por máquinas (Su et al., 2024).

Soma-se a esse cenário a dificuldade humana de detectar conteúdo enganoso em geral. Pesquisas já mostraram que humanos não conseguem separar satisfatoriamente notícias verdadeiras de notícias falsas (Bond Jr.; DePaulo, 2006; George; Keane, 2006), alcançando entre 50% e 63% de sucesso, dependendo do que seja considerado enganoso (Rubin; Conroy, 2011). Algumas explicações para isso são:

- a falta de conhecimento prévio das pessoas sobre determinados assuntos, fato este relacionado a questões de formação e educação formal, muitas vezes;
- a percepção humana da veracidade da notícia pode ser afetada pelo fato de as informações veiculadas estarem alinhadas com suas crenças e opiniões pré-existentes, o que tem sido chamado de “viés da confirmação” (Thakar; Bhatt, 2024);
- algumas pessoas recebem informações de fontes socialmente próximas, o que as leva a considerar legítimas as informações compartilhadas, além do fato de os usuários raramente verificarem as informações que compartilham (Tandoc Jr. et al., 2018);
- a dificuldade, a falta de interesse ou a falta de disponibilidade do usuário para seguir dicas que são amplamente divulgadas por sites de notícias e redes sociais, como o Facebook, cujas orientações são mostradas no Quadro 8.2, conforme reproduzidas por Santos (2022).

Quadro 8.2: Diretrizes do Facebook para identificação de notícias falsas (Santos, 2022).

1. **Seja cético com as manchetes.** Notícias falsas frequentemente trazem manchetes apelativas em letras maiúsculas e com pontos de exclamação. Se alegações chocantes na manchete parecerem inacreditáveis, desconfie.
2. **Olhe atentamente para a URL.** Uma URL semelhante à de outro site ou um telefone podem ser um sinal de alerta para notícias falsas. Muitos sites de notícias falsas imitam veículos de imprensa autênticos fazendo pequenas mudanças na URL. Você pode ir até o site para verificar e comparar a URL com a de veículos de imprensa estabelecidos.
3. **Investigue a fonte.** Certifique-se de que a reportagem tenha sido escrita por uma fonte confiável e de boa reputação. Se a história for contada por uma organização não conhecida, verifique a seção “Sobre” do site para saber mais sobre ela.
4. **Fique atento a formatações incomuns.** Muitos sites de notícias falsas contêm erros ortográficos ou apresentam layouts estranhos. Redobre a atenção na leitura se perceber esses sinais.



5. **Considere as fotos.** Notícias falsas frequentemente contêm imagens ou vídeos manipulados. Algumas vezes, a foto pode ser autêntica, mas ter sido retirada do contexto. Você pode procurar a foto ou imagem para verificar de onde ela veio.
6. **Confira as datas.** Notícias falsas podem conter datas que não fazem sentido ou até mesmo datas que tenham sido alteradas.
7. **Verifique as evidências.** Verifique as fontes do autor da reportagem para confirmar que são confiáveis. Falta de evidências sobre os fatos ou menção a especialistas desconhecidos pode ser uma indicação de notícias falsas.
8. **Busque outras reportagens.** Se nenhum outro veículo na imprensa tiver publicado uma reportagem sobre o mesmo assunto, isso pode ser um indicativo de que a história é falsa. Se a história for publicada por vários veículos confiáveis na imprensa, é mais provável que seja verdadeira.
9. **A história é uma farsa ou uma brincadeira?** Algumas vezes, as notícias falsas podem ser difíceis de distinguir de um conteúdo de humor ou sátira. Verifique se a fonte é conhecida por paródias e se os detalhes da história e o tom sugerem que pode ser apenas uma brincadeira.
10. **Algumas histórias são intencionalmente falsas.** Pense de forma crítica sobre as histórias lidas e compartilhe apenas as notícias que você sabe que são verossímeis.

8.1.3 Combate ao conteúdo enganoso

Como a propagação de conteúdo enganoso alcançou um ponto crítico, a partir das evidências de que as redes sociais têm uma parcela grande de responsabilidade pela proliferação desse tipo de conteúdo, somada à dificuldade humana de detectar tais conteúdos, surgiram iniciativas de combate a essa prática.

Agências de comunicação têm dado apoio a sites de checagem manual de fatos e a companhias de grande apelo digital. No Brasil, a checagem de fatos é em grande parte feita por jornalistas, em sites e agências como o Boatos.org, Fato ou Fake⁴, E-Farsas⁵, Aos Fatos⁶ e Agência Lupa⁷, entre outros. Eles costumam seguir uma metodologia de checagem de acordo com as diretrizes da Rede Internacional de Checagem de Fatos (em inglês, *International Fact-Checking Network* ou IFCN), que promove iniciativas de verificação de fatos⁸.

Estudos sobre a produção e a disseminação de notícias falsas têm sido realizados buscando-se mapear o fenômeno e seu comportamento. Vosoughi et al. (2018) relatam que, em seus estudos, cerca de 1% das notícias falsas mais populares foram difundidas de 1.000 a 100.000 pessoas, enquanto as notícias verdadeiras raramente chegam a 1.000 pessoas, o que pode ser explicado pelo apelo emotivo e pelo grau de novidade das notícias falsas.

No âmbito das pesquisas de PLN, há esforços para estudar as características do conteúdo enganoso e como detectá-lo automaticamente. Considerando toda a história da área, as tentativas são relativamente recentes, tanto do ponto de vista teórico quanto do prático, sendo majoritariamente focadas no processamento do conteúdo escrito. Segundo Hauch et

⁴<https://g1.globo.com/fato-ou-fake>

⁵<http://www.e-farsas.com>

⁶<https://aosfatos.org>

⁷<https://lupa.uol.com.br>

⁸<https://ifcncodeofprinciples.poynter.org>



al. (2012), a automatização do processo de detecção de conteúdo enganoso é atraente por no mínimo dois motivos: (i) esses sistemas podem ser mais objetivos do que os humanos, que são propensos a erros e vieses; e (ii) julgamentos online de várias pistas a partir de vídeos ou áudios podem sobrecarregar o humano e levar a atrasos e erros.

Em PLN, o uso de aprendizado de máquina, tanto da linha mais clássica quanto da linha profunda, incluindo os LLMs, são atualmente abordagens bastante comuns para a identificação de notícias falsas. Essas abordagens visam aprender automaticamente as características textuais (simples ou complexas) que podem se correlacionar com a veracidade/falsidade da informação, usando-as para identificar a classe de notícias novas. Há também outras abordagens, por exemplo, aquelas que buscam identificar fatos e checá-los automaticamente, assim como usar técnicas baseadas em recuperação e extração de informação.

Normalmente, essas abordagens baseiam-se em dados encontrados em *corpora* que contêm notícias falsas e verdadeiras, assim como metadados, que servem de base para o desenvolvimento/treino e avaliação/teste dos métodos de detecção de notícias falsas (para mais detalhes sobre *corpora*, ver Capítulo [Conjunto de dados, dataset e corpus](#)). A seguir, os principais *corpora* existentes para a tarefa de detecção de notícias falsas para a língua portuguesa são brevemente apresentados.

8.2 Corpora

A análise de notícias falsas tem se beneficiado da disponibilidade de variados *corpora*. Para o inglês, existem vários *corpora* em diferentes domínios, como política (Gruppi et al., 2020, 2021; Vlachos; Riedel, 2014; Wang, 2017), medicina (Dai et al., 2020), opiniões sobre produtos ou hotéis (Fornaciari; Poesio, 2013; Ott et al., 2011; Salminen et al., 2022) e tópicos sociais, como aborto (Pérez-Rosas; Mihalcea, 2014). Apesar de existir uma quantidade significativamente menor de *corpora* em português, nos últimos anos esse problema tem diminuído com o surgimento de *corpora* específicos para o português do Brasil, com características distintas que influenciam diretamente a eficácia dos modelos de detecção de notícias falsas.

A Tabela 8.1 lista os principais *corpora* de notícias falsas para o português do Brasil de que se tem ciência (na data da escrita desse capítulo), organizados cronologicamente conforme seu surgimento.

Tabela 8.1: *Corpora* de notícias falsas em português do Brasil.

<i>Corpus</i>	Quantidade de notícias	Categorias envolvidas	Período do conteúdo	Tipo de material
Fake.br (Monteiro et al., 2018; Silva et al., 2020b)	7.200	falsa, verdadeira	2016 a 2018	Notícias
FACTCK.BR (Moreno; Bressan, 2019)	1.309	falsa, meia-verdade, verdadeira	2016 a 2019	Notícias
Bracis2019FakeNews – Twitter (Faustini; Covões, 2019)	8.981	falsa, verdadeira	—	Postagens do X (Twitter)



<i>Corpus</i>	Quantidade de notícias	Categorias envolvidas	Período do conteúdo	Tipo de material
Bracis2019FakeNews – WhatsApp (Faustini; Covões, 2019)	177	falsa, verdadeira	—	Mensagens do WhatsApp
FakeTweet.Br (Cordeiro; Pinheiro, 2019)	299	falsa, verdadeira	—	Postagens do X (Twitter)
FakeWhastApp.BR (Cabral et al., 2021)	5.284	falsa, verdadeira	—	Mensagens do WhatsApp
FakeRecogna (Garcia et al., 2022)	11.902	falsa, verdadeira	2018 a 2021	Notícias
FakePedia (Charles et al., 2022)	12.398	falsa, verdadeira	2013 a 2021	Notícias
FakeTrueBR (Chavarro et al., 2023)	3.582	falsa, verdadeira	—	Notícias
Fake tweets Covid-19 (Geurgas; Tessler, 2024)	14.000	falsa, verdadeira	—	Postagens do X (Twitter)
FakeRecogna 2.0 (Garcia et al., 2024b)	52.800	falsa, verdadeira	2002 a 2023	Notícias

O primeiro *corpus* para o português de que se tem conhecimento é o Fake.br (Monteiro et al., 2018; Silva et al., 2020b), com notícias verdadeiras e falsas alinhadas, ou seja, cada notícia falsa tem uma notícia verdadeira relacionada. O Quadro 8.3 ilustra um caso de notícias alinhadas do *corpus*.

Quadro 8.3: Exemplo de notícia falsa e sua contraparte verdadeira no *corpus* Fake.br.

Notícia falsa	Notícia verdadeira
Polos magnéticos da Terra podem se inverter e causar colapso mundial: A Terra ficaria inabitável. Ao menos 100 mil pessoas morreriam por ano pela alta nos níveis de radiação espacial. Se o campo magnético continuar a diminuir e os polos magnéticos se invertem, a Terra pode acabar como Marte – um local seco, árido e incapaz de preservar a vida.	Inversão dos polos magnéticos da Terra pode ocorrer mais rápido do que o previsto. Segundo afirmações, essas ocorrências são, a princípio, indistinguíveis das verdadeiras mudanças nos polos. Apesar dessas reversões não representarem qualquer ameaça à humanidade, os especialistas alertam que poderão gerar falhas nos satélites que orbitam a Terra.

Como se pode ver na Tabela 8.1, o maior *corpus* conhecido é o FakeRecogna 2.0 (Garcia et al., 2024b), com mais de 50.000 notícias.

Os *corpora* possuem características variadas. Por exemplo, alguns deles fornecem informações temporais, enquanto outros não. Informações temporais são importantes porque já se percebeu na área que os sistemas computacionais de detecção de notícias falsas podem ser fortemente ancorados nos tópicos de determinado período. Na coluna “Período do conteúdo” da Tabela 8.1, é apresentado o período abrangido pelos dados apenas quando essa informação está disponível. Além disso, a informação é baseada nos metadados encontrados no *corpus*, o que pode ser diferente da descrição do artigo principal do *corpus*.

Por exemplo, Garcia et al. (2024b) citam que as notícias do *corpus* FakeRecogna 2.0 foram coletadas entre 2019 e 2023, mas, na análise dos metadados, foram encontradas notícias publicadas em 2002.

Os dados também variam em termos de aplicação. Existem *corpora* que incluem notícias em seu sentido mais estrito (o jornalístico), como o FakeRecogna, FakeRecogna 2.0 e o Fake.br. Outros usam dados relativos a postagens em redes sociais, como o FakeTweet.Br. Além disso, há *corpora* voltados para dados de aplicativos de mensagens instantâneas, como o WhatsApp, incluindo o Bracis2019FakeNews e o FakeWhastApp.BR.

Outro ponto de variação entre os *corpora* são as classes de veracidade avaliadas. A maioria trabalha com uma classificação binária, distinguindo entre notícias falsas e verdadeiras. Porém, o *corpus* FACTCK.BR contém três classes: falsa, meia-verdade e verdadeira.

Por fim, algumas bases de dados, como a Fake.br, fornecem, além dos textos, metadados e atributos linguísticos. Esses recursos adicionais enriquecem as possibilidades de análise, fornecendo aos pesquisadores informações estruturadas que podem ser diretamente utilizadas nos modelos de classificação. Embora seja possível extrair alguns desses atributos diretamente dos textos, sua inclusão prévia no *corpus* facilita o processo de pesquisa e gera maior consistência nos resultados.

Na próxima seção, são discutidos alguns métodos de detecção automática de notícias falsas, focando-se nos trabalhos existentes para a língua portuguesa, muitos dos quais se baseiam nos *corpora* aqui apresentados.

8.3 Métodos de detecção automática

Na busca por identificar notícias falsas de forma automática, diversos estudos na literatura exploram uma ampla gama de abordagens. Muitos desses estudos podem ser categorizados de acordo com a taxonomia do comportamento enganoso proposta por Zhou (2005), reproduzida na Figura 8.1. Essa taxonomia ajuda a organizar o comportamento enganoso, auxiliando (i) na sistematização de sinais e atributos que podem indicar se uma informação é falsa e (ii) na criação de métodos computacionais.

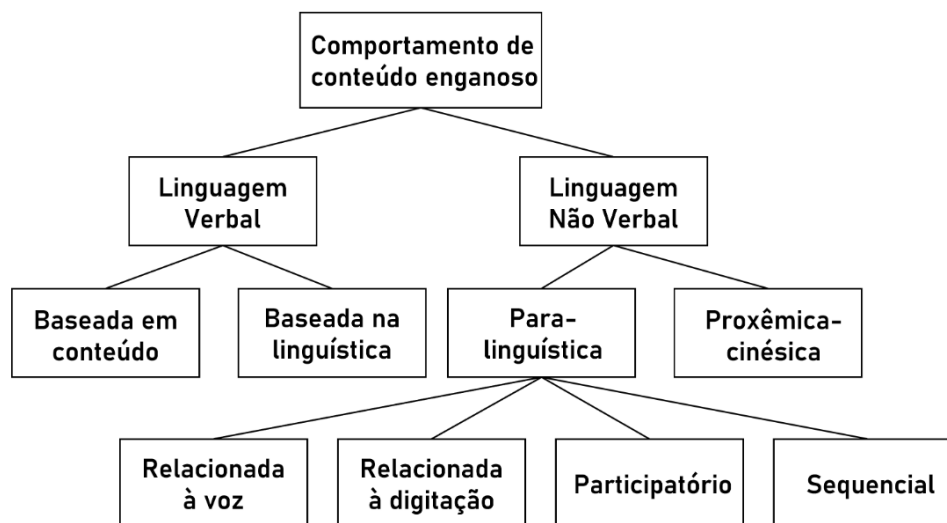
A taxonomia se divide em Linguagem Verbal e Não Verbal. Os indicadores não verbais focam nas características exibidas enquanto uma pessoa produz conteúdo enganoso, enquanto que os verbais estão ligados ao conteúdo escrito e falado da língua.

A paralinguística, um dos subgrupos da linguagem não verbal, foca em propriedades do discurso que não estão diretamente relacionadas ao conteúdo falado. Isso inclui propriedades relacionadas à voz, como o tom de voz e pausas na fala, relacionados ao teclado (velocidade de digitação, quantidade de vezes que apagou a mensagem, erros de digitação por proximidade de teclas), participatório (demora entre as respostas, mudança de assunto) e sequencial (de quem veio a iniciação de conversas). O segundo subgrupo da linguagem não verbal se refere ao comportamento proxêmico-cinésico do enganador. Ele envolve movimentos e posturas corporais, como gestos com as mãos e braços, expressões faciais e movimentos oculares, entre outros.

Apesar da linguagem não verbal ser importante, as pesquisas de PLN costumam atuar mais diretamente no grupo de Linguagem Verbal. Esse grupo é dividido em comportamento baseado em **conteúdo** e na **linguística**. Nas abordagens baseadas em linguística, são usadas pistas provenientes do texto para fazer a diferenciação entre conteúdo falso ou verdadeiro. Assume-se que os produtores de conteúdo enganoso podem inconscientemente deixar rastros linguísticos que sinalizam a falsidade. Por outro lado, a abordagem baseada



Figura 8.1: Taxonomia do comportamento enganoso.



Fonte: Adaptada de (Zhou, 2005)

em conteúdo busca simular a metodologia de checagem de fatos, avaliando cada informação apresentada, muitas vezes remetendo a abordagens baseadas em grafos de conhecimento.

A seguir, descrevem-se alguns métodos de detecção de notícias falsas, principalmente os aplicados para o português e que se alinham ao grupo da Linguagem Verbal. Para fins didáticos, os métodos são organizados em abordagens computacionais mais claramente discerníveis (na época da escrita deste capítulo), sendo que se busca apresentar pelo menos um exemplo típico pertencente a cada abordagem.

8.3.1 Aprendizado de máquina com atributos linguísticos

De acordo com Conroy et al. (2015), a maioria dos enganadores usa a linguagem estrategicamente para evitar que sejam descobertos. Nesse caso, o objetivo da abordagem linguística é procurar por exemplos de vazamentos – ou os chamados “sinais de engano preditivo” – presentes no conteúdo de uma mensagem. Isso demanda uma análise mais profunda do texto na busca de padrões que possam indicar se uma informação é enganosa ou não.

Para identificar esses padrões, a análise pode utilizar atributos linguísticos que possam diferenciar o conteúdo verdadeiro do falso, extraídos de elementos como classes gramaticais, semântica, erros ortográficos, expressividade e diversidade de conteúdo para representar as amostras textuais, a frequência de uso de certas palavras e o padrão de capitalização do texto.

Monteiro et al. (2018) e Silva et al. (2020b) apresentaram a primeira abordagem linguística para a detecção de notícias falsas em português, seguindo o que talvez possa ser definida como uma das abordagens mais clássicas para o problema. Usando o *corpus* Fake.br, o processo de detecção de notícias falsas foi modelado como uma tarefa de aprendizado de máquina, em que as notícias são as instâncias, a classificação é binária (notícia falsa ou verdadeira) e os atributos das notícias utilizados para a classificação incluem informações sobre os *tokens* presentes nas notícias e atributos linguísticos derivados.

Para representar os *tokens* de cada notícia, foram utilizadas a representação de *bag of words* (com ponderação via *Term Frequency-Inverse Document Frequency* ou TF-IDF) e a média das representações vetoriais dos *tokens* (com 300 dimensões) produzidas pelos métodos Word2Vec (versão skip-gram) (Mikolov et al., 2013b) e FastText (Bojanowski et al., 2017) treinados para o português (Hartmann et al., 2017) (Capítulo [Representação vetorial e semântica distribucional](#)). Em relação aos atributos linguísticos, foram utilizados (i) atributos sugeridos por Zhou et al. (2004), que são pausalidade, emotividade, incerteza e não imediatismo, e (ii) atributos mais gerais como diversidade, tamanho médio das sentenças, tamanho médio das palavras e número de erros ortográficos na notícia. A pausalidade indica a ocorrência de pausas no texto, que é computada como o número de sinais de pontuação sobre o número de sentenças. A emotividade mede a expressividade da linguagem, calculada como a soma do número de adjetivos e advérbios dividida pela soma do número de substantivos e verbos. A incerteza é baseada nas ocorrências de verbos modais e no uso da voz passiva. O recurso de não imediatismo é baseado na frequência de uso dos pronomes de 1ª e 2ª pessoa. A diversidade é calculada como o número total de *tokens* de conteúdo diferentes sobre o número total de *tokens* de conteúdo (ou seja, é uma versão mais refinada da conhecida taxa *type-token*).

Os autores avaliaram diferentes abordagens de pré-processamento dos dados e técnicas de Aprendizado de Máquina, incluindo, por exemplo, regressão logística, árvores de decisão, floresta aleatória e máquinas de vetores de suporte, além de testarem métodos de *ensemble* e *stacking*. Um ponto a destacar é que, como as notícias verdadeiras do *corpus* utilizado costumam ser maiores do que as falsas, os autores truncaram os maiores textos para equalizar os tamanhos e evitar vieses na classificação automática. De forma geral, os autores concluem que os melhores resultados foram produzidos pela técnica de *ensemble*, aproximando-se de 97% de acurácia na classificação.

Há várias outras iniciativas para o português que seguiram uma abordagem similar, com alguma variação no conjunto de atributos linguísticos utilizados. Veja, por exemplo, os trabalhos de Faustini; Covões (2019) e Okano; Ruiz (2019). Há também trabalhos que buscaram avaliar técnicas independentes de língua, incluindo o português entre seus experimentos, como os trabalhos de Abonizio et al. (2020) e Faustini; Covões (2020).

8.3.2 Modelos de linguagem

Modelos de aprendizado profundo têm sido utilizados em várias pesquisas que tratam do problema das notícias falsas. Em particular, modelos de linguagem, como o BERT e T5 (Capítulo [Modelos de linguagem](#)), podem ser empregados tanto para gerar representações vetoriais dos textos, que servem como entrada para métodos de classificação, quanto para atuar diretamente como classificadores.

Nesse contexto, um estudo que ilustra bem essa abordagem é o de Vieira et al. (2025), que avaliou vários modelos de geração de representação vetorial, comparando modelos tradicionais de classificação que incluíram floresta aleatória, máquinas de vetores de suporte e rede neural perceptron multicamadas. Para a geração das representações vetoriais, eles aplicaram os seguintes modelos especializados para o português: BERTimbau, Teeny-TinyLlama e Tucano. Também aplicaram modelos voltados para o inglês, como BERT, RoBERTa e BART. Por fim, aplicaram modelos multilíngues: mBERT, XLM-RoBERTa e mBART. Todos os métodos com os diferentes modelos de representação foram testados na base de dados Fake.Br. De acordo com os autores, a melhor combinação em termos de desempenho foi o mBART com o método de máquinas de vetores de suporte, atingindo



uma acurácia de 97,43%. Entretanto, ao considerar o tempo de processamento, a combinação BERTimbau com o mesmo método apresentou uma acurácia de 97,22% e um tempo de processamento significativamente menor (aproximadamente 86% inferior), oferecendo assim um bom equilíbrio entre precisão e eficiência.

Na mesma linha, Garcia et al. (2024b) avaliaram o desempenho de modelos tradicionais e de aprendizado profundo sobre o *corpus* FakeRecognia 2.0. Entre os modelos tradicionais, testaram regressão logística, perceptron multicamadas, *naive* Bayes, *optimum-path forest* (Papa et al., 2009), floresta aleatória e máquinas de vetores de suporte. No que diz respeito aos modelos de aprendizado profundo, testaram LSTM, GRU, CNN e BERT. Para gerar a representação vetorial dos textos nos experimentos com métodos tradicionais, utilizaram a frequência dos termos, a representação TF-IDF e o FastText. Nos experimentos com modelos de aprendizado profundo, utilizaram o FastText para LSTM, GRU e CNN, o BERTimbau (Souza et al., 2020a) como a versão do BERT para o português, e o modelo PTT-5 (Portuguese, Tagalog, Turkish, Tamil e Telugu) (Carmo et al., 2020b) para o T5.

É interessante destacar que a abordagem proposta pelos autores se diferencia de outras metodologias ao implementar uma etapa de sumarização das notícias verdadeiras antes de treinar os modelos de classificação (em vez de truncar os maiores textos, como comentado anteriormente neste capítulo). Dessa forma, a diferença de tamanho entre notícias verdadeiras e falsas é resolvida. Os autores adotaram tanto a sumarização extrativa quanto a abstrativa. Garcia et al. (2024b) afirmam que a sumarização extrativa é vantajosa, pois tende a ser imune a inconsistências e “alucinações”, já que o resumo final é composto pelas sentenças mais relevantes sem a geração de novas palavras ou expressões. Por outro lado, a sumarização abstrativa permite a produção de sentenças inéditas que podem diferir do texto original em termos de semântica e estrutura das sentenças, mas possui como vantagem ser mais semelhante à escrita humana.

Os melhores resultados obtidos pelos autores foram alcançados com o BERT, que apresentou acurácia aproximada de 98%. O modelo T5 apresentou resultados ligeiramente inferiores.

Além dos métodos de aprendizado profundo aplicados por Garcia et al. (2024b), outros estudos têm utilizado abordagens alternativas, como redes neurais de atenção hierárquica (Okano et al., 2020) e redes neurais de grafos (De Souza et al., 2024). No primeiro trabalho, a rede neural de atenção hierárquica alcançou uma acurácia de 90% (com textos truncados). No segundo trabalho, os autores propuseram um método que incorpora um mecanismo de atenção na abordagem de aprendizado não rotulado por propagação de rótulos utilizando uma rede neural baseada em grafo. O método foi avaliado em seis *corpora* de notícias falsas, representando diferentes cenários quanto à linguagem, fonte, tópicos e balanceamento. Nos testes realizados com o *corpus* Fake.br, o método alcançou acurácia de cerca de 78%.

Outros métodos já utilizados para a detecção de notícias falsas que empregaram textos em português do Brasil incluem variantes e aprimoramentos do BERT, como o modelo de linguagem DistilBERT. Um exemplo é o trabalho de Gôlo et al. (2023), que apresenta o MVAE-FakeNews (codificadores variacionais multimodais para detecção de notícias falsas). Este método multimodal representa os textos na detecção de notícias falsas por meio de aprendizado baseado em uma única classe, aprendendo uma nova representação a partir da combinação de modalidades promissoras para dados de notícias, combinando informações de linguagem, características linguísticas e tópicos, sendo que a modalidade de linguagem é derivada das informações do modelo DistilBERT. Neste trabalho, os autores utilizam o algoritmo de máquinas de vetores de suporte com uma classe para fazer a classificação. Os resultados indicam que o MVAE-FakeNews alcançou valores de acurácia melhores do que



diversos outros métodos, demonstrando que abordagens multimodais podem ser melhores do que abordagens unimodais que concatenam os dados de entrada. Nos experimentos com o *corpus* Fake.br, o melhor valor de acurácia obtido foi de 65% (utilizando 10% de notícias falsas rotuladas), com os métodos baseados em DistilBERT apresentando os melhores desempenhos.

O trabalho de Pires; Silva (2024) também avaliou variantes do modelo BERT, mais especificamente o BERTimbau e o mBERT, comparando seus resultados com o BERT. Eles testaram esses métodos nas seguintes bases de dados: Fake.Br, FakeTrueBR, FakeRecogna e Fakepedia. De acordo com os autores, os resultados obtidos indicam uma superioridade do BERTimbau sobre o BERT e sobre o mBERT na referida tarefa, com uma melhora média da medida F1 de 2,37% e 1,07%, respectivamente.

Outro trabalho que avaliou o mBERT é Fischer et al. (2022) comparando-o com outros modelos de classificação, como regressão logística, árvores de decisão, floresta aleatória, k-vizinhos mais próximos, SVM, naive Bayes, XGBoost, CNN, GRU e LSTM. O mBERT obteve uma medida F1 de 0,98 superando os outros métodos avaliados.

Em paralelo a essas abordagens, a aplicação de LLMs (Capítulo [Modelos de linguagem](#)) na detecção de notícias falsas tem ganhado atenção, especialmente em trabalhos em inglês, como os de Lucas et al. (2023) e Su et al. (2024), que discutem não apenas a detecção de notícias falsas com LLMs, mas também a geração desse conteúdo.

O uso de LLMs para detectar notícias falsas não é trivial, pois o treinamento desses modelos é custoso, resultando em atualizações pouco frequentes, o que muitas vezes os impede de fornecer informações atualizadas sobre determinados assuntos. No entanto, uma solução para isso é a técnica de *Retrieval-Augmented Generation* (RAG), que permite fornecer informações atualizadas aos LLMs. Essa técnica utiliza conceitos de recuperação de informação para buscar, em uma base de dados complementar, os conteúdos mais relevantes para uma pergunta (especificada na forma de um *prompt*) e, em seguida, passar essas informações ao LLM como contexto. Assim, o LLM gera respostas que, em vez de se basearem apenas em seu conhecimento prévio, são fundamentadas nos dados adicionais fornecidos.

Um exemplo relevante no contexto de notícias falsas é apresentado por Yue et al. (2024), que propõe o framework RARG (*Retrieval Augmented Response Generation*) para combater a desinformação online, especialmente no contexto da COVID-19. O RARG gera respostas baseadas em evidências científicas para refutar informações falsas, funcionando em duas etapas: (1) recuperação de evidências, onde documentos relevantes são coletados e reclassificados a partir de uma base com mais de um milhão de artigos acadêmicos; e (2) geração de respostas, em que um LLM é treinado para produzir respostas polidas e factuais com base nas evidências obtidas. Além disso, o modelo é otimizado com aprendizado por reforço com feedback humano (Capítulo [Modelos de linguagem](#)), garantindo o uso eficaz das evidências sem comprometer a qualidade das respostas.

Yue et al. (2024) fizeram experimentos com o RARG usando dados tanto de dentro quanto de fora do domínio da COVID-19 para a língua inglesa. Eles realizaram uma avaliação comparativa, de formas qualitativa e quantitativa, para analisar as respostas geradas por diferentes modelos. As métricas quantitativas utilizadas incluíram critérios de refutação, factualidade, polidez, e relevância da resposta em relação à alegação de entrada e à evidência. Foram avaliados modelos baseados em geração de texto, como BART, DialoGPT e GODEL, além de métodos de aprendizado por reforço, como PARTNER e MisinfoCorrect, e LLMs como Llama 2 e GPT-3.5. As análises revelaram que o modelo RARG obteve o melhor desempenho geral.



Para o português do Brasil, já existem algumas contribuições relevantes que indicam o potencial crescente dessa linha de pesquisa. Por exemplo, o estudo de Gôlo et al. (2024) explora modelos de código aberto, pouco aplicados ao contexto de notícias falsas. Os pesquisadores criaram um novo conjunto de dados de notícias políticas brasileiras e realizaram uma avaliação empírica de LLMs de código aberto e da OpenAI. O Gemma 2 do Google superou outros seis LLMs (Llama 3, Phi 3, Qwen 1, Qwen 2, Gpt 3.5 turbo e GPT 4), mostrando-se o mais promissor para essa tarefa, atingindo valores de F1 de 0,90. Além do desempenho de classificação, o estudo avaliou a capacidade dos LLMs de gerar explicações alinhadas ao raciocínio humano (com destaque para os modelos de melhor desempenho) e constatou que os modelos Gemma e Qwen produziram os melhores embeddings, separando com clareza notícias falsas e verdadeiras em visualizações via t-SNE.

Em outro estudo relacionado, Silva et al. (2025) avaliam o impacto de notícias falsas geradas por LLMs no desempenho de métodos de aprendizado de máquina na classificação. A pesquisa utilizou diversos modelos para classificar notícias falsas, incluindo algoritmos de aprendizado de máquina tradicionais (árvore de decisão, regressão logística, floresta aleatória, máquinas de vetores de suporte e gradiente descendente estocástico), um modelo BERTimbau ajustado, um LLM puro e dois LLMs combinados com RAG. Os experimentos foram realizados usando os corpora Fake.Br e FakeTrueBR pelo fato deles serem alinhados, isto é, cada notícia falsa possui sua contraparte verdadeira. Os pesquisadores avaliaram o desempenho desses modelos com notícias falsas sintéticas, criadas com o modelo de LLM Sabiá-3. Os resultados revelaram uma degradação significativa no desempenho dos modelos de aprendizado de máquina quando avaliados em condições de incompatibilidade, ou seja, quando treinados com conteúdo gerado por humanos e testados com notícias falsas geradas por LLMs, ou vice-versa. Por exemplo, a medida F1 do BERTimbau foi de 0,99 no experimento onde o conjunto de treino e teste eram compostos por notícias escritas por humanos, mas caiu para 0,69 quando o conjunto de treino era composto por notícias geradas por humanos e o conjunto de teste era formado por notícias geradas por LLMs. Observou-se ainda que classificadores baseados em LLM foram mais robustos sob condições de incompatibilidade, enquanto que o BERTimbau e os modelos tradicionais de aprendizado de máquina apresentaram melhor desempenho quando não havia essa diferença. Os autores afirmam que os resultados obtidos ressaltam a necessidade de abordagens mais resilientes e adaptativas diante da crescente sofisticação do conteúdo gerado por LLMs.

8.3.3 Checagem de conteúdo

Para abordagens de checagem automática de conteúdo, em que os fatos em si devem ser verificados (o que vai além de avaliar estilo de escrita e atributos linguísticos), Thorne; Vlachos (2018) sugerem que se extraiam triplas de informação dos documentos verdadeiros (por algum processo de extração de informação – ver Capítulo 3), normalmente formadas por <sujeito, verbo, objeto>, que constituem os fatos básicos mencionados. Nem sempre é trivial fazer isso, pois pode haver sentenças com número variável de argumentos (ausência de objeto ou mais de um objeto, por exemplo) ou com estruturas sintáticas mais complexas ou não usuais. Com base nessas triplas, pode-se construir um grafo de conhecimento, em que sujeitos e objetos dão origem a vértices do grafo e os verbos definem arestas entre eles. O grafo é, dessa forma, uma representação estruturada do conteúdo verídico, podendo ser utilizado para avaliação da veracidade de novas informações, o que pode ser feito de diferentes formas.

Santos; Pardo (2020) adaptam para o português a abordagem de Ciampaglia et al.



(2015). Nessa abordagem, o grafo de conhecimento é constituído de triplas extraídas da Wikipédia, considerada como uma fonte de informações verídicas sobre o mundo. Apresentada uma nova informação a ser checada, as triplas dessa informação são extraídas e analisadas com o apoio do grafo de conhecimento: caminhos distintos entre sujeitos e objetos, com diferentes distâncias, fornecem suportes factuais diferentes para as afirmações. São consideradas verdadeiras as afirmações cujos sujeitos e objetos apresentam conexões diretas entre si ou caminhos de distância relativamente curta. A proximidade entre os conceitos envolvidos no grafo reflete, portanto, a veracidade da informação.

Os autores testam a abordagem com afirmações históricas e geográficas, em especial, capitais de estados e países e cônjuges de presidentes brasileiros. Em média, os resultados atingem valores de acurácia de 74%. Essa abordagem, apesar de exigir análise linguística mais refinada para produção do grafo de conhecimento e de produzir resultados inferiores em relação às outras abordagens, é interessante por permitir rastrear a veracidade de cada fato e explicar os resultados. Como amplamente discutido atualmente na Inteligência Artificial, a explicabilidade dos métodos é uma característica muito relevante (ver, por exemplo, o Capítulo [Questões éticas em IA e PLN](#)).

Há várias outras estratégias de uso de grafos que não necessariamente buscam os fatos explicitamente. Santos; Pardo (2021), por exemplo, testam outras formas de construção e percurso em grafos (tanto para notícias verdadeiras como falsas) e o uso de medidas de redes complexas para caracterização estrutural de notícias falsas e verdadeiras, que se mostram promissoras (e às vezes de mais simples aplicação), mas com resultados inferiores aos obtidos anteriormente.

8.3.4 Recuperação de informação

Santos; Pardo (2020) exploram uma estratégia baseada no comportamento humano de checagem de informação: usa-se um sistema de recuperação de informação (no caso, o Google – ver Capítulo 2) para buscar pelo título da notícia de interesse, analisando-se os resultados retornados na busca. Caso haja indício de falsidade da notícia correspondente, ela é classificada dessa forma.

Os autores analisam apenas os links e os *snippets*⁹ retornados na 1ª página de resultados. Se houver link de algum site de verificação de notícias (como os citados no início desse capítulo, por exemplo, Boatos.org, Aos Fatos e Agência Lupa) ou houver algum termo indicativo de falsidade nos *snippets*, a notícia é classificada como falsa. Os termos indicativos considerados pelos autores são “fake”, “falso”, “mentira”, “calúnia”, “inverídico”, “enganoso”, “farsa”, “ilusório”, “ilegítimo”, “boato” e “rumor”, além de suas variações possíveis de gênero e número.

Os autores avaliam essa estratégia com o *corpus* Fake.br, indicando resultados de classificação com acurácia de 68%. Apesar do resultado limitado, a simplicidade desse método torna-o interessante, pois faz uso de serviços de web já disponíveis. Além disso, essa estratégia evidencia a necessidade de modelagens mais sofisticadas para o devido tratamento do fenômeno das notícias falsas.

8.3.5 Análise de credibilidade

Uma outra linha de pesquisa relacionada à detecção de notícias falsas é a análise de credibilidade. Newell et al. (2017) afirmam que a credibilidade das atribuições feitas

⁹Os *snippets* são os trechos de texto retornados junto a cada link na página de resposta do buscador web.

na notícia é fundamental para garantir a confiança nos meios de comunicação. Nesse sentido, Burriess (1988) comenta que o jornalismo deve lidar com fatos verificáveis, sendo princípio básico a seguir, pois o público que recebe as notícias depende (i) da reputação da organização noticiosa, (ii) da reputação do repórter e (iii) da informação dentro da própria história, a fim de determinar a exatidão de uma notícia. Por isso, é necessário buscar padrões linguísticos no texto e metadados que possam medir a credibilidade de uma notícia.

Um exemplo de trabalho que segue essa linha é de Couto et al. (2024), que analisou padrões de rede de fontes de notícias brasileiras de alta e baixa credibilidade, focando em redes de desinformação. Esses padrões emergem de diversas fontes, como registros de domínios, certificados TLS, infraestrutura de hospedagem e geolocalização. Para conduzir essa análise, os autores criaram um conjunto de sites de baixa credibilidade, identificando aqueles que publicaram ao menos uma notícia cuja veracidade foi contestada por uma agência de checagem de fatos signatária da *International Fact-Checking Network* (IFCN). A partir dessa base, 31 atributos foram computados.

Segundo os autores, os resultados revelaram que apenas 12,2% dos sites de baixa credibilidade estão registrados e hospedados no Brasil, em contraste com 47% dos sites de alta credibilidade. Além disso, sites de alta credibilidade tendem a ter uma vida útil mais longa, domínios registrados por períodos maiores e certificados TLS (em inglês, *Transport Layer Security*) com validade estendida. Foi também identificado que sites de baixa credibilidade são mais propensos a utilizar TLDs (em inglês, *Top Level Domain*) alternativos. Os autores visualizaram a separação entre sites de alta e baixa credibilidade em duas dimensões utilizando a técnica t-SNE, e treinaram um modelo de floresta aleatória que previu a credibilidade dos sites com bom desempenho, atingindo uma acurácia de aproximadamente 83%.

Além de analisar as fontes de notícias, também é possível analisar a credibilidade de usuários, usando abordagens que buscam identificar sinais de qualidade do conteúdo (Souza Freire et al., 2021) e os disseminadores de conteúdo enganoso, sejam robôs ou usuários comuns (Chakraborty et al., 2016; Speicher et al., 2018; Varol et al., 2017; Yang et al., 2020a). Esse tipo de abordagem se baseia em informações sobre os usuários para prever seu comportamento e as chances de que as notícias que compartilham sejam falsas, incluindo análises demográficas, como idade e localização. Essa linha de investigação é bastante usada em postagens de redes sociais, onde pode ser analisado o número de postagens, o perfil do usuário, sua rede de contatos e os dados das postagens em si. Esse nível de meta-análise pode ser combinado com os demais tipos de métodos para a identificação de conteúdo enganoso.

8.3.6 Outras iniciativas relacionadas

Sendo um campo de pesquisa vasto, como pode ser percebido pelos diversos tipos de conteúdo enganoso e abordagens computacionais existentes para detecção de notícias falsas, há algumas outras iniciativas que são importantes de serem citadas.

Há trabalhos que focam suas análises em atributos de inteligibilidade das notícias (Capítulo 6), buscando identificar padrões de escrita que possam sinalizar conteúdo falso. Santos et al. (2020b), por exemplo, avaliam dezenas de medidas de inteligibilidade textual, como índices de legibilidade, coesão, diversidade lexical, complexidade sintática, organização discursiva e informações psicolinguísticas, entre outros. Em experimentos com o *corpus* Fake.br, os autores verificaram que essas medidas contribuem para a discriminação entre



notícias falsas e verdadeiras.

Alguns trabalhos analisam o impacto da análise de sentimentos (Capítulo 16) para a determinação da veracidade do conteúdo. Alonso et al. (2021) apresentam uma discussão ampla sobre o assunto. Em um caso aplicado ao português, Jeronimo et al. (2020) evidenciam que a classificação de subjetividade pode auxiliar na tarefa de detecção de notícias falsas, argumentando que notícias verídicas que visam compartilhar informações factuais e imparciais tendem a usar uma linguagem mais objetiva que depende menos de pressupostos ou expressão sentimental.

Em uma vertente que utiliza análise linguística mais profunda, há trabalhos que buscam considerar padrões de organização do conteúdo textual que possam sinalizar a veracidade desse conteúdo, fazendo uso, por exemplo, de modelos discursivos, como a famosa *Rhetorical Structure Theory* (RST) (Mann; Thompson, 1988) (Capítulo Modelos discursivos). Buscam-se padrões de relações discursivas utilizadas e de distribuição de elementos mais e menos importantes (os “núcleos” e “satélites”, que são os termos utilizados pelo modelo RST) para caracterizar o conteúdo enganoso. Vargas et al. (2022c) apresentam uma revisão desses trabalhos.

Alguns trabalhos combinam diversas das linhas apresentadas neste capítulo. Por exemplo, o trabalho de Vargas et al. (2024) envolve diversas etapas para realizar raciocínio factual em nível sentencial, incluindo a identificação de conteúdo relevante para analisar, a busca de evidências que suportem as informações e a geração de explicação para os resultados obtidos, combinando métodos e recursos variados de PLN.

Por fim, para o leitor mais interessado, Vargas et al. (2025), Thakar; Bhatt (2024) e Farhangian et al. (2024) apresentam olhares complementares aos tópicos abordados neste capítulo, incluindo visão geral de métodos, comparações e discussões de desafios e tendências na tarefa de detecção automática de notícias falsas, podendo ser referências adicionais importantes.

8.4 Impactos para a sociedade

São muitos os impactos que a incidência de conteúdo enganoso traz para as sociedades em suas diversas camadas, em todos os domínios da vida. O Brasil não é exceção, infelizmente.

Algumas áreas, como a saúde, são extremamente sensíveis a isso. Um exemplo pode ser visto no estudo empírico conduzido por Galhardi et al. (2020) sobre a pandemia de COVID-19. Os autores analisaram os relatos de usuários enviados ao aplicativo brasileiro “Eu Fiscalizo” durante o período de março a abril de 2020. Foi observado que 65% dos relatos propagavam orientações caseiras e incorretas para evitar a disseminação do vírus. Eles também notaram que 20% continham métodos para curar a doença. Esse problema não foi exclusividade no Brasil. Por isso, o termo “infodemia” passou a ser usado em vários veículos de comunicação e até mesmo pelo Organização Mundial de Saúde, significando que o mundo estava enfrentando, além de uma pandemia do coronavírus, uma pandemia de desinformação (Salvi et al., 2021; Zarocostas, 2020).

A desinformação também pode causar prejuízos econômicos. Um caso foi a disseminação de uma informação falsa em 2020 sobre um investimento significativo em uma empresa brasileira. Essa notícia enganosa levou muitos investidores a aplicarem recursos na empresa, resultando em grandes prejuízos quando a fraude foi descoberta, o que provocou uma queda acentuada nos preços das suas ações¹⁰.

¹⁰<https://www1.folha.uol.com.br/mercado/2021/11/cvm-abre-processo-contr-ex-executivos-do-irb-por->



As notícias falsas podem gerar problemas ainda mais graves, chegando até a resultar em mortes. Existem vários casos que mostram esse perigo. Um caso impactante ocorreu em 2014, quando uma pessoa foi espancada até a morte após supostamente ser alvo de uma notícia falsa que a acusava de sequestrar crianças para rituais de magia negra¹¹. Em um caso mais recente, em 2024, outra pessoa foi morta após a disseminação de uma notícia supostamente falsa que afirmava que ela matava cães¹².

Notícias falsas também podem aumentar a polarização ao interferir no processo de troca de opiniões e evitar a formação de consenso dentro de uma sociedade. Isso pode ser evidenciado pelo estudo promovido por Azzimonti; Fernandes (2023), que simularam uma rede social sintética, baseada no Twitter (renomeada para X), para analisar a disseminação de desinformação. Os autores observaram que robôs, programados para promover visões tendenciosas, podem influenciar uma pequena parte dos usuários ao disseminarem notícias falsas. Essas notícias são propagadas por esses usuários, contaminando outros que confiam em suas opiniões. Isso dificulta a difusão de informações corretas e amplia a divisão e polarização no debate público.

Conclusões similares às obtidas no trabalho acima foram apresentadas por Oliveira (2023) no contexto brasileiro. A autora analisou a disseminação de informações falsas a partir da compreensão de cultura como uma teia de significados, que se constrói continuamente na interação discursiva das pessoas. O estudo explora as dinâmicas discursivas e comunicacionais durante as eleições brasileiras de 2022, apontando a relação entre a circulação de desinformação e a polarização crescente no Brasil. Segundo a autora, informações falsas produzem “verdades socialmente compartilhadas”, que nem sempre se confirmam na realidade factual, mas que moldam o comportamento das pessoas, gerando intolerância e dissonância cognitiva, fatores que contribuem para a polarização social. O estudo reforça que a disseminação de desinformação nas redes sociais facilita a criação de realidades paralelas, exacerbando os conflitos e fomentando divisões sociais. Oliveira (2023) argumenta que essas “verdades compartilhadas”, baseadas em conteúdos falsos, interferem na capacidade de diálogo e na coesão social, contribuindo para o adoecimento psicológico e o afastamento entre grupos, refletindo diretamente na polarização da sociedade.

Os diversos impactos sociais discutidos nesta seção, que vão desde a saúde pública até a qualidade dos debates públicos, as repercussões econômicas e a violência, evidenciam a relevância do desenvolvimento de técnicas para a detecção automática de notícias falsas. O contínuo avanço da Inteligência Artificial, juntamente com o surgimento de métodos que facilitam a criação de conteúdos enganosos cada vez mais sofisticados, aumentam a necessidade de estudos que busquem aprimorar os mecanismos automáticos capazes de identificar e mitigar esses problemas. A implementação eficaz dessas técnicas é essencial para auxiliar na proteção da sociedade dos efeitos prejudiciais da desinformação.

Visando transferir conhecimento e tecnologia para a sociedade, alguns estudos científicos resultam na produção de softwares disponibilizados para o público em geral. A pesquisa relatada por Monteiro et al. (2018), por exemplo, resultou no sistema FakeCheck¹³ (treinado com o *corpus* Fake.br), disponível online e podendo ser utilizado em qualquer navegador da web. Sua interface pode ser vista na Figura 8.2.

manipulacao-de-mercado.shtml

¹¹<https://g1.globo.com/sp/santos-regiao/noticia/2014/05/mulher-espancada-apos-boatos-em-rede-social-morre-em-guaruja-sp.html>

¹²<https://g1.globo.com/sp/mogi-das-cruzes-suzano/noticia/2024/02/24/Como-fake-news-sobre-assassinato-de-cachorros-causou-a-morte-de-um-jovem-e-a-prisao-de-7-pessoas-veja-o-que-se-sabe.gh.html>

¹³<https://nilc-fakenews.herokuapp.com/>



Figura 8.2: FakeCheck – um sistema para detecção automática de notícias falsas.

FakeCheck

Detector de Fake News

Como funciona?

Copie o texto de uma notícia, cole na caixa abaixo e clique em "Enviar". O sistema irá processar o texto para identificar características de escrita, como palavras usadas ou classes gramaticais mais frequentes, e utilizar essas características em um modelo de aprendizado de máquina que classificará a notícia em verdadeira ou falsa. Para mais informações sobre como o sistema funciona e sua taxa de acerto, clique [aqui](#). Você também pode utilizar o nosso [bot do WhatsApp](#).

ATENÇÃO: Utilize o texto completo da notícia! O texto deve ter pelo menos 100 palavras. O sistema pode não funcionar corretamente com apenas partes de notícias.

Esprema 200 ml do limão taiti. Dá mais ou menos um copo de limão. Beba todo o copo de uma vez só. Basta apenas uma dose. A carga viral vai acabar, bem como todos os sintomas, nas próximas horas. Nos casos de dengue, em que as plaquetas começam a baixar, o que pode levar à dengue hemorrágica, é preciso tomar os 200 ml de limão para erradicar a carga viral primeiro e logo depois fazer o seguinte procedimento: tomar de hora em hora 50 ml de caldo de cana. Pode tomar esse volume de caldo de cana por umas seis horas consecutivas. Esses dois procedimentos vão eliminar a carga viral e aumentar muito rapidamente as plaquetas no sangue.

Notícia

Modelo de Detecção

Palavras do Texto

ENVIAR

Resultado:

Essa notícia pode ser falsa. 😞 Busque fontes confiáveis.

Fonte: (Monteiro et al., 2018)

8.5 Considerações finais

Este capítulo abordou os principais aspectos relacionados às notícias falsas, incluindo os diferentes tipos de desinformação, os métodos mais utilizados na detecção, algumas bases de dados relevantes na língua portuguesa e os impactos significativos das notícias falsas na sociedade. Embora essa seja uma área de estudo altamente relevante, há ainda muitos desafios e aspectos que não foram explorados.

O capítulo se concentrou principalmente no problema sob a perspectiva do PLN, mas é importante reconhecer que o problema pode ser multimodal, ou seja, pode envolver não apenas texto, mas também imagens, vídeos e áudio. Dessa forma, a detecção eficaz de alguns tipos de notícias requer a integração de áreas como visão computacional, reconhecimento de voz e análise de redes sociais. A colaboração multidisciplinar é essencial para enfrentar a complexidade crescente do problema.

Outro desafio refere-se à disponibilidade de dados em quantidade e com a qualidade necessárias, dada a dificuldade de coleta do tipo de dado desejado, o que pode comprometer a modelagem linguístico-computacional dos fenômenos de interesse e o treinamento eficaz de sistemas. Nesse momento, é relevante explorar em mais detalhes alguns pontos relacionados à construção dos *corpora*, como a existência de possíveis vieses de classificação (Vargas et al., 2025), o aspecto temporal e os fenômenos que se pretende abordar.

A cobertura temporal das notícias dos *corpora* é uma questão de grande importância, sendo que se tem observado na área que muitos sistemas acabam não conseguindo lidar bem com tópicos novos que não existiam ou que não estavam em voga quando os dados de treinamento foram coletados. Por exemplo, no *corpus* Fake.br, não há notícias sobre a COVID-19 (uma vez que as notícias desse *corpus* foram coletadas antes da pandemia relacionada), e, como consequência, os sistemas com ele treinados podem não ser capazes de identificar com boa acurácia as notícias falsas e verdadeiras sobre o assunto. O tempo também pode afetar os trabalhos da área de maneiras inesperadas. Usando o mesmo *corpus* como exemplo, há uma notícia sobre a prisão de um político que era falsa na época em que a notícia foi veiculada; entretanto, depois de algum tempo, o político citado de fato foi preso. Sistemas que não modelem apropriadamente a questão temporal podem classificar incorretamente notícias assim. Sistemas baseados em LLMs com RAG, por exemplo, podem facilmente incidir em algum erro assim se restrições temporais não forem especificadas.

A importância do fator temporal foi enfatizada em um estudo recente de Wanderley et al. (2025), que aplicou métodos de detecção de desvio de conceito no *corpus* FakeRecognia 2.0 e mostrou que o fenômeno de desvio de conceito está presente em notícias brasileiras, especialmente nas notícias falsas. Os autores afirmam que pesquisas que ignoram a natureza temporal das notícias falsas provavelmente apresentam métricas de desempenho infladas. Segundo eles, isso expõe uma falha metodológica crítica nas práticas de pesquisa atuais e sublinha o perigo de implementar modelos que não conseguem se adaptar à evolução das narrativas de desinformação.

Também é importante destacar a limitação de que a maioria dos *corpora* existentes abrangem notícias inteiramente falsas ou verdadeiras, mas não cobrem outros tipos de fenômenos, como meias verdades e pós-verdades. Por um lado, é mais difícil identificar e coletar notícias desses tipos; por outro, pode ser bastante desafiador desenvolver sistemas de detecção desses fenômenos, que (ainda) acabam sendo relativamente pouco privilegiados na área. Outros desafios são pontuados no estudo qualitativo dos *corpora* em português apresentado por Baracho et al. (2025), que afirmam que comparado aos *corpora* da língua



inglesa, há problemas em termos de volume desses corpora em português, de diversidade temática, de multimodalidade e, em alguns casos, de balanceamento de classes.

Em relação ao desempenho dos sistemas computacionais, um ponto que necessita de mais discussão é o impacto dos erros na detecção de notícias falsas. Muitas vezes, em pesquisas na área, busca-se uma alta acurácia, maximizando os valores de verdadeiros positivos e verdadeiros negativos (usuais em avaliações de aprendizado de máquina) e, consequentemente, melhorando a precisão e a cobertura dos sistemas (Capítulo [Avaliação de tecnologias de linguagem](#)). Os erros cometidos são, frequentemente, discutidos de forma agnóstica nas pesquisas. Entretanto, se um sistema de detecção de notícias falsas é adotado no dia a dia, os tipos de erros podem ter importâncias diferentes. O que seria menos danoso para a sociedade: classificar erroneamente uma notícia verdadeira como falsa ou uma notícia falsa como verdadeira? Levados ao extremo, o primeiro caso pode ser classificado como uma tentativa de desqualificação do produtor do conteúdo ou de censura do próprio conteúdo, enquanto o segundo caso pode legitimar mentiras e trazer consequências desastrosas (como as citadas na seção anterior). Com os avanços da Inteligência Artificial, esse tipo de discussão é cada vez mais importante na área (Capítulo [Questões éticas em IA e PLN](#)), não se devendo esquecer de considerar fatores humanos, como a resistência do público em confiar em sistemas automatizados ou a tendência a politizar o uso dessas ferramentas.

Por fim, é importante salientar que a tecnologia é apenas um dos pilares para o combate às notícias falsas. A educação e a legislação são outros pilares: na frente educacional, é necessário preparar as novas gerações para lidar com a tecnologia, com os meios digitais e com a Inteligência Artificial, para que conheçam suas potencialidades e os riscos inerentes; em termos de legislação, é preciso avançar nas leis que permitam a responsabilização dos envolvidos na produção e na disseminação de notícias falsas, sendo este um tópico muito desafiador por envolver questões que podem perpassar interesses variados (por exemplo, interesses comerciais e políticos) e por caminhar em uma separação às vezes sutil entre liberdade de expressão, censura e ações possivelmente criminosas.

8.6 Exercícios de fixação

1. O capítulo apresenta diversas definições de conteúdo enganoso. Explique o que são notícias falsas e qual sua relação com os termos *misinformation* e *disinformation*, fornecendo um exemplo para cada uma das últimas duas categorias.
2. Pesquisas mostram que os humanos têm dificuldade em separar notícias verdadeiras de falsas. Enumere e explique três razões principais que justificam essa dificuldade humana.
3. Na abordagem de aprendizado de máquina com atributos linguísticos para detecção de notícias falsas, os pesquisadores buscam por “sinais de engano preditivo” no conteúdo de uma mensagem. Explique o conceito de “emotividade” e “pausalidade” como atributos linguísticos usando um exemplo de notícia. Depois explique como um sistema baseado predominantemente em atributos linguísticos poderia erroneamente classificar notícias verdadeiras como falsas.
4. O capítulo apresenta métodos de detecção de notícias falsas que se concentram na análise do conteúdo da notícia (atributos linguísticos, checagem de fatos) e outros que analisam a credibilidade da fonte (reputação do site, padrões de rede, características



do usuário). Discuta a conexão entre essas duas abordagens. Em quais cenários a análise da fonte seria mais eficaz que a análise do conteúdo, e vice-versa? Como a combinação de ambas poderia resultar em um sistema de detecção mais robusto e confiável?

5. O capítulo descreve diferentes métodos de detecção automática de notícias falsas. Compare brevemente a abordagem de “checagem de conteúdo” com a abordagem de “recuperação de informação”. Qual delas é mais facilmente explicável em seus resultados e por quê?
6. Um desafio importante na detecção automática é a questão temporal. Explique por que a temporalidade é um desafio para os sistemas de detecção e como a técnica de RAG, em conjunto com LLMs, pode oferecer uma solução para mitigar esse problema.
7. Com suas próprias palavras, crie uma notícia falsa sobre um assunto de seu interesse e teste alguns LLMs para checar se eles são capazes de identificar as notícias como falsas. Faça ajustes na escrita da notícia falsa e também no prompt utilizado, de forma a verificar a consistência dos modelos. Que lições você consegue tirar disso? Há diferenças entre os modelos? É possível enganar os LLMs, levando-os a validar notícias falsas? É possível ter plena confiança na avaliação dos LLMs?
8. Sistemas de detecção automática podem cometer dois tipos de erro: classificar uma notícia verdadeira como falsa (falso negativo) ou uma notícia falsa como verdadeira (falso positivo). Discuta sobre as implicações sociais e éticas de cada tipo de erro, considerando o impacto na liberdade de expressão, na confiança do público em fontes de informação e na segurança. Use exemplos de situações em que cada tipo de erro pode causar impacto negativo.
9. O capítulo foca no texto, mas ressalta que o problema da desinformação é multimodal, envolvendo imagens, vídeos e áudio. Se você fosse projetar uma plataforma de detecção automática de notícias falsas abrangente, que integrasse a análise de texto, imagem e áudio, quais seriam os principais desafios técnicos e de dados para essa integração, e quais disciplinas da Inteligência Artificial, além do PLN, seriam indispensáveis?
10. Algumas vertentes de conteúdo enganoso são muito desafiadoras para detecção automática. As pós-verdades são um desses casos. Explique o que são pós-verdades e como um método computacional poderia detectá-las.



Capítulo 9

PLN na Saúde

Adriana Pagano

Claudia Moro

Elisa Terumi Rubel Schneider

Lilian Mie Mukai Cintho

Yohan Gumiel

Publicado em: 26/09/2023

Atualizado em: 16/04/2026

9.1 Introdução

A área da saúde é uma das mais importantes em nossas vidas e, nos últimos anos, tem se beneficiado do uso da tecnologia para melhorar o diagnóstico, o tratamento e a gestão de pacientes. A aplicação de Processamento de Linguagem Natural (PLN) tem sido fundamental para avançar nessa área, pois permite a análise de grandes volumes de dados não estruturados gerados em ambientes clínicos (Turchioe et al., 2022).

O domínio da medicina abrange diversos tipos de texto, utilizados para distintas atividades produtoras de significado, que desenvolvemos em nosso convívio social. Chamamos essas atividades de socio-semióticas. Estudos da linguagem baseados em pesquisas antropológicas modelam essas atividades socio-semióticas em oito tipos (Matthiessen et al., 2008; Matthiessen, 2013).

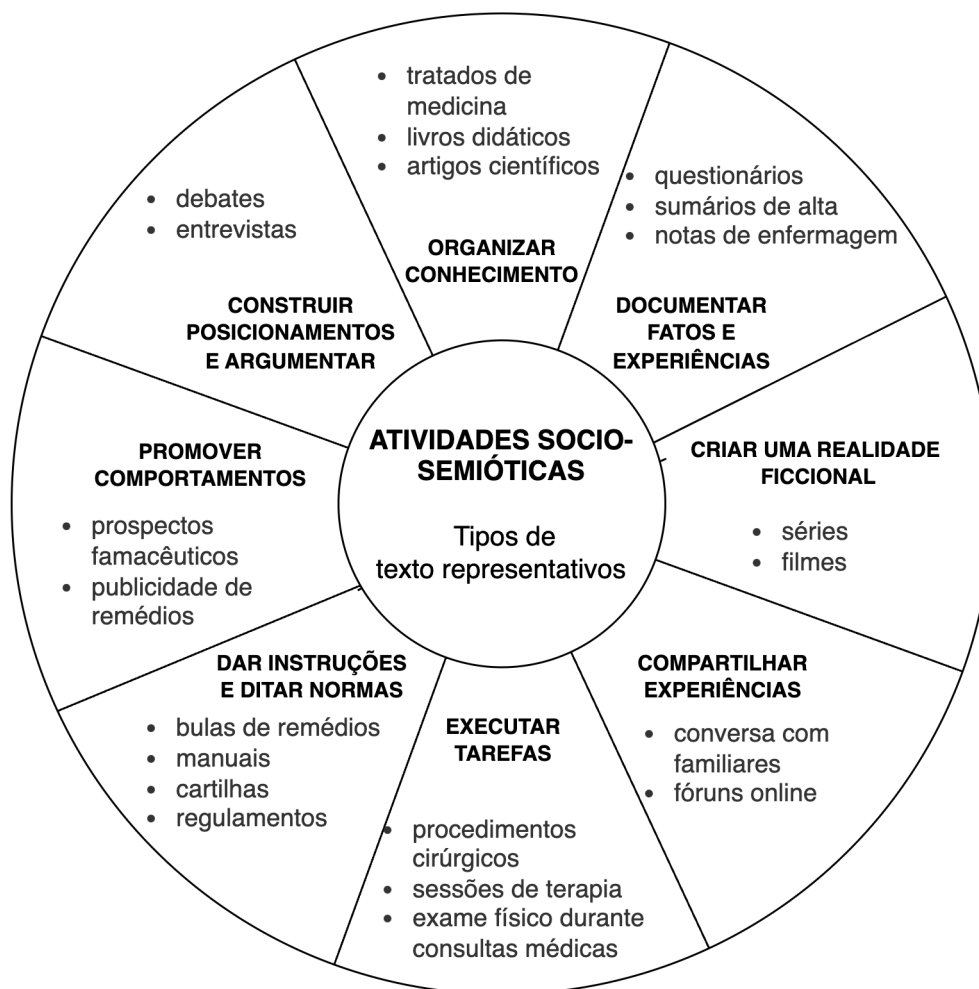
A Figura 9.1 mostra os oito tipos de atividades socio-semióticas e os tipos de texto mais representativos de cada um deles no domínio da medicina. Essas atividades são desenvolvidas por meio de textos escritos e falados, com funções específicas na nossa sociedade. Atividades nas quais a linguagem verbal tem um papel ancilar ou complementar são, por exemplo, a execução de procedimentos cirúrgicos, durante a qual ações podem ser verbalizadas ou não.

Mas, na grande parte das atividades humanas, a linguagem tem um papel constitutivo. Temos desde atividades que envolvem um uso especializado da linguagem para organizar a produção de conhecimento em tratados de medicina, livros didáticos e artigos acadêmicos, até atividades que envolvem um uso menos especializado, como o compartilhamento de experiências no âmbito privado, nas interações entre pacientes e familiares ou entre participantes de fóruns online sobre cuidados em saúde. Para a atividade de instruir e regular o comportamento, temos textos como bulas de medicamentos, cartilhas, normativas, manuais de instrução de equipamentos. Mesmo no domínio da medicina, há também textos pelos quais é construída uma realidade ficcional, como é o caso de séries e filmes que recriam interações em contextos médicos.

Uma atividade socio-semiótica muito relevante no domínio da medicina é documentar fatos e experiências, por meio de questionários aplicados ao paciente, registros de exames



Figura 9.1: Tipos de texto no domínio da medicina



clínicos e relatos de profissionais da saúde, nos quais são documentadas percepções sobre a saúde do paciente. Esses textos são conhecidos em PLN como **narrativas clínicas** e abrangem notas de evolução de enfermagem, sumários de alta, boletins médicos, e notas em texto livre em campos próprios do prontuário eletrônico do paciente. Cada um desses tipos de texto pode oferecer informações valiosas a serem obtidas por meio do PLN mais adequado às características do texto. Artigos acadêmicos, por exemplo, podem ser usados para a extração de ontologias, que são estruturas semânticas que permitem uma representação formal de conceitos, suas propriedades e relações. Essas ontologias podem ser usadas para facilitar a compreensão de termos técnicos e complexos em diferentes áreas da saúde, permitindo que as informações sejam compartilhadas de forma mais clara e precisa (Jiang et al., 2020). Também podemos identificar padrões e relacionamentos entre os dados e a construção de modelos preditivos (Lee et al., 2019).

Narrativas clínicas, por outro lado, são textos não estruturados que oferecem informações valiosas sobre a história do paciente, incluindo seus sintomas, histórico médico, estilo de vida e outras informações relevantes. A mineração desses dados pode ser usada para identificar padrões e relacionamentos entre os dados, permitindo uma melhor compreensão da condição do paciente e a construção de modelos preditivos para prever possíveis

complicações ou doenças (Wu et al., 2018).

Assim como nas narrativas clínicas, áudios produzidos por pacientes podem dar pistas de condições de saúde que afetam voz e fala, especialmente as que afetam o sistema respiratório. No Capítulo 11, o leitor encontra modelos e técnicas de classificação de áudio para a voz e fala de pacientes com problemas respiratórios.

9.2 O texto livre em narrativas clínicas

Com o advento do Registro Eletrônico de Saúde (RES)¹, como é denominado no Brasil, ou em inglês, o *Electronic Health Record* (EHR), a quantidade de dados gerados relativos à atenção aos pacientes aumentou significativamente. Os prontuários eletrônicos podem conter dados estruturados, semiestruturados ou não estruturados, todos eles oferecendo uma grande quantidade de informações sobre o paciente. A mineração desses dados pode ajudar a identificar tendências e padrões em relação a diagnósticos, tratamentos e resultados, permitindo uma melhor gestão do cuidado do paciente e um melhor planejamento da assistência (Shickel et al., 2017).

Os dados clínicos presentes nas narrativas clínicas em texto livre (dados não estruturados) apresentam características únicas que dificultam sua análise e interpretação. Esses dados são frequentemente apresentados em linguagem médica especializada, repleta de termos técnicos, jargões e abreviaturas que podem variar entre os distintos profissionais de saúde. Esses textos também podem conter erros de digitação, ortografia ou gramática, tornando a interpretação ainda mais complexa (Dalianis, 2018). A Figura 9.2 apresenta um exemplo de narrativa clínica adaptada para fins de ilustração. Nela podemos observar que as informações podem ser estruturadas de acordo com categorias destacadas com cores e rotuladas na legenda da figura.

No escopo do que chamamos narrativas clínicas, há diferentes tipos de texto, os quais apresentam desafios específicos em termos do tipo de linguagem e também da relevância das informações registradas. Por exemplo, as notas de evolução de enfermagem podem ser mais descritivas e detalhadas do que outros tipos de texto, enquanto os sumários de alta podem fornecer informações importantes sobre a condição atual do paciente e seu histórico de tratamento. Já as notas de ambulatório podem ser mais informais e fragmentadas, o que dificulta sua análise por modelos treinados com outros tipos de texto em outros domínios. Isso demanda a anotação manual de narrativas clínicas de forma contarmos com modelos mais refinados.

Como todo processo manual, a anotação de narrativas clínicas requer tempo e recursos, o que dificulta a construção de grandes *datasets* para treinamento de modelos de PLN. Como resultado, a aplicação de técnicas de aprendizado de máquina em dados clínicos sofre limitações pela disponibilidade de dados anotados manualmente (Koleck et al., 2019). Uma saída é utilizar modelos genéricos para pré-processamento, sendo a saída avaliada manualmente. Um exemplo deste tipo de trabalho é a anotação do *corpus* Depclin-Br, que vem sendo desenvolvida por uma equipe de cientistas da computação da PUCPR e de linguistas da Faculdade de Letras da UFMG. Trata-se de um conjunto de narrativas clínicas já anotadas em termos de entidades no domínio clínico e constituindo o *corpus* SemClinBr (Oliveira et al., 2022a). Uma parte desse *corpus* foi anotada morfossintaticamente com

¹No Sistema Único de Saúde (SUS), as informações dos usuários são coletadas e armazenadas por meio do Prontuário Eletrônico do Cidadão (PEC). Nele, há campos pré-determinados que podem ser preenchidos com texto livre.



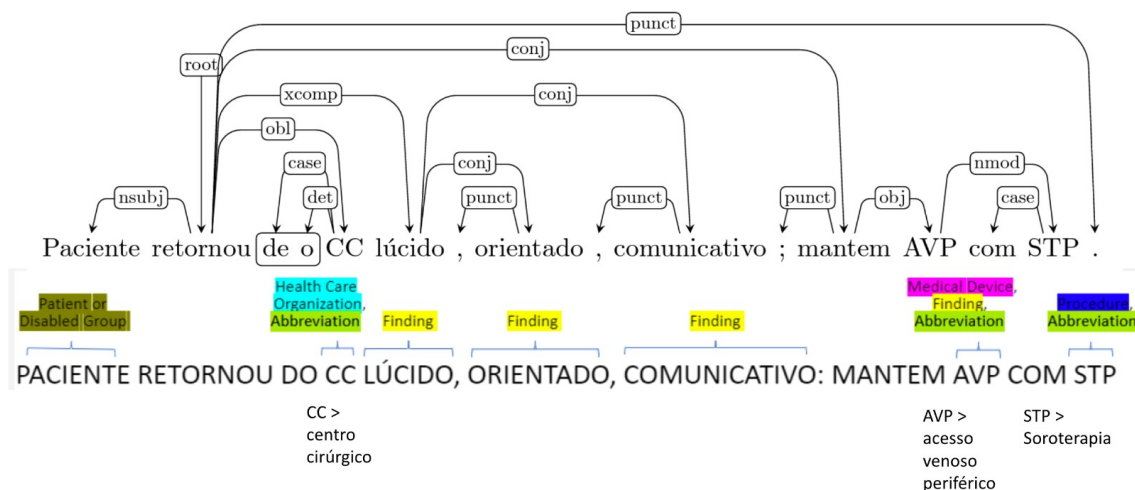
Figura 9.2: Exemplo de narrativa clínica elaborada para fins de ilustração. Na legenda, as categorias de informações que podem ser encontradas neste tipo de texto.

AP: HAS . Obesidade . DM. Tabagista.	Refere HAS desde 21 anos de idade, irregular controle da pressão arterial.
Ao exame: BEG, CHAAA, eupneico. ACV: BRNF, sem sopro. PA 159X100 mmHg, FC 89. AR: MV+ s/ RA . ABD : flácido , indolor a palpação , sem visceromegalias . MMII : edema bilateral ++/+4 .	
USG rins normal . Doppler MMII normais . ECG : BAV 1º G , BRD . TE Externo : baixa resposta cronotrópica , EEVV polimórficas frequentes , comportamento normal de PA . ECO TT Ext 19/06/20: AE 28 Septo 12 Parede Post 12 Massa VE 218 FEVE 70 % VE 48x29 Hipertrofia simétrica discreta do VE . Alteração do relaxamento do VE .	
Laboratório : 22/09/21: Hb:14,3 Leuco:4.960 Pla:204mil Colesterol:169 HDL:49 LDL:102 Triglic:91 Glicose:96 Ur:40 Cr:0,93 Na:142 K:4,3 TSH:2,08	
Oriento MEV , Restrição Hidrossalina . Oriento importância da adesão medicamentosa . Solicito ECO TT .	
Prescrevi : sinvastatina 40 mg comp . (1 comp vo 1xd) ; ácido acetilsalicílico 100 mg comp (1 comp vo 1xd) ; omeprazol 20 mg cáps; losartana 50 mg comp (1 comp vo 12/12h) ;	
Solicitei : Ecocardiograma com doppler colorido adquiridas ; hemoglobina glicada ; ácido úrico ; colesterol (sangue) ; CPK; creatinina / soro ; glicose / soro ; TGO	

Legenda	
Antecedentes pessoais	Exames laboratoriais
História pregressa	Orientações
Exame físico e sinais vitais	Medicamentos prescritos
Exames realizados	Planejamento

base num modelo genérico de português e a anotação revisada manualmente (Oliveira et al., 2022b). Essa primeira parte foi utilizada para refinamento do modelo genérico e anotação automática de uma segunda parte do *corpus*. Uma vez concluída a anotação, dados do *corpus* DepClinBr, anotado com relações de dependência, podem ser minerados e utilizados para caracterizar as entidades nomeadas previamente anotadas no SemClinBr. A Figura 9.3 ilustra a correlação de anotações morfosintáticas e entidades.

Figura 9.3: Correlação de anotações morfosintáticas e entidades.



A construção de *corpora* de narrativas clínicas (dados não estruturados) está sujeita a restrições técnicas e regulatórias, que dizem respeito à privacidade de dados. Essa especificidade limita a capacidade de construção de grandes *datasets* para treinamento de modelos de PLN (Chen; Chen, 2022). Como foi apontado, para contornar essa limitação,

são utilizados modelos genéricos da língua, os quais precisam ser refinados com dados específicos do domínio em um processo de *fine-tuning*, para melhorar ainda mais sua precisão e relevância (Lee et al., 2019).

A seguir, veremos alguns exemplos de aplicações da PLN em dados clínicos.

9.3 Aplicações de PLN na Saúde

9.3.1 Predição

Uma das principais tarefas de PLN na área médica é a predição, que pode ser aplicada em diversas demandas do cuidado em saúde, como diagnóstico, tratamento, evolução, alta médica hospitalar, detecção de quedas, detecção de depressão e outras. Essas demandas envolvem a classificação de dados clínicos, como narrativas de pacientes, prontuários eletrônicos, relatórios médicos e outros dados de saúde, para ajudar os médicos e outros profissionais de saúde a tomar decisões mais precisas. A predição de diagnóstico, por exemplo, pode ajudar a identificar doenças em estágios iniciais, permitindo tratamentos mais eficazes e prevenindo complicações. A predição de tratamento pode ajudar a personalizar o tratamento para cada paciente, maximizando sua eficácia e minimizando efeitos colaterais. A detecção de quedas e depressão pode ajudar a prevenir acidentes e melhorar a qualidade de vida dos pacientes. Em resumo, a tarefa de predição é essencial para a aplicação bem-sucedida de PLN na área de saúde (Yan et al., 2022).

Alguns exemplos de trabalhos envolvendo predição e classificação em textos clínicos em português são (Gonçalves et al., 2023; Santos et al., 2021a; Silva et al., 2023a; Yang et al., 2022).

9.3.2 Desidentificação

Um aspecto crucial na aplicação de PLN na área médica é a desidentificação dos dados dos pacientes, associada a processos de anonimização ou pseudonimização. Esta envolve a remoção de informações que possam identificar o paciente, como nome, endereço, número de telefone e outras informações pessoais. A anonimização é necessária para garantir a privacidade dos pacientes e cumprir as regulamentações de proteção de dados, como a Lei Geral de Proteção de Dados (LGPD) no Brasil² e a General Data Protection Regulation (GDPR) na União Europeia³.

A anonimização de dados clínicos é um processo desafiador, uma vez que esses dados contêm informações altamente sensíveis e complexas, como histórico médico, sintomas, exames, tratamentos e outros detalhes que podem identificar um paciente. Portanto, é necessário utilizar técnicas avançadas de PLN, como o uso de modelos de linguagem, para remover as informações sensíveis e garantir a privacidade dos pacientes (Jones et al., 2020).

Existem diversas técnicas que podem ser utilizadas na desidentificação dos dados clínicos, dependendo do tipo de informação que deve ser removida e do nível de anonimização desejado, por exemplo:

- Substituição de nomes próprios e outros identificadores pessoais por símbolos ou pseudônimos aleatórios;

²Lei Geral de Proteção de Dados Pessoais (LGPD), Lei nº 13.709/2018. Disponível em: <https://www.gov.br/pt-br/aceso-a-informacao/legpd>

³Data protection in the EU. Disponível em: https://commission.europa.eu/law/law-topic/data-protection/data-protection-eu_en



- Remoção de informações geográficas específicas, como endereço e CEP;
- Substituição de datas de nascimento e outras informações temporais por intervalos ou idades aproximadas;
- Remoção de informações de contato, como números de telefone e endereços de e-mail;
- Remoção de informações de identificação de instituições, como o nome de hospitais e clínicas.

Além dessas técnicas, também é possível utilizar métodos mais avançados de PLN, como a detecção e remoção de termos médicos específicos ou o uso de técnicas de de-identificação baseadas em modelos de linguagem, que tentam preservar a integridade semântica dos dados, mesmo após a remoção ou substituição das informações pessoais.

A desidentificação dos pacientes permite que os dados clínicos sejam utilizados para fins de pesquisa e análise, sem comprometer a privacidade dos pacientes. Isso é fundamental no avanço da medicina, permitindo a análise de grandes volumes de dados na descoberta de padrões e tendências em doenças, tratamentos e outros aspectos da saúde (Liu et al., 2017). Em (Santos et al., 2021b) temos um exemplo de trabalho para o português nessa tarefa.

9.3.3 Extração de conceitos clínicos

A busca e extração de conceitos clínicos relevantes é uma tarefa essencial na aplicação de PLN na área médica. Essa tarefa envolve a identificação de entidades relevantes nos dados clínicos, como sintomas, diagnósticos, tratamentos, medicamentos e outros termos específicos da área da saúde. Essa identificação geralmente é feita por meio de técnicas de NER (do inglês, *Named Entity Recognition*) (Capítulo 3), que permitem a identificação e classificação automática de entidades em textos não estruturados. A Figura 9.4 ilustra um exemplo de entidades do tipo *Problema* reconhecidas em uma narrativa clínica elaborada para fins de ilustração.

Figura 9.4: Exemplo de entidades do tipo Problema (em azul) encontradas em narrativa clínica.

Extração de conceitos clínicos: problemas

Paciente DM tipo 2, HAS e obesidade, com queixa de descontrole glicêmico necessitando procurar o PS 3x em 2 semanas. Refere que não consegue perder peso mesmo realizando dieta alimentar e caminhadas e faz uso regular dos medicamentos prescritos. Relata que apresenta disúria e hematúria há 2 dias. Apresenta PA 120X80 mmHg, sem edemas de MMII. Faz uso de atenolol 25 mg 2X/dia, metformina 850 mg 2X/dia. Solicitado exames laboratoriais (hemoglobina glicada, colesterol total, LDL e triglicérides e exame de Urina I). Aumento metformina 850 mg para 3X/dia e prescrevo Ciprofloxacino 250 mg cada 12 horas, durante 7 dias. Oriento a manter os medicamentos em uso e a modificar o estilo de vida e retornar para acompanhamento ambulatorial.

Além da identificação de entidades, outras técnicas de PLN também podem ser utilizadas para a busca e extração de conceitos clínicos relevantes, como a detecção de negação e a resolução de ambiguidades. A detecção de negação, por exemplo, é útil para identificar quando um sintoma é negado pelo paciente ou um diagnóstico dado pelo médico nega alguma condição. A precisão na detecção de negação é fundamental para a interpretação dos dados clínicos (Nath et al., 2022).



Outra técnica importante na busca e extração de conceitos clínicos é o mapeamento de terminologia, que consiste na associação dos termos clínicos encontrados nos textos com um conjunto de termos padronizados, como a Classificação Internacional de Doenças (CID) ou o *Systemized Nomenclature of Medicine* (SNOMED CT). Isso permite uma melhor organização e interpretação dos dados clínicos, facilitando a análise e a tomada de decisão médica (Fennelly et al., 2021).

A busca e extração de conceitos clínicos relevantes é fundamental para a análise de dados clínicos em larga escala, permitindo a identificação de padrões e tendências em doenças, tratamentos e outros aspectos da saúde. Além disso, essas técnicas de PLN também podem ser utilizadas para a construção de sistemas de suporte à decisão médica, que auxiliam os profissionais de saúde na escolha de tratamentos mais adequados para cada paciente (Demner-Fushman et al., 2009).

9.3.4 Relações temporais

Uma linha do tempo do paciente é uma representação gráfica que organiza as informações clínicas de um paciente de maneira cronológica. O interesse pela pesquisa em extração de relações temporais provém da característica longitudinal dos dados presentes nos RES. Esses registros contêm múltiplos textos clínicos referentes ao mesmo paciente, escritos em diferentes momentos (Gumiel et al., 2021).

A extração de relações temporais concentra-se na organização sequencial de menções em um texto, sendo essas menções eventos médicos ou expressões temporais.

No contexto clínico, eventos médicos são circunstâncias clínicas de relevância, cujo escopo é delimitado pelo contexto da aplicação. Por exemplo, para a extração de informações significativas para o diagnóstico, pode ser apropriado delimitar eventos como menções a tratamentos passados, sinais, sintomas, medicamentos em uso e exames realizados pelo paciente com os respectivos resultados. Já as expressões temporais envolvem menções de tempo, como a duração de um sintoma ou indicações de quando o paciente realizou determinada cirurgia. É notável que as expressões temporais só têm significado quando associadas a algum evento, enquanto os eventos podem fazer sentido quando relacionados entre si.

A fim de extrair essas menções do texto, são empregadas técnicas de Processamento de Linguagem Natural (PLN), como a Reconhecimento de Entidades Nomeadas. A tarefa de NER consiste em identificar e classificar automaticamente eventos e expressões temporais.

Com eventos e expressões temporais devidamente identificados, aplica-se a extração de relações temporais, uma técnica de PLN que se concentra na conexão de eventos entre si ou com expressões temporais. Desse modo, cada entidade acaba sendo relacionada a um período de tempo específico.

Ao considerar relações temporais no contexto clínico, diversas áreas de pesquisa emergem. Doenças crônicas, por exemplo, apresentam uma natureza longitudinal que torna a temporalidade extremamente relevante, pois existem fluxos de dados do paciente contínuos e extensos, nos quais podem ser extraídos padrões significativos (Sheikhalishahi et al., 2019). A progressão de uma doença e os eventos a ela associados são registrados cronologicamente, onde certos eventos são relevantes apenas em momentos específicos, como problemas médicos identificados durante um exame físico em uma consulta ambulatorial ou sintomas relatados (Sheikhalishahi et al., 2019). No caso de tratamento ineficaz de hipertensão com monoterapia, por exemplo, busca-se terapias com medicamentos combinados. Portanto, algumas informações sobre a progressão de doenças podem ser mais



facilmente discernidas por meio da extração de relações temporais (Gumiel et al., 2021).

A aplicação prática de uma linha do tempo na área da saúde pode ser utilizada para analisar a evolução do quadro clínico do paciente ao longo do tempo, identificar possíveis tendências e realizar previsões. Além disso, a linha do tempo do paciente pode ser integrada a sistemas de suporte à decisão médica, contribuindo para a seleção de tratamentos mais adequados para cada paciente.

9.3.5 Sumarização

A sumarização de evoluções clínicas é uma tarefa de PLN que tem como objetivo extrair as informações mais relevantes de um conjunto de dados clínicos, de forma a produzir uma versão resumida e legível dessas informações. A Figura 9.5 exibe um exemplo fictício de uma narrativa clínica sumarizada.

Figura 9.5: Exemplo fictício de uma narrativa clínica sumarizada, na qual as informações mais importantes foram mantidas.

Exemplo de Sumarização

Paciente hipertenso, dislipidêmico e sobrepeso, internou por 2 dias na clínica médica devido à crise hipertensiva e queixa de formigamento no MSD. Solicitado exames laboratoriais, ECG e cateterismo cardíaco. Sem apresentação de alterações enzimáticas e do ECG. Cateterismo sem lesões obstrutivas graves na CD, DP e VP, TCE sem lesões, DA e Dg1 sem lesões, CX, Mg1 e VPE sem lesões. Medicado e manteve a pressão arterial estável e sem queixas de dor. Mantido os medicamentos em uso: Atenolol 100 mg 1x/dia, Anlodipino 5mg 2xd e Atorvastatina 20mg a noite. Recebe alta hospitalar com a pressão arterial controlada e sem sintomas para dor. Oriento para a adesão medicamentosa, mudança de estilo de vida e acompanhamento ambulatorial.

Para realizar a sumarização de evoluções clínicas, são utilizadas técnicas de sumarização automática de texto (Capítulo 5), que podem ser baseadas em abordagens extrativas ou abstrativas⁴. Na abordagem extrativa, as frases mais importantes do texto original são selecionadas e combinadas para formar um resumo. Já na abordagem abstrativa, o resumo é gerado a partir da síntese das informações do texto original, gerando uma nova versão que não necessariamente contém as mesmas palavras e frases do texto original.

Para realizar a sumarização de evoluções clínicas, são utilizadas técnicas de processamento de linguagem natural, incluindo NER para identificar as entidades relevantes, PoS (*Part-of-Speech*) para identificar as partes do discurso e gramática do texto e também técnicas de análise sintática e semântica.

Essa tarefa de PLN é muito útil para os profissionais da área da saúde, pois permite que eles analisem brevemente as informações mais importantes dos pacientes, como histórico de doenças, exames realizados, tratamentos prescritos, entre outras informações clínicas (Gulden et al., 2019).

⁴Para projetos de sumarizadores em português, visite: <https://sites.icmc.usp.br/taspardo/sucinto/>



9.4 Recursos de PLN na área clínica

Como vimos, a aplicação de PLN na área clínica tem revolucionado a maneira como os dados médicos são processados e analisados. Com a crescente complexidade das informações clínicas e o volume crescente de dados gerados, a utilização de técnicas avançadas de PLN se torna essencial para extrair *insights* valiosos e melhorar a qualidade dos cuidados de saúde.

Nesta seção, vamos explorar alguns recursos de PLN em português utilizados no contexto clínico, com ênfase nos *corpora* e modelos.

9.4.1 Corpora

9.4.1.1 Semclinbr-corpus

SemClinBr (Oliveira et al., 2022a) é um *corpus* semanticamente anotado que contém um conjunto de dados clínicos em português brasileiro, com anotações detalhadas.

O *corpus* é composto por 1.000 notas clínicas provenientes de diversas especialidades médicas, tipos de documentos e instituições. Estas notas foram cuidadosamente anotadas para incluir informações valiosas sobre entidades e relações dentro dos textos clínicos.

SemClinBr inclui 65.117 entidades e 11.263 relações anotadas. As anotações cobrem uma variedade de aspectos clínicos, como condições médicas, medicamentos e procedimentos, além de incluir dicionários de negação e abreviações médicas.

A qualidade das anotações foi avaliada de forma intrínseca e extrínseca. A avaliação mostrou um alto nível de concordância entre anotadores, com índices variando de 0,71 a 0,92, e os resultados obtidos ao aplicar o corpus em tarefas de PLN demonstraram a utilidade e a confiabilidade das anotações.

SemClinBr fornece um recurso valioso para a pesquisa em PLN clínico, ajudando a desenvolver e avaliar novos algoritmos e técnicas.

Sendo um dos primeiros *corpus* clínico disponível em português brasileiro, o SemClinBr preenche uma lacuna significativa no campo e promove avanços na pesquisa biomédica e no desenvolvimento de tecnologias de PLN para o idioma.

O *corpus* está disponível para pesquisa acadêmica e pode ser solicitado neste endereço: <https://github.com/HAILab-PUCPR/SemClinBr>.

9.4.1.2 Detecção de Quedas

O *corpus* Detecção de Quedas (Santos et al., 2019a) é projetado para auxiliar na detecção de incidentes de quedas, que são uma das maiores categorias de relatórios de eventos adversos em hospitais.

A detecção eficiente desses eventos é crucial para melhorar a compreensão dos incidentes e, consequentemente, a qualidade do cuidado ao paciente.

O *corpus* contém 1.078 notas de progresso desidentificadas, que são registros clínicos detalhados sobre os pacientes durante sua estadia no hospital.

As notas foram anotadas para identificar incidentes de quedas, permitindo a avaliação de modelos de linguagem na tarefa de detectar tais eventos.

A avaliação do *corpus* foi realizada com vários modelos de linguagem, incluindo redes neurais recorrentes (RNNs) de última geração. Os experimentos mostraram que a abordagem de aprendizado profundo superou os trabalhos anteriores na detecção de eventos de queda.



O *corpus* anotado está disponível para fins de replicação em <https://github.com/nlp-pucrs/fall-detection>, permitindo que outros pesquisadores reproduzam os experimentos e utilizem os dados para melhorar os sistemas de detecção de eventos adversos.

9.4.1.3 Casos Clínicos - Neurologia

O *corpus* de Casos Clínicos - Neurologia (Lopes et al., 2019) é formado por 281 textos de casos clínicos coletados dos números 1 e 2 do volume 17 da revista clínica Sinapse, publicada pela Sociedade Portuguesa de Neurologia.

Os textos foram pré-processados com ferramentas do NLPPort, um kit de ferramentas de PLN para o português, baseado no OpenNLP.

Cada texto foi tokenizado com TokPort, etiquetado com TagPort e lemas para cada par token-PoS foram obtidos com LemPort. Após o pré-processamento, foi realizada a anotação manual de entidades nomeadas clínicas.

O *corpus* está disponível para pesquisa acadêmica e pode ser solicitado neste endereço: <https://github.com/fabioacl/PortugueseClinicalNER>.

9.4.2 Modelos

9.4.2.1 BioBERTpt

Os modelos da família BioBERTpt (Schneider et al., 2020) são modelos adaptados do modelo BERT multilíngue para o português brasileiro, com foco em tarefas de PLN nas áreas clínica e biomédica.

Essas versões foram ajustadas utilizando um *corpus* de narrativas clínicas e artigos científicos biomédicos em português.

Os modelos foram avaliados em duas coleções anotadas de textos clínicos para reconhecer entidades nomeadas. Comparado com modelos BERT existentes, o BioBERTpt apresentou uma melhoria de 2,72% no F1-score, superando o modelo base em 11 de 13 entidades avaliadas.

O desenvolvimento de BioBERTpt é um passo importante na área, ao melhorar o desempenho de tarefas de PLN em idiomas com menos recursos e pesquisa, como o português brasileiro, que frequentemente é sub-representado em modelos de PLN avançados.

Os modelos estão disponíveis em: <https://huggingface.co/pucpr>.

9.4.2.2 BioBERTptRT

O BioBERTptRT (De Oliveira et al., 2024) é uma versão do BioBERTpt projetado para a classificação de sentenças de notas clínicas, categorizando sentenças de acordo com o padrão SOAP (Subjetivo, Objetivo, Avaliação e Plano), o que ajuda a estruturar e organizar as notas clínicas de maneira mais eficaz.

O modelo foi treinado e ajustado usando uma base de dados privada de 10.000 registros de saúde anonimizados contendo 234.673 notas clínicas, divididas em 1.183.345 sentenças únicas. Além disso, 100.021 sentenças foram rotuladas manualmente para uso no ajuste fino dos modelos.

Esse treinamento especializado ajudou o modelo a entender melhor o contexto e a terminologia médica em português.

O modelo BioBERTptRT demonstrou um desempenho superior em comparação com outros modelos BERT em português, alcançando resultados superiores em precisão, acurácia, revocação e F1-score na tarefa de classificação de sentenças.



9.4.2.3 GPT2-Bio-Pt

GPT2-Bio-Pt (Schneider et al., 2021) é um modelo baseado em GPT-2 (*Generative Pre-trained Transformer 2*) especializado para o idioma português, desenvolvido para apoiar tarefas de PLN na área clínica e biomédica.

Um GPT-2 genérico em português foi ajustado para o domínio biomédico com a técnica de transferência de aprendizado, utilizando textos biomédicos escritos em português.

O modelo GPT-2 ajustado para o domínio biomédico foi testado em um conjunto de dados público, manualmente anotado para a tarefa de detecção de quedas de pacientes. O modelo especializado demonstrou um desempenho superior ao modelo GPT-2 genérico em português, com um aumento de 3,43 pontos em *F1-score*.

O modelo está disponível para pesquisa acadêmica e pode ser acessado neste endereço: <https://huggingface.co/pucpr/gpt2-bio-pt>.

9.4.2.4 BERT para Classificação ICD-10

(Coutinho; Martins, 2022) desenvolveram um modelo BERT especializado para a atribuição de códigos ICD-10 (Classificação Internacional de Doenças) para causas de morte, usando descrições em texto livre.

O modelo foi projetado para atribuir códigos ICD-10 a causas de morte com base em descrições de texto livre encontradas em certificados de óbito, relatórios de autópsia e boletins clínicos fornecidos pelo Ministério da Saúde de Portugal.

A abordagem inclui um procedimento de pré-treinamento que incorpora conhecimento específico do domínio clínico, permitindo a adaptação do BERT para a tarefa de codificação ICD-10.

O modelo também adotou uma estratégia de ajuste fino para lidar com o problema de desequilíbrio de classes, que é comum em tarefas de classificação quando algumas categorias são muito mais frequentes do que outras.

9.4.2.5 CardioBERTpt

CardioBERTpt (Schneider et al., 2023) é um modelo baseado em BERT adaptado especificamente para tarefas de PLN na área da cardiologia em português.

O modelo CardioBERTpt aproveita a transferência de aprendizado a partir de registros eletrônicos de saúde de um hospital terciário brasileiro especializado em doenças cardiológicas. Esta abordagem permite que o modelo capture nuances e terminologias específicas do domínio cardiológico.

Esse modelo demonstrou que a representatividade dos dados e um volume alto de dados de treinamento podem melhorar significativamente os resultados para tarefas clínicas.

O modelo está disponível para pesquisa acadêmica e pode ser acessado neste endereço: <https://huggingface.co/pucpr-br/cardiobertpt>.

9.5 Para onde estamos caminhando?

Embora a tecnologia de PLN na área clínica tenha avançado significativamente nos últimos anos, ainda existem vários desafios a serem superados. Alguns desses desafios incluem:

- Garantir a qualidade dos dados clínicos utilizados para treinar e testar os modelos de PLN, incluindo a devida anonimização e a padronização dos termos utilizados,



assegurando a ética e a privacidade dos dados clínicos;

- Desenvolver modelos de PLN capazes de lidar com textos clínicos mais complexos e heterogêneos, como notas de enfermagem, laudos médicos e textos escritos por pacientes;
- Integrar os modelos de PLN em sistemas de informação em saúde existentes, garantindo a interoperabilidade e a segurança dos dados;
- Garantir a aceitação e a adoção dos modelos de PLN pelos profissionais de saúde, demonstrando sua utilidade e eficácia na prática clínica.

É importante destacar que, embora o PLN possa ser útil na análise e interpretação de dados clínicos, ele não pode substituir a experiência e o conhecimento clínico de um médico ou de outros profissionais de saúde. A tecnologia pode ser uma ferramenta valiosa para auxiliar na tomada de decisões clínicas, mas não pode substituir o julgamento clínico humano. Ressalta-se que o desenvolvimento de tecnologias de PLN na área clínica seja visto como uma forma de complementar e melhorar o cuidado ao paciente, e não como uma substituição aos profissionais de saúde.



Capítulo 10

Detecção de Transtornos de Saúde Mental a partir de Texto

*Ivandr  Paraboni
Helena de Medeiros Caseli*

Publicado em: 20/11/2024

10.1 Introdu  o

Transtornos de sa de mental como depress o e ansiedade s o problemas bem conhecidos e uma crescente fonte de preocupa  o na sociedade moderna. Segundo dados do Global Burden of Disease Study de 2019¹, a depress o   o segundo fator que mais impacta na queda da expectativa de vida da popula  o global. Aliado   gravidade do quadro, a Organiza  o Mundial da Sa de (OMS) alerta que apenas um quarto dos indiv duos que sofrem algum transtorno de sa de mental recebem o atendimento adequado (WHO, 2021). No caso espec fico do Brasil, de acordo com Lopes et al. (2022), a preval ncia da depress o no pa s cresceu 36,7% de 2013 a 2019, indo de 7,9% em 2013 para 10,8% em 2019. Durante a pandemia de Covid-19, outra pesquisa (Zhang et al., 2021) realizada em maio de 2020 com 483 adultos constatou que 70,3% tinha sintomas de depress o e 67,2% tinham sintomas de ansiedade.

Ao mesmo tempo, diversos estudos demonstram que indiv duos com transtornos de sa de mental s o usu rios regulares de redes sociais em propor  o similar   popula  o em geral, que   estimada em cerca de 70% entre indiv duos de meia idade e at  97% entre indiv duos mais jovens (Aschbrenner, 2020; Aschbrenner et al., 2018; Birnbaum et al., 2017; Brunette et al., 2019), e que indiv duos com estes transtornos frequentemente recorrem  s redes sociais em busca do suporte de outros usu rios com problemas semelhantes (Bucci et al., 2019).

Este cen rio, e a observa  o de que transtornos de sa de mental frequentemente se refletem na linguagem empregada pelos indiv duos que sofrem dessas condi  es, levaram a um n mero expressivo de estudos no PLN e  reas relacionadas, tendo como foco a detec  o de transtornos como depress o, ansiedade, bipolaridade, anorexia, idea  o suicida e automutila  o em redes sociais (Cohan et al., 2018; Lin et al., 2020a; Mann et al., 2020; Shen et al., 2017; Shen; Rudzicz, 2017; Trotszek et al., 2018; Yates et al., 2017; Yazdavar et al., 2017). Este cap tulo visa apresentar um levantamento geral de estudos deste tipo, que procuram identificar casos de maior gravidade e eventualmente sinalizar a necessidade de um indiv duo buscar ajuda por meio da detec  o computacional de transtornos de

¹<https://ghdx.healthdata.org/gbd-2019>



depressão e ansiedade a partir de textos em português, em geral provenientes de redes sociais.

10.1.1 Definição do problema

A detecção computacional de indivíduos com transtornos de saúde mental a partir de textos provenientes de redes sociais e outras fontes é uma tarefa complexa e não totalmente resolvida na pesquisa em PLN, sendo tipicamente modelada como um problema de aprendizado de máquina (AM) supervisionado. Em abordagens deste tipo, um *corpus* de publicações (e.g., puramente textuais ou multimodais) rotulado com algum tipo de informação relativa à saúde mental (e.g., depressivo ou não depressivo) é utilizado para treinamento e teste de modelos de classificação do fenômeno de interesse.

Com base em um *corpus* textual rotulado com informações de saúde mental, há diversas formulações possíveis para o problema computacional a ser resolvido. Dentre estas, as mais comuns são a detecção de transtornos de saúde mental (e.g., decidir se um indivíduo é depressivo ou não, como em (Yates et al., 2017)), e a detecção de sintomas (e.g., decidir se um texto expressa um sintoma tradicional de depressão, como em (Yadav et al., 2020)). Outras formulações menos frequentes para o problema incluem o objetivo de determinar o grau de severidade de um determinado transtorno (Mann et al., 2020), ou diferenciar o discurso de indivíduos depressivos de outros meramente interessados em saúde mental (Santos et al., 2020c), ou identificar uma granularidade maior de sintomas e fatores protetivos e de risco para a depressão (Mendes; Caseli, 2024). Há, ainda, outra formulação popularizada pela série de desafios computacionais eRisk (Losada; Crestani, 2016) que é a tarefa de detecção precoce de risco de depressão e outros transtornos, em que o objetivo consiste em prever o estado de saúde mental de um indivíduo com base no menor volume possível de publicações em ordem cronológica, ou seja, o mais cedo possível.

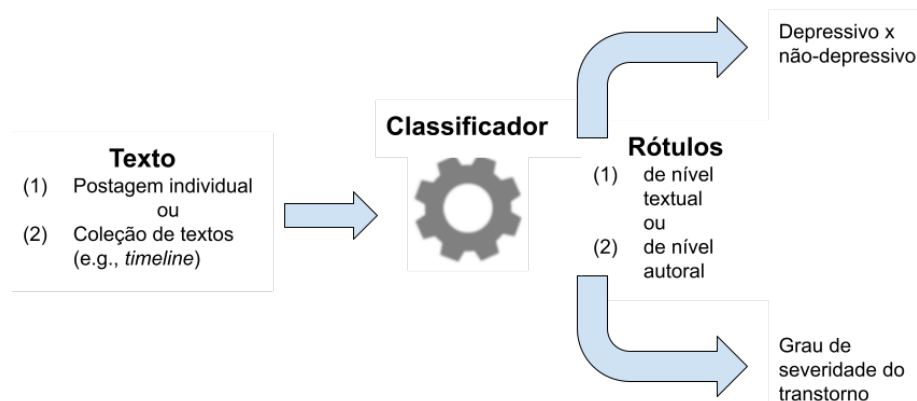
Outro aspecto importante da formulação do problema é o objeto da classificação. Parte dos estudos da área é dedicada à classificação de estados de saúde mental de indivíduos (e.g., usuários de rede social). Este tipo de abordagem, aqui denominada de **nível autorral**, toma por base uma coleção de publicações de um indivíduo (e.g., uma sequência de postagens em rede social, ou *timeline*) que pode ou não conter informações relevantes para a tarefa. Abordagens orientadas ao indivíduo são empregadas quando o objetivo é estabelecer o estado de saúde mental (e.g., depressivo *versus* não depressivo) de uma pessoa. Por outro lado, abordagens aqui denominadas de **nível textual** modelam o problema de classificação de textos individuais (e.g., postagens de rede social) independentemente da sua autoria. Estas abordagens tendem a ser mais utilizadas quando o objetivo é filtrar ou selecionar textos de maior relevância para uma aplicação da área de saúde mental, como a detecção de sintomas de depressão em redes sociais. A Figura 10.1 ilustra estas possibilidades considerando o exemplo da detecção de depressão, e estas abordagens são discutidas em mais detalhes nas seções a seguir.

10.1.1.1 Classificação de nível autorral

A classificação de nível autorral recebe como entrada uma coleção de textos e outros dados produzidos por (ou relativos a) uma determinada pessoa (e.g., uma coleção de postagens em redes sociais, meta-dados etc.), e objetiva determinar o estado de saúde mental (e.g., depressivo ou não depressivo) desta pessoa. A base de dados de treino e teste nestes casos contempla indivíduos com um transtorno de saúde mental conhecido (ou pertencentes à classe positiva, na terminologia de AM), que podem ser selecionados por meio de validação



Figura 10.1: Formulações do problema de classificação de depressão em nível textual e autoral.



externa (e.g., por meio de avaliação médica especializada ou uso de questionários, como em (Mann et al., 2020)), obtendo-se, por exemplo, um *corpus* de textos de indivíduo com e sem depressão. Este método de aquisição de dados, apesar de confiável, tende a ser de alto custo, o que na prática pode limitar o número de casos (ou instâncias) obtidos.

Como alternativa à validação externa, estudos da área de PLN têm explorado a ocorrência de **autorrelatos** em redes sociais, como em “Fui ao psicólogo hoje e ele me diagnosticou com esquizofrenia” para identificar rapidamente grandes volumes de indivíduos da classe positiva com custo mínimo. Autorrelatos deste tipo – que servem apenas para a seleção de usuários, mas que não são incluídos nos dados do *corpus* propriamente dito – são comuns em redes sociais, e embora sejam naturalmente mais sujeitos a imprecisões ou falsidade (o que de resto pode ser problemático também no caso da validação externa), é geralmente aceito que estas dificuldades tendem a ser compensadas pelo maior volume de dados obtido por este método (Coppersmith et al., 2015).

Dado entretanto que autorrelatos fornecem evidência apenas da possível **presença** de um transtorno, mas não da sua **ausência** – e que portanto não podem ser tomados por base para seleção de indivíduos que deveriam compor a classe negativa do aprendizado – faz-se necessário utilizar neste caso uma formulação distinta para o problema. Mais especificamente, ao invés de utilizar um conjunto de usuários (por exemplo) sem depressão, utiliza-se um grupo de usuários de controle selecionados aleatoriamente (e que portanto também inclui um certo número de indivíduos com depressão não detectáveis) de proporção várias vezes superior ao da classe positiva (normalmente proporcional à incidência do transtorno na população em geral). Assim, o problema computacional a ser resolvido não consiste em distinguir indivíduos com e sem depressão (e na verdade a informação sobre ausência de depressão é inexistente nos dados), e sim em **identificar indivíduos com probabilidade acima da média da população geral de virem a receber um futuro diagnóstico**. Esta probabilidade, que entre indivíduos da classe positiva será próxima de 100%, é necessariamente muito mais baixa em um grupo de controle aleatório, e pode assim ser capturada com métodos computacionais variados como o aprendizado profundo.

Em virtude destas características, o uso de autorrelatos para identificar indivíduos (provavelmente) diagnosticados com um transtorno de saúde mental configura um problema de AM fortemente desbalanceado. Mesmo com estas dificuldades, entretanto, a seleção de indivíduos com transtorno de saúde mental por meio de seus próprios autorrelatos é

possivelmente o método de coleta de dados mais popular na pesquisa na área de PLN (Cohan et al., 2018; Losada et al., 2019; Santos et al., 2024b; Yates et al., 2017).

10.1.1.2 Classificação em nível textual

A classificação de nível textual recebe como entrada um texto individual (e.g., uma postagem em rede social) e objetiva verificar se este texto expressa algum tipo de informação relacionada à saúde mental (e.g., se é de caráter depressivo ou não) sem levar em conta a identidade do seu autor. Modelos deste tipo são comumente empregados para detectar sintomas de depressão (Yadav et al., 2020), verificar a pertinência do texto para detecção de um transtorno (Santos; Paraboni, 2024), entre outras aplicações. A classificação em nível textual pode também ser tratada como um pré-requisito para a classificação de nível autoral, procurando filtrar textos de maior interesse para a tarefa em meio a uma grande massa de dados ruidosos (e.g., provenientes de redes sociais). Uma estratégia possível de combinação destes dois níveis de processamento é, por exemplo, o uso de aprendizado baseado em múltiplas instâncias (Mann et al., 2022).

Ao contrário da classificação em nível autoral, em que atribui-se um rótulo a uma coleção de textos (e.g., uma *timeline* de rede social) com base em um relato feito pelo próprio autor (ou usuário), a classificação em nível textual pressupõe a disponibilidade de textos rotulados individualmente com o tipo de informação a ser aprendida. Isso demanda um esforço adicional de anotação de *corpus*, ou algum método de rotulação automática implementado com uso de grandes modelos de linguagem (-cap-modelos-linguagem) e engenharia de *prompts* (Santos; Paraboni, 2023; Santos; Paraboni, 2024).

Um exemplo de projeto que trabalhou na classificação em nível textual foi o Amive² (Seção 10.5). Nesse projeto, postagens anônimas de páginas de segredos universitários foram coletadas por meio da busca por palavras-chave associadas à depressão e ideação suicida. Essas postagens foram, então, rotuladas com 21 sinais relacionados a depressão (sintomas, fatores de risco e protetivos) por estudantes e profissionais das áreas de psicologia, medicina e terapia ocupacional. Contudo, a anotação em uma granularidade maior, em dados de conteúdo sensível, por especialistas humanos é um trabalho bastante oneroso que demanda tempo e deve ser feito com cuidado para não afetar também a saúde mental dos anotadores.

10.2 Conjuntos de dados para detecção de depressão/ansiedade

A detecção de depressão e ansiedade a partir de textos faz o uso de *corpus* especialmente coletados e rotulados para este propósito. Recursos linguístico-computacionais deste tipo já foram desenvolvidos em quantidade expressiva para o idioma inglês, em alguns casos como subproduto de eventos do tipo *shared task* como o desafio CLPsych-2015 (Coppersmith et al., 2015) e a série eRisk (Losada et al., 2019). Iniciativas para o português incluem o estudo em (Mann et al., 2020) e o *corpus* SetembroBR (Santos et al., 2024b). Nesta seção apresentamos um levantamento de alguns recursos linguísticos-computacionais existentes na área. Para cada um dos estudos selecionados, a Tabela 10.1 apresenta o domínio, idioma e modalidade de dados, o número de instâncias de depressão/ansiedade, a relação entre usuários de controle (C) e diagnosticados (D) e o método de pareamento das classes. Detalhes adicionais são discutidos a seguir.

²<https://www.amive.ufscar.br/>



Tabela 10.1: Exemplos de *corpora* para detecção de depressão/ansiedade

<i>Corpus</i>	Domínio	Idioma	Modalidade	Dep.	Ans.	C/D	Pareamento
CLPsych-2015	Twitter/X	En	texto	477	-	1	idade, gênero
RSDD	Reddit	En	texto	9210	-	11,6	comportamento
Shen et al.	Twitter/X	En	texto	1402	-	1	na
eRisk-2017	Reddit	En	texto	135	-	5,6	na
eRisk-2018	Reddit	En	texto	214	-	7	na
SMHD	Reddit	En	texto	14139	8783	9	comportamento
Mann et al.	Instagram	Pt	texto, imagens	82	-	1,7	na
SetembroBR	Twitter/X	Pt	texto, rede soc.	1283	1767	7	gênero, data, tamanho
Amive	Facebook	Pt	texto	780	-	-	na

Com base neste levantamento, observa-se que os *corpora* existentes tendem a ser baseados principalmente nos domínios Reddit e Twitter/X, e que a maioria dos recursos listados são do tipo textual e dedicados ao idioma inglês. Para o português temos o já citado estudo em (Mann et al., 2020), que se destaca por ser o único a considerar o domínio Instagram e baseado no idioma português, e também por incluir dados de imagens; o *corpus* SetembroBR (Santos et al., 2024b) a ser discutido em mais detalhes na Seção 10.4 e único do gênero a incluir informações relativas à rede social dos usuários representados na base de dados; e o *corpus* do projeto Amive, apresentado em detalhes na Seção 10.5. Além disso, observa-se que os *corpora* SetembroBR e SMHD são os únicos que contemplam explicitamente casos de transtorno de ansiedade³. É importante observar, entretanto, que diversos métodos de detecção de transtornos de saúde mental discutidos na próxima seção fazem uso de dados multimodais (e.g., imagens e características da rede social etc.) mas que este tipo de informação, via de regra, não é parte integrante de um *corpus* publicamente disponibilizado para pesquisa.

No que diz respeito ao volume de dados, observa-se que os *corpora* considerados apresentam grande variação de tamanho (aqui medida em quantidade de indivíduos diagnosticados), e que recursos baseados em dados da rede social Reddit tendem a ser maiores. Isso é explicável pela organização dos dados em grupos de discussão (e.g., sobre depressão) nessa plataforma e, conseqüentemente, pela maior facilidade de identificação de indivíduos diagnosticados. Dados no domínio Twitter/X, por outro lado, são consideravelmente menos estruturados, o que torna mais complexa a tarefa de seleção de indivíduos com transtorno de saúde mental. No caso do *corpus* do projeto Amive, a quantidade apresentada na tabela indica o total de *posts* do Facebook coletados para serem anotados pelos especialistas em saúde mental a partir de busca por palavras-chave relacionadas à depressão e ideação suicida. Neste sentido, esse *corpus* difere dos demais uma vez que as instâncias não se referem a indivíduos diagnosticados, mas sim a postagens que podem ter sinais relacionados à depressão. Pela mesma especificidade desse *corpus*, não faz sentido indicar a proporção C/D nesse caso.

Observa-se de modo geral uma grande variação também na proporção entre indivíduos de controle e diagnosticados, indicada na coluna C/D da tabela, e que é um aspecto crucial da modelagem do problema de classificação a ser resolvido. Dentre os estudos selecionados, o *corpus* CLPsych-2015 é o único, dentre os mais amplamente utilizados na área, que se limita a um balanceamento artificialmente ideal, criado com o objetivo de simplificar o uso

³Como descrito na Seção 10.5, ansiedade é um dos sintomas contemplados no projeto Amive (Preocupação, medo ou ansiedade). Contudo, no escopo do projeto Amive, ele foi considerado um sinal relacionado à depressão e não foi usado especificamente para classificação do transtorno de ansiedade.

de métodos de AM no contexto da competição (ou *shared task*) de mesmo nome, enquanto que em todos os demais procurou-se considerar de alguma forma a necessidade de modelar um cenário mais realista. Dentre as estratégias adotadas, observa-se a modelagem de um grupo de controle de dimensão bastante superior ao do grupo de diagnosticados (Cohan et al., 2018; Losada et al., 2017, 2018; Santos et al., 2024b; Yates et al., 2017), o uso de múltiplas seleções de indivíduos aleatórios a partir de uma grande massa de dados (Shen et al., 2017), ou simplesmente o uso de inventários clínicos (Mann et al., 2020) como forma de selecionar indivíduos livres de depressão (a um custo possivelmente elevado, refletido no menor volume de indivíduos obtidos, porém dispensando o uso de um grupo de controle aleatorizado).

Outra questão fundamental na construção de um *corpus* deste tipo é o pareamento entre indivíduos diagnosticados e suas contrapartidas no grupo de controle, já que diferenças indesejáveis entre os dois conjuntos podem levar ao aprendizado de padrões espúrios. Por exemplo, caso indivíduos diagnosticados e de controle sejam extraídos de populações diferentes (e.g., com publicações feitas em períodos muito distintos), é possível que diferenças de vocabulário, decorrentes da evolução natural da língua, sejam tratadas como características relevantes para distinção entre as duas classes. Assim, como forma de assegurar que os dois grupos sejam minimamente comparáveis, o *corpus* CLPsych-2015 se destaca como o único a garantir pareamento de gênero e faixa etária entre indivíduos de ambas as classes, o que minimiza possíveis diferenças de vocabulário entre os dois grupos. O pareamento entre classes também é uma preocupação do *corpus* RSDD e sua extensão SMHD, em que indivíduos diagnosticados e de controle são selecionados com base no seu comportamento na rede social (e.g., grupos de discussão dos quais participam, frequência de postagens etc.). Este tipo de pareamento é um quesito importante no caso da plataforma Reddit, mas não possui contrapartida direta no domínio Twitter/X. No *corpus* SetembroBR, baseado na plataforma Twitter/X, foi realizado o pareamento por datas aproximadas de publicação, volume de postagens e gênero dos usuários. No caso do *corpus* do projeto Amive, uma vez que as postagens são anônimas, o pareamento com base em informações como gênero e idade não seria possível. Nesse caso, como o foco era o público alvo de estudantes universitários, buscou-se minimizar o impacto de algum viés variando-se as páginas de Segredos Universitários a partir das quais as postagens foram coletadas.

Com exceção do *corpus* SetembroBR a ser discutido em mais detalhes na Seção 10.4, os recursos aqui discutidos não levam em conta a noção de detecção **precoce** de transtornos, ou seja, anterior ao momento em que o indivíduo é oficialmente diagnosticado por um profissional da área de saúde. Em todos os casos discutidos, os dados constantes do *corpus* contemplam publicações feitas antes e depois do diagnóstico ou tratamento do indivíduo, e podem assim incluir um número potencialmente grande de indicadores a esse respeito. Por exemplo, uma vez que um indivíduo é diagnosticado ou começa a fazer tratamento para depressão, pode ser natural que ele passe a falar (ou falar mais) sobre este assunto, ou que relate as conversas que tem com o psicólogo, os efeitos da medicação etc. Todos estes indicadores, que em maior ou menor grau estão presentes em todos os *corpora* aqui discutidos, podem não só reduzir o grau de realismo do desafio computacional, como talvez reduzir a própria utilidade prática do método desenvolvido. Além disso, cabe observar que esta limitação está presente até mesmo nos *corpora* disponibilizados na série de desafios e-Risk (Losada et al., 2017, 2018, 2019), que mesmo possuindo o propósito explícito de detecção de primeiros sinais de transtornos de saúde mental como depressão, não fazem nenhuma distinção efetiva entre dados publicados antes ou depois do seu diagnóstico oficial, limitando-se apenas ao desafio computacional de classificar estes indivíduos com base no

menor volume de dados possível seguindo a ordem cronológica das publicações.

Finalmente, é importante destacar que *corpora* textuais rotulados com informações sobre saúde mental geralmente não são recursos computacionais totalmente acessíveis, como seriam *corpora* voltados a outros tipos de aplicação de PLN. Além da natureza sensível do problema e das considerações éticas associadas, algumas redes sociais como a plataforma Twitter/X proíbem a reprodução e divulgação de conteúdo gerado por seus usuários como forma de proteger a privacidade e o direito de arrependimento destes. Assim, *corpora* de tweets como SetembroBR, quando disponibilizados, são apenas coleções de identificadores numéricos a partir dos quais as postagens integrais podem ser recuperadas diretamente a partir da API da plataforma, caso ainda sejam disponibilizadas pelos seus autores. Este acesso, via de regra, acarreta algum esforço de programação e, no caso da plataforma Twitter/X, deixou de ser gratuito em fevereiro de 2023.

10.3 Modelos computacionais de detecção de depressão/ansiedade

A detecção de transtornos de saúde mental apresenta os desafios comuns à boa parte das aplicações de PLN. Em especial, observa-se que a distinção entre indivíduos diagnosticados e aqueles que pertencem ao grupo de controle com base em características linguísticas é naturalmente dificultada pelo fato de que ambos os grupos podem discutir qualquer tipo de assunto, incluindo questões relacionadas à saúde mental (deles próprios, de terceiros etc.), em uma variedade de contextos reais e imaginários, fazendo uso de humor, sarcasmo, metáforas, conjecturas e muitas outras formas de expressão. Além disso, no caso do aprendizado a partir de textos provenientes de redes sociais, observa-se que os dados de entrada são normalmente esparsos, isso é, com grande incidência de ruído na forma de postagens sem relação com o problema a ser tratado. É comum, por exemplo, deparar-se com *timelines* de usuários do Twitter/X que, apesar de sabidamente diagnosticados com depressão, não fazem qualquer referência a questões de saúde mental.

O que possui um texto para que seja classificado como relacionado a um determinado transtorno de saúde mental? De um lado, são conhecidos diversos indicadores linguísticos estabelecidos na literatura médica (Trifu et al., 2017). Indicadores deste tipo incluem o uso frequente de pronomes de primeira pessoa, tempo passado em verbos de ação, inversão da ordem de palavras em tópicos, ênfases, presença de sentenças curtas, impessoais, truncadas ou ríspidas, elipses, tautologias, repetições e ausência de comparação. O uso de termos absolutistas (e.g., todo, sempre etc.) também é mais fortemente associado a indivíduos com depressão (Johnstone, 2018), assim como o uso de expressões de conotação negativa em certas redes sociais (Rickard, 2018). Finalmente, alguns estudos exploram também o uso de informações não textuais como conexões da rede social, frequência e horário de publicações e outros (Choudhury et al., 2013; Shrestha; Spezzano, 2019).

Indicadores linguísticos de transtornos de saúde mental são normalmente modelados como sequências de *tokens* ou outras representações textuais tradicionais do PLN. Embora de valor indiscutível (por exemplo, para capturar o uso de referências pronominais etc.), a eficácia destes indicadores depende em grande parte da disponibilidade de um conjunto de treino suficientemente grande, onde todos os exemplos possíveis tenham ocorrido em número significativo para aprendizado automático. Assim, por exemplo, um modelo de AM reconhece a correlação entre sequências de *tokens* como “eu me sinto péssimo”, ou “psicóloga disse que eu tenho que me cuidar melhor” e saúde mental. Por outro lado,

exemplos mais raros, mesmo que altamente significativos, tendem a ser ignorados. Este é o caso por exemplo do uso de linguagem figurativa como em “acho que vou pular da ponte” como uma expressão de ideação suicida.

Uma possível forma de compensar a insuficiência de exemplos (que é em certa medida inerente a qualquer abordagem supervisionada) seria o uso de grandes modelos de linguagem como base de conhecimento auxiliar. Neste sentido, estudos como (Santos; Paraboni, 2024) demonstram que modelos como GPT (Brown et al., 2020c) podem capturar com facilidade noções complexas como a do exemplo anterior, e que este conhecimento pode ser combinado com abordagens tradicionais (e.g., supervisionadas) para detecção de transtornos de saúde mental em dois níveis, ou seja, considerando tanto indicadores linguísticos de baixo nível como sequências de *tokens* e conceitos de alto nível como sintomas. Detalhes desta estratégia são discutidos na Seção 10.4.

O Quadro 10.1 categoriza uma série de exemplos de estudos da área de acordo com o tipo de problema considerado (d=depressão, a=ansiedade, s=sintomas de depressão, v=grau de severidade da depressão, *=outros), tarefa (det=detecção, e.d = detecção precoce, exp=explicação), domínio (Reddit, Twitter/X etc.), idioma (En=inglês, Ch=chinês, Pt=português), características de aprendizado (e=*embeddings*, t=tópicos, n=informações de rede, u=informações referentes ao usuário, l=atributos psicolinguísticos LIWC, p=*part-of-speech*, d=dicionário de domínio, s=sentimentos/emoções, i=imagens, b=*bag of words*, m=metadados, t=informação temporal, ou texto simples), e métodos de aprendizado (e.g., CNN=redes neurais convolucionais, LSTM=*long short-term neural networks*, MLP=*multilayer perceptron*, NN=outras arquiteturas neurais, LR = regressão logística, RF=Random Forest, DT=árvore de decisão, NB=Naive Bayes, XG=XGBoost etc.). Detalhes adicionais são discutidos a seguir.

Quadro 10.1: Exemplos de modelos computacionais de detecção de depressão e ansiedade

Estudo	Alvo	Tarefa	Domínio	Líng.	Atributos	Abordagem
(Yates et al., 2017)	d	det	Reddit	En	e	CNN
(Yazdavar et al., 2017)	s	det	Twitter/X	En	t	LDA
(Shen et al., 2017)	d	det	Twitter/X	En	n, u, i, t, d, l, e	LR
(Shen; Rudzicz, 2017)	a	det	Reddit	En	e, t, l	MLP
(Loveys et al., 2017)	a,*	det	Twitter/X	En	s	VADER
(Cohan et al., 2018)	d,a,*	det	Reddit	En	b, e	FastText
(Song et al., 2018)	d	exp	Reddit	En	d, s, t, m, p	MLP+RNN
(Trotzek et al., 2018)	d	e.d	Reddit	En	e, p, m, t, d	CNN
(Nascimento et al., 2018)	d	det	Blogs	Pt	s	DT
(Shen et al., 2018)	d	det	Weibo	Ch	s, p, t, m, i, u, n	NN
(Arora, 2019)	d,a	det	Twitter/X	En	d, s, t	ensemble
(Aragón et al., 2019)	d	e.d	Reddit	En	s	SVM
(Carneiro, 2019)	d	e.d	Reddit	En	t, m, n	RF
(Santos et al., 2020c)	d,a,*	det	Twitter/X	Pt	e, b	LR
(Burdizzo et al., 2020)	d	e.d	Reddit	En	b	SS3
(Lin et al., 2020a)	d	det	Twitter/X	En	e, i	CNN
(Hitzler, 2020)	d	det	Twitter/X	En	b, s, t, i, n, l, u	RF
(Mann et al., 2020)	v	det	Instagram	Pt	i, e	NN
(Souza et al., 2021b)	a,d	det	Reddit	En	e	LSTM+CNN
(Ansari; Ji, 2022)	d	det	Reddit, Twitter/X	En	e,s	LR+LSTM
(Mann et al., 2022)	v	det	Instagram	Pt	e	BERT,LSTM
(Santos et al., 2024b)	d,a	det	Twitter/X	Pt	e, n	BERT,LSTM
(Santos et al., 2023)	d,a	det	Twitter/X	Pt	e	BERT,MoE
(Santos; Paraboni, 2023)	d	det	Twitter/X	Pt	texto	ChatGPT
(Oliveira; Paraboni, 2024)	d,a	det	Twitter/X	Pt	e, n	BERT,LSTM,LR
(Oliveira et al., 2024b)	d,a	det	Twitter/X	Pt	e, n	BERT,LSTM,LR
(Santos; Paraboni, 2024)	d	det	Twitter/X	Pt	texto	GPT
(Mendes; Caseli, 2024)	d	det	Facebook	Pt	e, l, p, d, b	SVM,LR,NB,RF,XG,BERT

No que diz respeito ao tipo de transtorno a ser detectado, observa-se que estudos focados na depressão (d) tendem a ser muito mais frequentes do que os focados no transtorno de ansiedade (a). Além disso, observa-se que o estudo de (Arora, 2019) trata da questão da

ansiedade depressiva, um transtorno misto que combina sintomas de depressão e ansiedade, e que o trabalho de (Souza et al., 2020b) investiga a comorbidade destes dois transtornos. Alguns estudos enfocam ainda a detecção de sintomas (Mendes; Caseli, 2024; Yazdavar et al., 2017) ou grau de severidade (Mann et al., 2020) destes transtornos.

Considerando-se a tarefa computacional a ser implementada, é possível observar que a maioria dos estudos identificados trata da detecção (det) de depressão/ansiedade ou sua detecção com base no menor volume possível de evidência (e.d), que é uma formulação específica do problema adotada na série de *shared tasks* eRisk (Losada et al., 2019), e/ou baseada nos *corpora* derivados destas competições. A preocupação com a necessidade de explicar os passos que levaram ao resultado da classificação é o foco do estudo de (Song et al., 2018).

Assim como no caso dos *corpora* discutidos na seção anterior, os estudos identificados são, com poucas exceções, baseados em dados nos domínios Twitter/X ou Reddit. Vários destes estudos desenvolvem seus próprios conjuntos de dados (especialmente quando considerando características não textuais da rede social ou outras não disponibilizadas em *corpora* de uso público), enquanto outros reutilizam um *corpus* de referência na área, como o RSDD (Yates et al., 2017), SMHD (Cohan et al., 2018) ou eRisk (Losada et al., 2017) no domínio Reddit, ou ainda o *corpus* de tweets discutido em (Shen et al., 2017).

Também conforme as características dos *corpora* discutidos na seção anterior, a maioria dos estudos identificados é dedicada ao idioma inglês (En). Há entretanto iniciativas isoladas para diversos outros idiomas, e que não são discutidos aqui porque fogem ao escopo deste livro. Estas iniciativas incluem os idiomas árabe (Almouzini et al., 2019), japonês (Tsugawa et al., 2015), coreano (Park et al., 2013), romeno (Briciu; Lupea, 2018), russo (Semenov et al., 2015), e tailandês (Katchapakirin et al., 2018). No caso da língua portuguesa (Pt), identificamos os estudos de (Mann et al., 2020, 2022), que tratam da questão da detecção de depressão em um *corpus* multimodal no domínio Instagram com base em questionários.

No que diz respeito aos modelos e métodos computacionais empregados, observa-se exemplos de estudos que exploram praticamente todo tipo de informação disponibilizada em redes sociais, incluindo dados textuais, temporais, estruturais da rede, e ainda características demográficas e comportamentais dos usuários, geralmente de forma combinada. A variedade de métodos de aprendizado também é significativa, com uma certa preferência por modelos neurais. Assim como em outras áreas do PLN, estudos mais recentemente passaram a utilizar modelos de língua pré-treinados do tipo BERT (Devlin et al., 2019), ELMo (Peters et al., 2017) e similares (Mann et al., 2022; Mendes; Caseli, 2024; Santos et al., 2024b), ou baseados em grandes modelos de linguagem e engenharia de *prompts* (Santos; Paraboni, 2023; Santos; Paraboni, 2024).

As próximas seções focam nos estudos especificamente desenvolvidos para o idioma português brasileiro, uma vez que este é o foco deste livro.

10.4 SetembroBR: detecção de depressão/ansiedade em português

Como forma de ilustrar o desenvolvimento de modelos de detecção de transtornos de saúde mental a partir de dados provenientes de redes sociais em Português, esta seção discute o caso do projeto SetembroBR, dedicado à detecção de depressão e transtorno de ansiedade no Twitter/X do Brasil.



A Seção 10.4.1 descreve o *corpus* construído, e a Seção 10.4.2 apresenta algumas abordagens computacionais para a tarefa em questão.

10.4.1 *Corpus* SetembroBR

O *corpus* SetembroBR é uma base de dados multimodal específica para o português, enfocando a detecção precoce de transtornos de depressão e ansiedade a partir de dados textuais e não-textuais no domínio Twitter/X⁴. A construção deste recurso partiu de um estudo-piloto em (Santos et al., 2020c), cuja definição da tarefa foi posteriormente reformulada e apresentada em sua forma final em (Santos et al., 2024b).

O *corpus* contempla *timelines* de usuários e meta-dados associados, como informações de suas relações na rede social e outros. Estes indivíduos foram selecionados com base em seus autorrelatos contendo indicação explícita do momento do diagnóstico (e.g., ‘Na semana passada o meu psicólogo me informou que eu tenho depressão’) feito por um profissional da área médica (e.g., psicólogos etc.).

Nesta abordagem, assim como em vários outros estudos da área (Cohan et al., 2018; Coppersmith et al., 2015; Yates et al., 2017), a seleção de usuários é feita de forma cuidadosa consultando-se a plataforma Twitter/X em busca de autorrelatos que correspondem a uma ampla gama de expressões regulares de interesse, e seguida de inspeção manual. Diferentemente de estudos prévios, entretanto, na presente abordagem a coleta dos dados (tweets) propriamente dita é feita também com inspeção manual, explorando-se a ordenação cronológica da plataforma Twitter/X de modo a restringi-los ao conjunto de dados anteriores ao evento relatado. Um exemplo de como esta porção de dados dita ‘útil’ (para fins de classificação) é delimitada é ilustrada pelo marcador [end] na Quadro 10.2 com base em um autorrelato indicado pelo marcador [msg].

Quadro 10.2: Timeline de um usuário com marcador [end] indicando o término da porção de dados a ser considerada na predição de depressão, em que todos os tweets abaixo do ponto [end] são descartados.

Data	Marcador	Texto
Mon March 25	[end]	Deixei meu celular em casa e agora a bateria está morta.
Tue March 26		Eu assisti esse filme duas vezes no ano passado.
Thu March 28		tão feliz que finalmente comprei meus óculos novos LOL
Mon April 1		Vou dormir agora. Amanhã é um grande dia.
Wed April 3		@usuário você nunca me contou isso.
Mon April 8	[msg]	Pensando em ligar para ela de novo hoje à noite...
Fri May 3		Oiiii! como você está?
Sun May 5		Mês passado o psiquiatra me diagnosticou com depressão :(

Neste exemplo, observa-se que no dia 5 de maio (na parte inferior da *timeline*), o usuário relata um diagnóstico recebido em uma data não especificada do mês anterior (abril). Assim, todos os tweets anteriores ao mês de abril, até o ponto indicado como [end], são coletados para suporte à tarefa de detecção precoce de depressão. O restante dos dados (a partir de 1o. de abril e incluindo o próprio autorrelato em [msg]) são descartados. Neste mesmo exemplo é possível ainda observar que, como em uma típica *timeline* de rede

⁴A escolha do domínio Twitter/X é motivada pela sua maior popularidade no Brasil.



social, a grande maioria das postagens tende a apresentar pouca ou nenhuma relação com questões de saúde mental.

O *corpus* coletado segundo esta metodologia, detalhada em (Santos et al., 2024b), consiste em um conjunto de 46,8 milhões de tweets (totalizando cerca de 555 milhões de *tokens*) em Português, produzidos por 18.819 usuários únicos. Dentre estes usuários, 3.163 são considerados, em função de seus autorrelatos externos ao *corpus*, como tendo um futuro diagnóstico de depressão, ansiedade ou ambos. Os demais usuários formam o grupo de controle na proporção de 7 indivíduos aleatórios para cada indivíduo diagnosticado.

De acordo com as restrições da plataforma Twitter/X, o *corpus* é disponibilizado para reuso no formato de lista de identificadores numéricos de cada tweet divididos em conjuntos de treino e teste padronizados, a partir dos quais o pesquisador interessado pode recuperá-los utilizando-se a API da rede social. Estão disponíveis⁵ informações para *download* e a rotulação do subconjunto Depressão com uso do modelo GPT 3.5., discutida na seção a seguir.

10.4.2 Modelos desenvolvidos no projeto SetembroBR

A seguir enumeramos alguns exemplos de uso do *corpus* SetembroBR aplicados à tarefa de detecção de depressão/ansiedade utilizando métodos computacionais variados. Para detalhes de cada abordagem e seus resultados, sugere-se consultar as publicações referenciadas em cada caso.

10.4.2.1 Abordagens supervisionadas tradicionais

A disponibilização do *corpus* SetembroBR incluiu a descrição de uma série de sistemas *baseline* construídos com uso de redes do tipo BiLSTM e do modelo de língua BERT (Devlin et al., 2019) e outros existentes na literatura da área para predição de depressão e transtorno de ansiedade a partir de dados textuais e não textuais. Dentre as arquiteturas consideradas, destaca-se o uso de LSTM que recebe como entrada uma representação em nível de usuário construída com uso de BERT. Nesta estratégia, a representação de indivíduo consiste de sequências consecutivas de 10 tweets iniciadas em uma posição aleatória da *timeline* do usuário a cada época. Os tweets de entrada são transformados em uma representação criada com uso do modelo BERTimbau (Souza et al., 2020a), e a última camada é usada como representação final da sequência. Esta arquitetura é seguida de uma camada BiLSTM com função de ativação ReLu que alimenta uma sequência de camadas totalmente conectadas com regularização *dropout*, e usando *softmax* como função de ativação na camada de saída. O desbalanceamento de classes é tratado com uso de uma função binária de entropia cruzada com pesos balanceados.

Esta arquitetura foi posteriormente aprimorada com uso de um modelo BERT treinado especificamente para o domínio do Twitter/X em português, denominado BERTabaporu (Costa et al., 2023). Este modelo foi treinado em uma massa de 238 mil tweets, totalizando cerca de 2,9 bilhões de *tokens*. Os resultados deste modelo foram comparados aos obtidos pelo tradicional BERTimbau (Souza et al., 2020a), treinado em cerca de 2,7 bilhões de *tokens*, para as tarefas de predição de transtornos de saúde mental e outras em . O modelo BERTabaporu propriamente dito está disponível para reuso no repositório Hugging Face⁶.

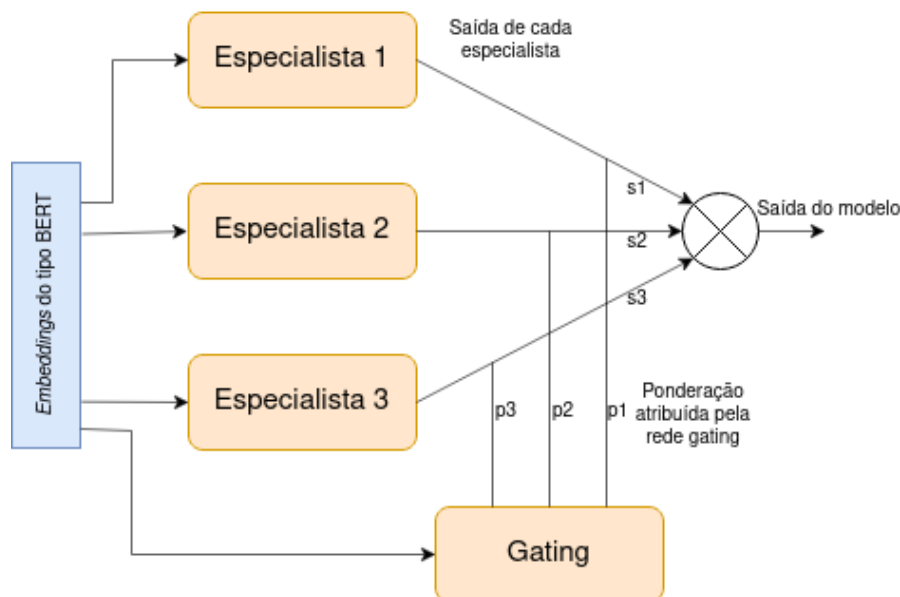
⁵https://drive.google.com/drive/folders/1MXFRs0u8iF1RNUWABTA0Oz8_Ix1skqZT?usp=sharing

⁶<https://huggingface.co/pablocosta/bertabaporu-large-uncased>



Finalmente, estes modelos foram combinados em uma arquitetura de comitê de máquina dinâmico de mistura de especialistas apresentada em (Santos et al., 2023), e que é baseada em (Jacobs et al., 1991). Esta arquitetura é constituída de três modelos especialistas com uma rede *gating* ponderadora na qual cada modelo especialista, assim como sua própria rede *gating*, consistem em uma rede do tipo LSTM que recebe como entrada uma representação em nível de usuário construída pelo modelo de representação distribuída BERT (Devlin et al., 2019). A Figura 10.2, extraída de (Santos et al., 2023), ilustra esta arquitetura.

Figura 10.2: Arquitetura de mistura de especialistas em (Santos et al., 2023)



Os experimentos apresentados em (Santos et al., 2023) confirmam a superioridade do modelo baseado em mistura de especialistas (MoE) usando BERT. Para a tarefa de detecção de depressão, MoE apresenta F1 médio de 0,65, enquanto a abordagem de engenharia de características apresenta F1 médio de 0,56. Para a tarefa de detecção de transtorno de ansiedade, são obtidos F1 médio de 0,60 e 0,53, respectivamente.

10.4.2.2 Abordagem *zero-shot* usando ChatGPT

Como alternativa às abordagens supervisionadas tradicionais, o estudo em (Santos; Paraboni, 2023) investigou a tarefa de predição de depressão com base nos dados do *corpus* SetembroBR em modalidade *zero-shot* com uso da ferramenta ChatGPT⁷. Para este fim, o modelo foi instruído a avaliar sequências aleatórias de 80 tweets anonimizados cada. A avaliação consistia em atribuir um escore em uma escala de 0 a 10 de acordo com a presença de indicadores linguísticos de depressão no texto, em que 0 significa nenhum indício e 10 significa indícios muito fortes. Com base nos textos assim rotulados, foi calculado então um *threshold* mínimo para diferenciar as classes positiva e negativa com base em dados de treinamento (de resto não empregados na presente abordagem *zero-shot*), obtendo-se assim o rótulo da *timeline* (ou seja, do autor dos textos) como um todo.

⁷<https://chat.openai.com/chat>

A abordagem *zero-shot* baseada em ChatGPT foi comparada com a arquitetura BERT tradicional discutida na seção anterior, obtendo resultado médio equivalente (F1 médio de 0,65 para a abordagem ChatGPT e 0,66 para BERT supervisionado), embora com grande diferença nas classes individuais. A abordagem baseada em ChatGPT apresentou ampla superioridade para a classe positiva (Diagnosticados) com F1 de 0,70 contra apenas 0,48 da abordagem BERT, enquanto que a abordagem supervisionada baseada em BERT foi superior para a classe Controle com F1 de 0,84 contra 0,60 da abordagem ChatGPT. O estudo em (Santos; Paraboni, 2023) argumenta que essa diferença é explicável pelo fato de que o modelo de linguagem subjacente ao ChatGPT não possui uma definição clara do que seria a população média representada pela classe Controle do *corpus*, enquanto que a abordagem supervisionada possui um grande volume de exemplos da classe, e teria assim melhores condições de distingui-la da classe positiva.

Como resultado, observa-se que a abordagem *zero-shot* pode ser considerada superior tanto pela acurácia obtida para a classe de interesse (que é a positiva, não a aleatória), mas igualmente pelo fato de não fazer uso de *corpus* rotulado. Cabe destacar, entretanto, que os resultados em (Santos; Paraboni, 2023) são preliminares e não podem ser comparados diretamente com os estudos anteriores dado que a abordagem foi testada apenas em um subconjunto dos dados de teste do *corpus* em virtude do custo associado ao uso da ferramenta ChatGPT.

10.4.2.3 Engenharia de *prompts* com supervisão

Inspirado nos resultados iniciais obtidos com uso da ferramenta ChatGPT, o estudo em (Santos; Paraboni, 2024) explorou a questão da natureza do conhecimento útil para detectar depressão a partir de dados textuais, e a necessidade de filtrar o ruído proveniente da rede social, ou seja, as postagens ditas irrelevantes para a tarefa.

Conforme discutido na Seção 10.3, observa-se que parte da solução para o problema parece ser, por um lado, encontrada no uso de características de teor mais linguístico, como sequências de *tokens*, o uso de certas formas pronominais e afins. Por outro lado, abordagens puramente linguísticas são menos úteis na captura de indicadores de mais alto nível, ou mais próximos da aplicação, como no caso de sintomas de depressão. Casos deste tipo, que exigem um processamento mais profundo (ou um maior ‘entendimento’ do texto) são tratados de forma eficiente com uso de *prompts* submetidos a um grande modelo de linguagem. O estudo realizado tratou assim de combinar estas duas visões do problema, ou seja, a abordagem supervisionada e a engenharia de *prompts*.

O estudo em (Santos; Paraboni, 2024) utilizou o modelo de língua GPT 3.5. (Brown et al., 2020c) para avaliar uma amostra aleatória de tweets do *corpus* SetembroBR de acordo com a relevância para a previsão de depressão. Como resultado, estes dados foram categorizados como sendo de alta (1), média (2) ou baixa (3) relevância para a saúde mental. Esta rotulação automática foi então utilizada para treinar um classificador de postagens de alta, média e baixa relevância de acordo com o modelo de língua T5 (Raffel et al., 2020), e este classificador foi então aplicado à rotulação do conjunto completo de 19,42 milhões de postagens do sub*corpus* de depressão. Este conjunto de rótulos em nível de postagem, provisoriamente chamada de SetembroBRGPT foi parcialmente disponibilizada (no momento apenas para o subconjunto Depressão) para reuso juntamente com o *corpus* original (rotulado ao nível de usuário).

Como esperado, a maioria das publicações (65,2% na classe Diagnosticado e 71,3% na classe Controle) são consideradas de baixa relevância para a predição de saúde mental de



acordo com o modelo GPT. Por outro lado, publicações de alta relevância são relativamente raras (5,4% na classe Diagnosticados e 3,0% na classe Controle). No entanto, tão ou mais importante do que esta distribuição é a observação de que, ao consultar um LLM sobre a relevância de uma publicação para a saúde mental, estamos nos concentrando em grande parte na semântica, isto é, em sintomas e outros sinais clínicos bem conhecidos de depressão, mas isto não quer dizer que outros fatores possam ou devam ser ignorados. Em outras palavras, o *prompt* utilizado permite ao modelo identificar uma vasta gama de publicações que podem sugerir, por exemplo, distúrbios alimentares ou uso de linguagem negativa, ambos sabidamente relacionados à depressão (American Psychiatric Association, 2013), mas não considera explicitamente indicadores linguísticos de depressão mais refinados, como o uso de pronomes de primeira pessoa (Trifu et al., 2017), termos absolutos (Johnstone, 2018) e outros. Estes indicadores, que podem estar presentes em qualquer publicação de baixa ou alta relevância para a depressão, são também importantes preditores do estado de saúde mental, e sugerem que nenhuma publicação deveria em princípio ser descartada com base apenas nos resultados do LLM.

Como um meio de manter os dados de treino integralmente disponíveis para o classificador e, ao mesmo tempo, distinguir mensagens mais e menos relevantes (que mostram claramente diferentes distribuições no *corpus*, conforme discutido em (Santos; Paraboni, 2024)), foi proposta uma abordagem que usa os rótulos fornecidos pelo LLM para dividir os dados de treinamento em subconjuntos de alta, média e baixa relevância, a partir dos quais foram criados três vetores individuais de contagens TF-IDF reduzidos com seleção univariada. Além disso, como forma de capturar também a noção de ordem das postagens de diferentes graus de relevância, a esta representação foi anexado ainda um vetor de contagens de bigramas dos rótulos alto/médio/baixo computados. Esta adição foi motivada pela observação de que certas sequências de postagens (por exemplo, uma ocorrência de vários tweets consecutivos e altamente relevantes para a detecção de depressão) podem ser indicativas do estado de saúde mental do indivíduo.

A arquitetura proposta procura assim capturar (i) a noção de postagens mais ou menos importantes (e.g., postagens expressando sintomas de depressão etc.), (ii) os indicadores linguísticos de baixo nível que podem estar presentes em qualquer postagem, incluindo as que não tenham nenhuma relação com saúde mental, e ainda (iii) a noção de sequências de postagens consecutivas com maior ou menor importância para a tarefa. O conjunto final de 55.000 características empregadas para esse fim (i.e., 15.000 contagens de *tokens* mais 40.000 contagens de bigramas de rótulos) é ilustrado na Figura 10.3.

Figura 10.3: Modelo de 15.000 tokens e 40.000 bigramas de rótulos alto/médio/baixo.

Tokens (TF-IDF)			Bigramas de rótulos alto/médio/baixo
alto	médio	baixo	
6000	6000	3000	40000

Os resultados reportados em (Santos; Paraboni, 2024) apontam uma pequena vantagem da abordagem baseada em *prompts* (F1 médio de 0,66) sobre a arquitetura de mistura de especialistas em (Santos et al., 2023) (com F1 médio de 0,65). É possível, entretanto, que esta abordagem possa ser aperfeiçoada com uso de modelos do tipo BERT ao invés dos modelos de bigramas utilizados pelo componente supervisionado da arquitetura.

10.5 Amive: detecção de sintomas de depressão em português

O projeto Amive⁸ – um acrônimo para **Amigo** virtual especializado – foi desenvolvido na UFSCar⁹ (Universidade Federal de São Carlos), de 2021 a 2023, por uma equipe multidisciplinar de pesquisadores e estudantes da computação e da área de saúde mental (psicologia, terapia ocupacional e psiquiatria). O projeto Amive visava combinar dados fisiológicos coletados por sensores presentes em um relógio inteligente (smartwatch) com as informações textuais das postagens de redes sociais para: (i) a identificação de um Possível Perfil Depressivo (PPD) e (ii) a oferta de intervenções personalizadas projetadas para o público-alvo de estudantes universitários (Cássia Alves et al., 2024).

Para tanto, um conjunto de sinais associados à depressão e seus fatores de risco e protetivos foram definidos pelos especialistas em saúde mental da equipe, conforme descrito na Seção 10.5.1. Focando no público-alvo de estudantes universitários, o *corpus* foi construído a partir de postagens em páginas do Facebook utilizadas por esse público, conforme descrito na Seção 10.5.2.

10.5.1 Sinais associados à depressão

Focando especificamente no uso das informações textuais para a etapa de identificação de PPD, a equipe de especialistas em saúde mental definiu um conjunto de 18 sintomas de depressão e 3 sinais adicionais (19, 20 e 21) descritos brevemente a seguir:

1. **Agitação ou inquietação** – Característica de não conseguir ficar parado fisicamente ou diante de uma situação. Percepção de que o corpo está agitado, não conseguir controlar o movimento dos membros (balançar a perna, andar de um lado para o outro, incapacidade de ficar sentado, agitar as mãos, puxar ou esfregar a pele, roupas ou outros objetos). Mente agitada e inquieta (muitos pensamentos, incômodo mental). Necessidade de fazer algo, de se mover.
2. **Alteração de peso ou apetite** – Relato de ganho ou perda de peso. Comer em excesso ou comer bem menos que o habitual. Muita ou pouca vontade de comer.
3. **Alteração de sono** – Caracterizado por não dormir de forma adequada. Dormir muito pouco. Dormir em excesso. Ter insônia. Acordar muitas vezes durante a noite. Sentir sono durante o dia. Acordar e não conseguir voltar a dormir. Cansaço e fadiga por não ter dormido.
4. **Alteração na eficiência ou funcionalidade** – Caracterizado pela diminuição na eficiência durante a realização de alguma tarefa. Perda de desempenho. Gastar mais tempo do que o habitual para realizar tarefas. Efetuar maior esforço para realizar tarefas. Falta de iniciativa. Interromper uma atividade. Desistir de realizar tarefas.
5. **Cansaço, desânimo, desencorajamento, fadiga, perda de energia ou lentificação** – Nesta categoria de sintomas juntamos diversos sintomas semelhantes, sendo que podem aparecer em conjunto na mesma sentença ou separadamente, devendo ser anotados de qualquer forma (em conjunto ou separados). Caracterizados por cansaço excessivo, falta de energia física ou mental, desgaste físico ou mental, fraqueza, falta de ânimo, falta de vontade, não ter motivação. Falta de coragem para resolver ou

⁸<https://www.amive.ufscar.br/>. Projeto apoiado pela FAPESP (# 20/05157-9).

⁹<https://www.ufscar.br/>



fazer algo. Perda de interesse para realizar alguma tarefa. Não ter estímulo. Perda do entusiasmo. Desejo de desistir diante de obstáculos percebidos como insuperáveis. Perda da agilidade. Demorar para realizar tarefas. Discurso, pensamento ou movimentos corporais lentificados. Maiores pausas antes de responder. Fala diminuída em termos de volume, inflexão, quantidade ou variedade de conteúdos. Não querer falar ou fazer. Preguiça. Falta de pressa, de vigor, de dinamismo.

6. **Desamparo, prejuízo social ou solidão** – Caracterizado pelo sentimento de não ter apoio e ajuda de outros. Sentimento de abandono. Necessidade de apoio. Prejuízo nas interações sociais. Isolamento. Não dar atenção a outras pessoas. Não querer ir para o trabalho, nem para escola para não ter que encontrar com pessoas, não fazer questão de participar de grupos, não querer encontrar amigos, perda do papel social. Desinteresse em estar com pessoas.
7. **Desesperança** – Caracterizado pela falta de esperança. Não enxergar perspectivas positivas seja no presente ou futuro. Não acreditar que algo pode ser bom, que o futuro pode ser melhor, que a própria pessoa pode melhorar. Sentimento de que não há saída ou que não resta mais nada.
8. **Desvalia ou baixa autoestima** – Caracterizado pelo sentimento de não ter valor. Inutilidade. Aversão a si mesmo. Sentimento de inferiorização, depreciação ou diminuição do valor pessoal. Sentimento de não ser merecedor de algo. Não reconhecer as próprias qualidades. Avaliações negativas e irrealistas do próprio valor. Sensação de fracasso. Imagem negativa sobre si mesmo. Não reconhecer a própria beleza (interior e exterior).
9. **Dificuldade para decidir** – Caracterizado pela dificuldade de tomar decisões. Incerteza. Dúvidas. Dificuldade para fazer escolhas.
10. **Deficit de atenção ou memória** – Caracterizado pela alteração da capacidade de concentração e memória. Não conseguir manter a atenção em algo. Distração excessiva. Não conseguir prestar atenção adequadamente. Perda de memória. Ter dificuldade para lembrar do que precisa fazer, do que fez. Esquecimento repentino e frequente. Capacidade diminuída para pensar.
11. **Irritação ou agressividade** – Caracterizado por sentir-se irritado, impaciente. Agressão verbal ou física. Provocação, hostilidade. Raiva persistente. Ataques de raiva.
12. **Perda ou diminuição do prazer ou da libido** – Caracterizado pela perda ou diminuição de prazer pela vida ou por atividades que antes eram agradáveis. Diminuição ou perda de satisfação pessoal ou com a vida. Diminuição ou perda do sentimento de realização. Não sentir alegria ou felicidade na realização de tarefas que antes geravam estes sentimentos. Caracterizado pela alteração no desejo sexual. Perda de vontade sexual. Diminuição do interesse sexual.
13. **Preocupação, medo ou ansiedade** – Neste marcador juntamos preocupação, medo e ansiedade, que são sintomas muito parecidos, podem aparecer em conjunto ou separadamente, mas a marcação se dará independente de estarem separados ou juntos. Podem ser caracterizados por: Preocupações exageradas e contínuas e/ou pensar excessivamente em algo e/ou pensamento fixo em algo e/ou medo em excesso



e/ou antecipação de algo que ainda não aconteceu e/ou angústia e/ou sentimento de que algo negativo vai acontecer e/ou avaliação negativa de um acontecimento que ainda não aconteceu e/ou catastrofização e/ou identificar perigo ou ameaça em algo que não apresenta perigo ou ameaça e/ou crises de pânico.

14. **Sentimento de culpa** – Sentimento de ser responsável por algum dano ou prejuízo causado a outra pessoa ou a si mesmo. Arrependimento por alguma atitude ou decisão tomada.
15. **Sentimento de vazio** – Caracterizado por algo que falta. Não conseguir perceber sentimento algum.
16. **Sintoma físico** – Qualquer sintoma físico apresentado. Dor de cabeça. Dor muscular. Batimentos cardíacos alterados para mais ou para menos. Respiração acelerada. Sudorese. Tremedeira. Tontura. Sensação de desmaio. Enjoo. Enxaqueca. Dormência. Falta de ar. Calor excessivo. Entre outros. Boca seca, flatulência, indigestão, diarreia, cólicas, palpitação, cefaleia, hiperventilação, suspiros. Alteração de pressão arterial.
17. **Suicídio ou auto-extermínio** – Caracterizado por ideias suicidas. Pensamentos recorrentes de morte. Planejamento da própria morte. Tentativa de suicídio. Desejo passivo de não querer acordar pela manhã. Acreditar que os outros estariam melhor se o indivíduo estivesse morto. As pessoas mais gravemente suicidas podem ter colocado seus negócios em ordem (p. ex., atualizar o testamento, pagar as dívidas), podem ter adquirido materiais necessários (p. ex., corda ou arma de fogo) e podem ter estabelecido um local e momento para consumarem o suicídio. Intenso desejo de pôr fim a um estado emocional extremamente doloroso.
18. **Tristeza ou humor depressivo** – Caracterizado pela tristeza, que pode durar mais do que o habitual. Sentir-se para baixo, desanimado, sem vontade. Negatividade, pessimismo. Melancolia. É importante que a gravidade e/ou a duração do sintoma de tristeza / humor depressivo sejam considerados na anotação.
19. **Fator de risco** – Fator relacionado ao ambiente ou a uma visão geral do mundo. Bullying. Sofrer muita pressão no trabalho. Racismo. Machismo. Pobreza extrema. Adoecimento ou comorbidade. Desemprego, problemas de moradia.
20. **Fator protetivo, cuidado em saúde e bem-estar** – Recursos que auxiliam no enfrentamento de problemas. Resiliência. Apoio de amigos. Apoio de familiares. Estar em acompanhamento ou realizar atividades que promovam o bem estar, que dão prazer ao usuário, sentimento de pertença ou sentido à vida.
21. **Morte ou suicídio de outro** – Aqui, com esta etiqueta, gostaríamos de deixar anotado quando uma pessoa está se referindo à morte de outras pessoas, estas postagens podem ser caracterizadas pelo relato de tentativas de suicídio ou o suicídio cometidos por outra pessoa, um familiar, conhecido ou pessoa pública, também devem ser anotados relatos de sentimento de luto e perda de pessoas queridas ou que a pessoa tenha apreço.

A motivação para uma investigação mais granular, e não apenas binária, de sinais de depressão em postagens textuais foi derivada do fato de que embora se caracterize por um



sofrimento psicológico clinicamente significativo que traz vários impactos nas funcionalidades e na qualidade de vida dos indivíduos, a depressão se manifesta de forma diferente nos indivíduos. Nesse sentido, a detecção de sintomas de depressão e fatores associados pode auxiliar os profissionais de saúde mental a oferecer a melhor forma de auxílio aos indivíduos identificados como PPD.

10.5.2 *Corpus* do projeto Amive e anotação manual

Um conjunto de postagens de páginas públicas do Facebook foi construído a partir da busca, em páginas de segredos universitários¹⁰, por expressões e palavras-chave: “suicídio”, “depressão”, “me corto”, “vontade de viver”, “me matar” e “quero morrer”. As postagens foram coletadas por meio da plataforma Crowdtangle, com data de postagem entre Janeiro de 2012 e Dezembro de 2021. 780 postagens foram selecionadas para anotação manual. Dado o conteúdo sensível, antes da anotação, um processo manual de anonimização adicional foi realizado para substituir referências a lugares, instituições, eventos, datas e cursos, por etiquetas genéricas como <CIDADE>.

As 780 postagens foram anotadas por 4 estudantes de psicologia, psiquiatria ou terapia ocupacional de gênero feminino e masculino. A tarefa de anotação consistia em selecionar trechos de texto característicos de um ou mais dos 21 sinais. Um rótulo especial também foi criado para indicar se a postagem como um todo era indicativa de um PPD: <PPD>. Assim, o *corpus* anotado resultante possui tanto a granularidade fina de instâncias para os 21 sinais de depressão quanto a anotação binária tradicionalmente usada na literatura.

A Tabela 10.2 traz a quantidade de instâncias anotadas para cada um dos 21 sinais de depressão. Vale destacar que um dos sinais mais frequentes foi o de “Suicídio ou auto-extermínio” o que pode ter sido decorrente da escolha pelas expressões e palavras-chave usadas na coleta das postagens. Destaca-se também a grande quantidade de instâncias para os fatores de risco e protetivos, indicando a relevância deles no contexto de transtornos de saúde mental em universitários. Por fim, como esperado, o sintoma mais frequente é o mais característico do quadro depressivo: “Tristeza ou humor depressivo”.

Tabela 10.2: Quantidade de instâncias (#) anotadas para os 21 sinais de depressão definidos no projeto Amive, em ordem crescente

Sinal	#
Agitação ou inquietação	5
Deficit de atenção ou memória	13
Alteração de peso ou apetite	15
Perda ou diminuição do prazer da libido	17
Sintoma físico	24
Dificuldade para decidir	26
Alteração de sono	27
Sentimento de vazio	34
Sentimento de culpa	48
Irritação ou agressividade	88
Cansaço, desânimo, desencorajamento, fadiga, perda de energia ou lentificação	106

¹⁰As páginas de segredos universitários publicam postagens anônimas de estudantes enviadas por meio de um formulário, também anônimo. Após o envio, as postagens ainda passam por uma moderação antes de serem publicadas.



Sinal	#
Desesperança	112
Alteração na eficiência ou funcionalidade	113
Preocupação, medo ou ansiedade	140
Desvalia ou baixa autoestima	168
Suicídio ou auto-extermínio	244
Desamparo, prejuízo social ou solidão	283
Tristeza ou humor depressivo	341
Morte ou suicídio de outro	47
Fatores protetivos	167
Fatores de risco	286

10.5.3 Identificação automática de sinais de depressão

Mendes; Caseli (2024) investigaram a identificação automática dos 21 sinais de depressão usando estratégias de aprendizado de máquina baseadas em atributos – SVM, regressão logística, naive Bayes, RandomForest e XGBoost – bem como o ajuste fino de modelos neurais – BERTimbau, mBERT, metalBERT e mentalRoBERTa. Entre os atributos investigados estão o uso de dicionários específicos (LIWC, AnewBR), categorias da PHQ-9, etiquetas de *part-of-speech* e traços morfológicos, métricas geradas pelo NILCMetrix e *embeddings* estáticas (Word2Vec, FastText, GloVe e LexVec, TF-IDF) e contextualizadas (BERTimbau).

De modo geral, as estratégias de melhor desempenho na identificação dos sinais de depressão foram as baseadas no ajuste fino dos modelos neurais, em especial o BERTimbau que foi o modelo de melhor desempenho para 8 dos 20 sinais¹¹. Contudo, vale ressaltar que apenas 5 sinais obtiveram uma precisão média de 70% ou mais: Sintoma físico (70,0%), Suicídio ou auto-extermínio (71,2%), Tristeza ou humor depressivo (74,5%), Preocupação, medo ou ansiedade (79,4%) e Alteração de sono (90%).

Entre as principais conclusões de (Mendes; Caseli, 2024) tem-se o fato de que o desempenho dos modelos computacionais está mais associado à complexidade da identificação de alguns sinais do que à quantidade de instâncias disponíveis para treinamento. Por exemplo, com apenas 17 instâncias de um sinal aparentemente simples de ser expresso como “Alteração de sono” o modelo alcançou 90% de precisão enquanto que sinais que geralmente são expressos em uma linguagem mais complexa, mais abstrata e menos direta, como “Desesperança”, mesmo com 5 vezes mais instâncias, foi identificado pelo melhor modelo com apenas 38% de precisão.

10.6 Considerações finais

A detecção de transtornos de saúde mental ou seus sintomas a partir de texto é uma tarefa computacional complexa e não totalmente resolvida do ponto de vista da pesquisa no PLN. Alguns **desafios**, exemplificados pelos projetos SetembroBR e Amive discutidos ao longo deste capítulo, são resumidos a seguir.

Em relação ao **quantidade e qualidade da anotação**, tem-se o alto esforço demandado pela anotação humana especializada, o que na prática pode impactar no número de

¹¹O sinal de “Agitação ou inquietação” não foi avaliado nos experimentos reportados em (Mendes; Caseli, 2024) por conter apenas 5 instâncias.



instâncias obtidas, como no caso do projeto Amive. A sensibilidade dos dados também demanda atenção extra no processo de anotação manual, considerando **questões éticas** como a não publicização dos dados e o cuidado para não afetar a saúde mental dos próprios anotadores. Como alternativa, tem-se a anotação a partir de autorrelatos ou usando LLMs. Nesses casos, contudo, a confiabilidade das anotações pode não ser tão alta e, no caso dos LLMs, há que se considerar os **custos** computacional e financeiro envolvidos.

Há também que se considerar a presença de **vieses** como o inerente nos textos de indivíduos da classe positiva (e.g. depressivos ou ansiosos), que após serem diagnosticados podem passar a falar mais sobre o assunto. Por outro lado, os textos das redes sociais também permitem que uma grande variedade de assuntos sejam tratados nas postagens, o que acarreta em **ruídos** em relação ao problema de pesquisa em foco que para ser tratado deveria ter como base de dados apenas textos representativos de questões de saúde mental. Desse modo, é importante observar que os fenômenos de interesse, e que seriam indicativos de saúde mental, podem não estar totalmente representados na rede social. A frequência e teor das postagens de um indivíduo podem variar drasticamente a depender de múltiplos fatores externos, e que não são relacionados ao seu estado de saúde mental.

A rede social é, portanto, um **retrato naturalmente incompleto**, que talvez traga fragmentos de informação útil para a tarefa, ou talvez nem isso. Como consequência, não é razoável supor que um modelo computacional que receba como entrada apenas as postagens de um indivíduo possa obter altos níveis de acurácia na detecção de um transtorno de saúde mental ou de seus sintomas. Assim, modelos deste tipo são melhor entendidos como ferramentas de apoio ao usuário da rede social, aos seus pais ou responsáveis, ou mesmo a um profissional da área de saúde.

Considerando-se a natureza específica da atividade em redes sociais, observa-se também que modelos preditivos de saúde mental precisam lidar com **grandes volumes** de publicações, ou seja, *timelines* longas. Sequências longas deste tipo são uma limitação natural para métodos baseados em *fine-tuning* de modelos pré-treinados como BERT, e este problema é agravado pelo fato de que os dados normalmente incluem um grande número de publicações que podem ser pouco relevantes para a tarefa.

Por fim, vale mencionar um grande desafio para as pesquisas nessa temática que é a **não disponibilização pública** de dados de saúde mental. *Corpora* textuais rotulados com informações sobre saúde mental geralmente não são recursos computacionais totalmente acessíveis, como seriam *corpora* voltados a outros tipos de aplicação de PLN.

Esforços para contornar algumas destas dificuldades incluem, mais recentemente, o uso de métodos baseados em *prompt* de grandes modelos de linguagem. Entretanto, estudos iniciais deste tipo, apesar de promissores, ainda não obtiveram resultados significativamente superiores ao *fine-tuning* de modelos BERT.

Agradecimentos

Esse capítulo contou com apoio FAPESP # 2021/08213-0 e # 2020/05157-9 (Projeto Amive). O primeiro autor também agradece ao Centro de Inteligência Artificial (C4AI-USP) e ao apoio da Fundação de Apoio à Pesquisa do Estado de São Paulo (processo FAPESP # 2019/07665-4) e da IBM Corporation. A segunda autora agradece os especialistas do Comitê de Especialistas do Amive: Jair Borges Barbosa Neto, Larissa Campagna Martini, Augusto Rozendo Mendes, Tais Bleicher, Heloísa Cristina Figueiredo Frizzo, Ana Teresa Delfino, Karina Antonialli e Elaine Nabeth Louzada Torres.



Capítulo 11

Classificação de Áudio aplicada à Saúde

Detecção de Problemas Respiratórios

Marcelo Matheus Gauy

Larissa Cristina Berti

Flaviane Romani Fernandes Svartman

Beatriz Raposo de Medeiros

Celso Ricardo Fernandes de Carvalho

Marcelo Gomes de Queiroz

Marcelo Finger

Publicado em: 16/04/2026

11.1 Introdução

Classificação de áudio é um tipo de tarefa de aprendizado de máquina que consiste em atribuir uma classe a cada um dos áudios de um conjunto de dados. Um exemplo clássico de conjunto de dados para esta tarefa é o AudioSet (Gemmeke et al., 2017). O AudioSet consiste em aproximadamente 5000 horas de áudios do Youtube distribuídos em 527 classes (anotadas). Alguns exemplos de classe presentes no AudioSet são fala, respiração, tosse, ruídos de barcos e aviões, música clássica, entre outros. Modelos de aprendizado de máquina e, em particular, de aprendizado profundo, são utilizados para processar o sinal de áudio, seja ele em forma de onda ou espectrograma (Brigham; Morrow, 1967) e classificá-lo em uma das 527 classes. Trata-se, portanto, de uma tarefa de classificação, via aprendizado supervisionado, clássica.

O sucesso das técnicas de aprendizado de máquina (especialmente às de aprendizado profundo) tornou possível que pesquisadores buscassem usar essas técnicas na classificação de áudios de pessoas com problemas de saúde que possam afetar sua voz e fala (Fagherazzi et al., 2021). Dentre estes, destacam-se especialmente as condições de saúde que afetam o sistema respiratório de alguma maneira (Landry et al., 2025). Este capítulo tratará prioritariamente de modelos e técnicas de classificação de áudio para a voz e fala de pacientes com problemas respiratórios. Algumas especificidades distinguem este tipo de problema de classificação de uma tarefa clássica de classificação de áudio. Essas especificidades são dificuldades típicas no uso de aprendizado de máquina no contexto da saúde. A principal delas é a dificuldade, tanto por motivos práticos quanto por motivos éticos, de coletar dados de pacientes em hospitais. Modelos complexos de classificação necessitam, normalmente, de grande volume de dados para atingirem desempenho satisfatório. Em geral, é um grande desafio coletar áudios de pacientes no volume demandado por modelos do estado da arte de aprendizado profundo. Além do volume de dados coletados, a diversidade do conjunto de dados coletados também é um limitante importante em muitos casos. Esta diversidade vai desde possíveis comorbidades em pacientes, a outros fatores como



variações da doença ao longo do tempo. Além disso, o tratamento de áudios em situações reais pode levar a grandes variações nos tipos de ruído presentes, o que pode fazer com que modelos de excelente performance no conjunto de dados original falhem uma vez que ocorram alterações nestes ruídos.

O problema de lidar com o baixo volume de dados é tratado, em geral, através de técnicas clássicas de transferência de aprendizado. Modelos de classificação de áudio são pré-treinados em grandes volumes de áudio, como o AudioSet e, posteriormente, refinados para o conjunto de dados de menor volume em questão (Huang et al., 2022; Kong et al., 2020; Niizumi et al., 2022). A Seção 11.3.2.1 apresentará uma visão detalhada desta e outras técnicas para lidar com conjuntos pequenos de dados. O problema da baixa diversidade de dados presentes no conjunto de áudios de pacientes é mais complexo. No momento, não existem técnicas satisfatórias para tratar deste problema de forma geral. Isto torna a maior parte dos modelos de aprendizado profundo aplicados à saúde pouco úteis, uma vez que foram retirados de seu (restrito) contexto original (Cohen et al., 2021). Uma discussão sobre esse problema e como potencialmente lidar com ele será feita na Seção 11.4.

Descreveremos nas próximas subseções alguns problemas clássicos de saúde que afetam a voz e fala e podem ser detectados com modelos de classificação de áudio. Neste capítulo, focaremos principalmente na língua portuguesa, embora muitas das técnicas presentes aqui não sejam específicas para o português.

11.1.1 Detecção de COVID-19 através da voz e tosse

Durante a pandemia, diversas iniciativas de pesquisa (em nível internacional) foram criadas com o objetivo de detectar COVID-19 de forma automática, com aprendizado de máquina, através da voz, tosse e sons da respiração de pacientes (Brown et al., 2020a; Despotovic et al., 2021; Erdoğan; Narin, 2021; Han et al., 2021; Hassan et al., 2020; Laguarta et al., 2020; Mouawad et al., 2021; Orlandic et al., 2021; Pinkas et al., 2020). Esses trabalhos apontaram a possibilidade de detectar COVID-19 através de sinais de áudio, atingindo entre 70% e 90% de acurácia na detecção da doença. A maioria dos conjuntos de dados utilizados foram construídos por contribuição colaborativa (*‘crowdsourcing’*). As técnicas de aprendizado de máquina variam desde modelos mais simples com *features* construídos manualmente até o uso de redes neurais recorrentes.

Os trabalhos citados acima têm foco principal na língua inglesa, uma vez que os áudios coletados são de falantes do inglês. Trabalhos similares para o português foram feitos no escopo do projeto SPIRA-BM (Finger et al., 2025), um projeto temático da FAPESP de número 23/00488-5 cujas informações podem ser encontradas em <https://bv.fapesp.br/p/t/auxilios/113158>. O SPIRA-BM é uma sequência de outro projeto FAPESP (número 20/06443-5) chamado SPIRA (Aluísio et al., 2022). O SPIRA-BM, bem como o projeto SPIRA, seu predecessor, possuem uma natureza ligeiramente diferente, uma vez que o foco do SPIRA-BM está na insuficiência respiratória (sintoma típico de COVID-19), e não na detecção da COVID-19 em si. Inclusive, o projeto SPIRA-BM busca detectar não apenas a insuficiência respiratória proveniente de COVID-19, mas também a proveniente de outras fontes. A próxima subseção tratará dos resultados obtidos pelo projeto até meados de 2025 para o caso de insuficiência respiratória.



11.1.2 Insuficiência Respiratória e seus efeitos na voz e fala

A insuficiência respiratória (IR) é um dos sintomas causados pela COVID-19 e a principal responsável pela hospitalização de pacientes de COVID-19. Por esse motivo, o projeto SPIRA reuniu médicos, fisioterapeutas, linguistas, fonoaudiólogos e cientistas da computação com o intuito de investigar a possibilidade de sua detecção automática através da voz e fala de pacientes. O projeto SPIRA-BM expandiu o escopo do projeto SPIRA original para além da IR proveniente de COVID-19, incluindo IR proveniente de outras fontes. No escopo desses projetos foram realizadas duas coletas de dados, uma na época da pandemia (*corpus* SPIRA que pode ser encontrado no [Github](#)) e outra um pouco depois das vacinas (*corpus* SPIRA 2.0 que não foi tornado público por questões de privacidade dos pacientes). Esses *corpora* consistem de sinais de áudio (voz e fala) de pacientes com IR e controles saudáveis e permitem a construção de modelos de aprendizado de máquina para a detecção da IR através do sinal de áudio. Mais detalhes sobre esses e outros *corpora* serão descritos na Seção 11.2. O projeto concluiu, através de análises dos áudios presentes, que os pacientes faziam pausas mais numerosas e mais longas que os controles, com uma distribuição, muitas vezes, em desacordo com o que seria esperado de um falante do português (Fernandes-Svartman et al., 2022). Em termos de parâmetros de voz, foi verificado que parâmetros relacionados à frequência fundamental são diferentes entre pacientes e controles (Berti et al., 2023).

Mais ainda, foram construídos diversos modelos de aprendizado profundo (Casanova et al., 2021b; Gauy et al., 2023; Gauy; Finger, 2021; Matheus Gauy et al., 2026) que buscam a detecção da IR nos *corpus* SPIRA e SPIRA 2.0. Em (Casanova et al., 2021b), os autores construíram uma rede convolucional, chamada de SpiraNet, que recebe o MFCC-grama¹. De fato, ambos MFCC-grama e espectrograma foram testados como entrada da rede convolucional construída. O desempenho da rede ao receber MFCC-grama foi substancialmente superior. Com o MFCC-grama, a rede atingiu aproximadamente 87% de acurácia no *corpus* SPIRA. Posteriormente, a rede convolucional foi substituída por um Transformer (Vaswani et al., 2017b) que recebia MFCC-grama, chamado de MFCC-Transformer (Gauy; Finger, 2021). Esta alteração da rede levou a uma performance de aproximadamente 96% no mesmo *corpus* SPIRA. Posteriormente, o grupo SPIRA buscou usar técnicas de transferência de aprendizado que permitissem aumentar a complexidade dos modelos usados e, também, permitissem o uso do espectrograma direto, sem passar pelo MFCC. Assim, os pesquisadores usaram redes convolucionais e redes baseadas no Transformer que foram pré-treinadas no AudioSet. Esses novos modelos atingiram performance acima de 98.5% (Matheus Gauy et al., 2026), tanto no caso de IR proveniente de COVID-19 (*corpus* SPIRA), quanto no caso em que a IR provém de diversas causas potenciais (*corpus* SPIRA 2.0). Uma observação importante é que os modelos não apresentam o mesmo desempenho caso sejam retirados de seu contexto original de treinamento, com os modelos treinados no *corpus* SPIRA, não obtendo um bom desempenho no SPIRA 2.0 e vice versa (Gauy et al., 2023). Apesar destas dificuldades, as técnicas propostas pelo projeto serão de grande valor quando treinadas em *corpus* mais diversos. Como exemplo, trabalhos do projeto SPIRA foram vencedores no desafio de detecção de COVID-19 ComParE 2021 (Casanova et al., 2021a). Além disso, usando técnicas de interpretabilidade de modelos de aprendizado profundo, foi possível verificar que os modelos do projeto

¹Em linhas gerais, o MFCC é o resultado do cálculo do espectro dos áudios, via FFT, duas vezes. Na primeira, obtemos o espectrograma. Uma segunda aplicação da FFT sobre o espectrograma, produz o MFCC-grama. Mais detalhes são apresentados na Seção 11.3.1.



SPIRA dão atenção principalmente à voz e fala dos pacientes e não para os ruídos de fundo presentes nos áudios (Silva et al., 2024).

Considerando o sucesso obtido na detecção de IR pelo projeto, bem como o sucesso na detecção do COVID-19, foi criado o projeto temático SPIRA-BM que busca aprofundar a compreensão dos modelos de IR, potencialmente construindo modelos capazes de adicionalmente detectar a causa da IR, e também amplia o projeto para considerar outros problemas respiratórios, como asma e tabagismo. Na próxima subseção, discutiremos um pouco sobre esses problemas.

11.1.3 Outros problemas respiratórios

Inspirado no sucesso da detecção de IR, o projeto SPIRA-BM propôs estudar outros problemas respiratórios e seus efeitos na voz e na fala. Além da insuficiência respiratória, o projeto SPIRA-BM tem outras duas linhas gerais propostas. São elas: asma e tabagismo. Neste capítulo apresentamos esses problemas, o método de coleta de dados, os objetivos e seus impactos. Os resultados são parte do projeto temático SPIRA-BM (veja no link <https://bv.fapesp.br/pt/auxilios/113158>).

Consideremos o caso da asma: trata-se de uma doença crônica das vias aéreas, caracterizada por limitação ao fluxo aéreo expiratório que está associado a sintomas respiratórios como sibilância, dispneia, aperto no peito e tosse que variam ao longo do tempo quanto à sua frequência, ocorrência e intensidade. Esses sintomas são, frequentemente, desencadeados por fatores como exercício, exposição a alérgenos e/ou irritantes, alterações climáticas ou infecções respiratórias virais. O Brasil tem a oitava maior prevalência de asma no mundo, atingindo 12% da população brasileira. Na fase de criança e adolescência, a asma é mais prevalente em meninos enquanto, na fase adulta, tem prevalência maior entre mulheres. Em termos das tarefas de classificação de áudio que podemos considerar estão: i) a classificação entre asma não controlada e asma controlada; ii) a detecção de disfunção de prega vocal; iii) a previsão da probabilidade de exacerbação da asma dentro de um período de dois dias (estudo longitudinal). Os dois primeiros são problemas clássicos de classificação (binária) de áudio, com os dados de pacientes com asma controlada ou não controlada e com ou sem disfunção da prega vocal sendo coletados pelo projeto. A terceira tarefa é um estudo longitudinal, em que os pacientes com asma serão acompanhados ao longo do tempo e realizarão coletas de dados periódicas, com o intuito de permitir uma análise do risco de exacerbação da asma em um futuro próximo. Tal quadro de exacerbação, muitas vezes, leva à hospitalização. Estas tarefas, bem como outras análogas que possam ser consideradas, possuem o potencial, de auxiliar médicos e enfermeiros no tratamento preventivo da asma, se antecipando, por exemplo, a uma possível exacerbação dos sintomas.

No caso do tabagismo, busca-se: i) classificar o sujeito entre fumante ou não fumante; ii) estimar o nível de carga tabagística (quantidade de cigarros por dia vezes o número de anos como fumante); iii) determinar o nível de monóxido de carbono exalado pelos pacientes através do sinal de áudio. As tarefas ii) e iii) podem ser formuladas como problemas de regressão e, se necessário, quantizadas em problemas de classificação. Tarefas de regressão tendem a ser mais difíceis que tarefas de classificação, em particular, para modelos de análise de áudio. Um exemplo desta dificuldade pode ser encontrado em um trabalho do projeto SPIRA-BM que busca detectar o nível de saturação de oxigênio dos pacientes com insuficiência respiratória através do sinal de áudio (Matheus Gauy et al., 2026). Segundo Karelitz et al. (2017), a medida do monóxido de carbono no ar exalado



(COex) é uma avaliação simples e de baixo custo, além de ser um método não invasivo e objetivo que avalia a exposição recente a cigarro ou fumaça de cigarro. É bastante útil para detectar fumantes diários e mostra correlação com o consumo diário de cigarros. Dessa forma, estamos propondo o desenvolvimento de um método rápido e barato que auxilie os médicos e enfermeiros em programas de cessação de tabagismo.

11.1.4 Organização do capítulo

A Seção 11.2 apresentará uma descrição dos diversos *corpora* existentes atualmente, bem como suas limitações. Incluiremos os *corpora* construídos no escopo do projeto SPIRA e diversos outros *corpora*. Será feita uma divisão entre os *corpora* para COVID-19 e IR. Também serão brevemente apresentadas algumas características dos dados em processo de coleta de asma e tabagismo. A Seção 11.3 descreverá as técnicas de classificação de áudios para problemas respiratórios usados em geral. Trataremos das estratégias usuais de pré-processamento dos áudios e dos modelos de aprendizado profundo usados, descrevendo suas arquiteturas e técnicas para tratar de conjuntos pequenos de dados. A Seção 11.4 apresentará os potenciais impactos práticos do desenvolvimento de modelos de aprendizado de máquina para a detecção de problemas respiratórios e das limitações existentes nos modelos atuais e possíveis estratégias para o seu aprimoramento.

11.2 Corpora de áudios

Nesta seção, descreveremos os *corpora* existentes para detecção de problemas respiratórios. Ressaltamos que não temos como objetivo fazer uma descrição completa, escolhendo apenas alguns dos *corpora* principais. Para revisões mais detalhadas dos *corpora* existentes indicamos, por exemplo, (Shuja et al., 2021) (para a COVID-19) e (Kapetanidis et al., 2024; Xia et al., 2022) (para uma descrição um pouco mais geral). Ainda nesta seção, descreveremos os *corpora*: SPIRA (também chamado de primeira fase do projeto SPIRA), SPIRA 2.0 (também chamado de segunda fase do projeto SPIRA) e alguns *corpora* focados na COVID-19. Por último, comentaremos brevemente sobre *corpora* em construção no escopo do projeto SPIRA-BM para a insuficiência respiratória, asma e tabagismo.

11.2.1 Corpus de detecção de insuficiência respiratória

Os dois principais *corpora* já coletados no escopo do projeto SPIRA para a detecção de insuficiência respiratória são o SPIRA e o SPIRA 2.0.

11.2.1.1 Corpus SPIRA

Este *corpus* foi coletado pelo projeto SPIRA durante a pandemia da COVID-19 e consiste de dados de pacientes com insuficiência respiratória (causada pela COVID-19) e controles saudáveis. Os dados dos pacientes de insuficiência respiratória foram coletados em dois hospitais universitários da cidade de São Paulo. Os dados dos controles foram coletados através de um aplicativo na internet para a coleta dos áudios. A coleta consistiu em 4 áudios de cada paciente ou controle. O primeiro foi o consentimento do falante para o estudo. Usualmente, esse tipo de consentimento é feito por escrito, mas excepcionalmente na pandemia, foi permitido o consentimento por fala. O segundo áudio coletado foi uma frase elaborada por linguistas, para ser longa o bastante para conter pausas naturais, mas



simples o bastante para evitar que dificuldades de leitura levassem os pacientes a ter dificuldades de locução. A frase proposta foi: ‘O amor ao próximo ajuda a enfrentar o coronavírus com a força que a gente precisa’. A função do segundo áudio é verificar se os pacientes com insuficiência respiratória necessitam fazer pausas adicionais, não previstas gramaticalmente em português brasileiro, para recuperar o ar, durante a fala. O terceiro áudio consiste em uma parlenda conhecida pelos falantes brasileiros como ‘batatinha quando nasce...’. O objetivo desse áudio é averiguar propriedades da fala quando as pausas estão predeterminadas. O último áudio consiste em cantar ‘Parabéns para você’.

Uma observação importante é que o projeto SPIRA coletou áudios de pacientes e controles de ambientes distintos. Em particular, os ruídos de fundo presentes nos áudios foram bastante diferentes e muito pronunciados no caso dos pacientes (pois foram gravados em hospitais). Para evitar que esses ruídos introduzissem vieses nos modelos de aprendizado de máquina, foram coletados alguns áudios adicionais no início de cada coleta nos hospitais. Esses áudios contêm apenas o ruído ambiente presente, sem a voz ou fala de nenhum paciente. O intuito dessa coleta adicional foi permitir que os ruídos ambientes fossem adicionados aos áudios de controle, dificultando a tarefa da rede em classificar os áudios baseada apenas no ruído presente. A Seção 11.3.1 tratará de maneiras de lidar com os diferentes tipos de ruído presentes nos áudios de pacientes e controles.

Ao todo, foram coletados mais de 6000 áudios de controle e 536 áudios de pacientes de insuficiência respiratória. Foi realizada uma triagem dos áudios de pacientes, para evitar, principalmente, que áudios contendo voz e fala dos coletores fossem mantidos no conjunto final. A triagem também foi feita, no caso dos estudos (Berti et al., 2023; Fernandes-Svartman et al., 2022) para excluir os dados em que fossem detectadas pausas ocasionadas por problemas de letramento. Como o número de controles total é muito maior que o de pacientes, também é feito um balanceamento do conjunto de dados, selecionando os controles (e pacientes) de forma a manter distribuições similares de homens e mulheres entre ambos². Para a construção de modelos de aprendizado de máquina, os áudios remanescentes foram divididos em conjuntos de treino, validação e teste. A distribuição final dos dados está expressa na Tabela 11.1. O *corpus* original pode ser encontrado no [Github](#). O conjunto de dados de controle completo (sem a triagem - mais de 6000 áudios e aproximadamente 18 horas de Português Brasileiro) pode ser encontrado no [Zenodo](#).

Tabela 11.1: *Corpus* SPIRA

Dados		Treino	Validação	Teste
Controle	Masculino	59	8	22
	Feminino	84	8	26
	Duração média (s)	8.15	7.75	7.77
Pacientes	Masculino	83	8	28
	Feminino	66	8	32
	Duração média (s)	13.18	10.78	9.43
Áudios totais		292	32	108
Duração total (s)		3110	296	983

Fonte: (Casanova et al., 2021b)

²Segundo os autores, foi observado também um desbalanceamento das faixas etárias dos pacientes e controles, mas o tamanho do conjunto de dados tornou o balanceamento desse fator impraticável.



11.2.1.2 **Corpus SPIRA 2.0**

Este *corpus* possui uma estrutura similar ao anterior. A principal diferença é que ele foi coletado após o auge da pandemia e posterior às doses principais de vacinação. Assim, os dados de insuficiência respiratória não são dominados por pacientes de COVID-19, sendo provenientes de diversas causas. Outra diferença crucial é que tanto os dados dos controles quanto pacientes foram coletados em hospitais. Os dados foram coletados em 4 hospitais do estado de São Paulo. São eles: a Beneficência Portuguesa, o Hospital da Unimar, ambos em São Paulo, e a Santa Casa de Marília (SC) e CEES-Marília em Marília. Esta diferença torna o tratamento de ruído bem menos difícil que no *corpus* SPIRA. Apesar disso, os ruídos ambiente dos hospitais foram novamente coletados. Quanto aos outros áudios coletados de cada falante, novamente, o consentimento do estudo pelo falante foi em forma de áudio. A frase foi ligeiramente alterada para uma similar, que permitisse comparação posterior: ‘O amor ao próximo ajuda a enfrentar *essa fase* com a força que a gente precisa’. A parlenda (com pausas predeterminadas) foi mantida. A canção ‘Parabéns para você’ foi substituída por uma vogal sustentada (vogal a, reproduzida por alguns segundos). A função da vogal sustentada é avaliar parâmetros de voz dos falantes, em oposição aos parâmetros de fala³.

Uma desvantagem desse *corpus* sobre o anterior é seu tamanho. Por ter sido coletado após o auge da pandemia, a quantidade de pacientes nos hospitais que sofrem de insuficiência respiratória é bastante menor que durante a pandemia. Uma vantagem, porém, é que a insuficiência respiratória presente provém de diversas causas, não estando restrita apenas à COVID-19. Outra vantagem é que tanto pacientes e quanto controles foram coletados nos mesmos ambientes. Esse fato torna o tratamento do ruído bem mais simples.

Ao todo, foram coletados 116 áudios de controle⁴ e 26 dados de pacientes com insuficiência respiratória de causas diversas. A Tabela 11.2 apresenta a distribuição dos dados de pacientes e controles por gênero usada em (Gauy et al., 2023). Novamente, existe um desbalanceamento. Como se trata de poucos dados, o balanceamento não é possível. Em (Gauy et al., 2023; Matheus Gauy et al., 2026), foi feito apenas o balanceamento dos conjuntos de validação e teste, mantendo o conjunto de treino com todos os dados remanescentes. Esse conjunto de dados não é público, como diversos outros conjuntos de dados de saúde. Observe que o *corpus* SPIRA 2.0 possui a limitação de ser excessivamente pequeno, não chegando a 20 minutos de duração total.

Tabela 11.2: *Corpus* SPIRA 2.0

Dados		
Controle	Masculino	36
	Feminino	80
	Duração média (s)	7.41
Pacientes	Masculino	14
	Feminino	12
	Duração média (s)	8.14
Áudios totais		144

³Aqui, temos uma distinção entre parâmetros de voz, que correspondem às características acústicas da elocução humana, e parâmetros de fala, que envolvem a produção de linguagem natural.

⁴Segundo os autores, foram removidos do conjunto de dados, aqueles pacientes que apresentaram sintomas de COVID longa, uma vez que eles poderiam enviesar os resultados (Robotti et al., 2021).

Fonte: (Gauy et al., 2023)

11.2.1.3 Limitações atuais dos *corpus* de insuficiência respiratória

Além das limitações específicas já citadas em cada um dos *corpora*, temos limitações gerais, que estão presentes em diversos *corpora* de saúde. Estas limitações são relativas a baixa diversidade do *corpus* de áudios coletado, em parte pelo número total de áudios, em parte pela limitada distribuição espacial e temporal da coleta realizada. Os áudios coletados estão concentrados em São Paulo, e em um período bem específico de tempo, não contemplando, por exemplo, variações ao longo do tempo (ou de localização geográfica) na distribuição das doenças. Isto impede que modelos *end-to-end* de aprendizado de máquina, eficazes na prática, sejam construídos. Um exemplo desta dificuldade pode ser observada pelos dois *corpora* coletados. Modelos treinados no *corpus* SPIRA apresentam um desempenho pior que uma previsão aleatória, se testados no SPIRA 2.0 e vice-versa (Gauy et al., 2023).

11.2.2 *Corpus* de detecção de COVID-19

Durante a pandemia, diversos pesquisadores no mundo coletaram dados de voz, fala e tosse de pacientes para a detecção de COVID-19. Enquanto a maior parte da pesquisa no Brasil focou na tarefa de detecção de insuficiência respiratória (dados apresentados na Seção 11.2.1), pesquisadores no resto do mundo focaram principalmente na detecção de COVID-19 diretamente. Diversos *corpora* foram construídos por meio de coleta colaborativa (crowdsourcing) e foram apresentados em (Bhattacharya et al., 2023; Brown et al., 2020a; Despotovic et al., 2021; Orlandic et al., 2021). Embora existam problemas inerentes à coleta de dados de forma colaborativa (Garcia-Molina et al., 2016), essa é uma forma bastante eficaz de construir *corpus* de volume maior. Neste capítulo, apresentaremos dois *corpora* principais: o Coswara (Bhattacharya et al., 2023) (língua inglesa) e o ComParE (Schuller et al., 2021) (multilingual). Outros *corpora* foram construídos e podem ser encontrados nos artigos originais (Erdoğan; Narin, 2021; Han et al., 2021; Hassan et al., 2020; Laguarta et al., 2020; Mouawad et al., 2021; Pinkas et al., 2020; Watase et al., 2023).

11.2.2.1 *Corpus* Coswara

Este *corpus* foi construído durante a pandemia, entre abril de 2020 e fevereiro de 2022 (Bhattacharya et al., 2023). Ao todo, foram coletados áudios de 2635 indivíduos, dos quais 1819 são negativos para o Coronavírus, 674 são positivos para o Coronavírus e 142 contraíram e se recuperaram da COVID-19. Os áudios foram gravados através de um aplicativo online (<https://coswara.iisc.ac.in/>). Os participantes informaram o seu consentimento para a coleta de dados e responderam um questionário contendo informações demográficas (incluindo idade, gênero, localização, se são fumantes ou não e se a gravação foi feita com máscara), estado de saúde atual (incluindo sintomas como tosse, febre, diarreia, dificuldades de respiração, entre outros, e comorbidades como pneumonia e asma) e teste de COVID-19. Após responder o questionário, os participantes gravaram nove áudios de diferentes tipos. Dois deles tratam da respiração. Um é superficial, definido como alguns ciclos de respiração rápida sem forçar os pulmões. O outro é profundo, um tipo de respiração mais lenta que trabalha mais os pulmões. Os dois áudios para a tosse envolvem uma tosse superficial induzida, que não força o



pulmão, e uma tosse profunda, que envolve mais esforço pulmonar. Três tipos de vogais sustentadas foram gravados, com as vogais (como em *boot*), (como em *beet*) e (como em *bat*), sendo escolhidas. Para a fala, foi solicitado que os participantes contassem de 1 a 20, em velocidade normal (primeiro) e rápida (segundo). Assim, foram totalizados nove áudios de cada paciente. O tempo de coleta médio por participante foi 7 minutos. O total de horas gravadas foi de aproximadamente 65. Importante observar que uma fração significativa ($> 95\%$) dos participantes que reportaram teste positivo de COVID-19 foi gravada em hospitais e centros de saúde, tornando estas anotações de alta qualidade.

Os áudios coletados foram escutados por humanos para serem manualmente anotados entre qualidade excelente (nenhum ruído ambiente), moderados (pouco ruído ambiente) e ruim (ruído ambiente significativo). Aproximadamente 78% dos áudios foram considerados excelentes, com 12% considerados moderados e 10% considerados ruins. Os dados originais estão disponibilizados no [Zenodo](https://zenodo.org/record/5444441).

11.2.2.2 Corpus ComParE COVID-19

O *COMputational PARalinguistics challenge* de 2021 (COMPARE 2021) envolveu 4 tarefas, duas delas focadas na detecção de COVID-19. A primeira focou em dados de tosse e a segunda foi direcionada para a fala. A construção do *corpus* é similar ao Coswara e foi feita nos trabalhos (Brown et al., 2020a; Han et al., 2021). Novamente, foi usada a técnica de coleta colaborativa. Os dados foram coletados através de um aplicativo que podia ser acessado por diversas plataformas (móvel ou internet). Os participantes responderam um questionário com informações demográficas, médicas e o sintomas reportados. Para o desafio ComParE, foram incluídos apenas os áudios de tosse e fala dos participantes com a informação do teste de COVID-19. Para a primeira tarefa, focada em tosse, foram gravados 725 áudios de 343 participantes, totalizando uma hora e meia. Cada participante gravou entre uma e três tosses forçadas. Para a segunda tarefa, focada na fala, foram gravados 893 áudios de 366 participantes, totalizando pouco mais de três horas. Cada participante gravou de um a três áudios recitando a frase: “*I hope my data can help to manage the virus pandemic*” em um idioma (inglês, alemão, italiano, etc). O *corpus* ComParE não inclui áudios de português.

Para as tarefas de detecção de COVID no desafio ComParE 2021 foi proposta uma divisão entre treino, validação e teste. A Tabela 11.3 apresenta esta divisão. É importante observar que cada participante pode aparecer mais de uma vez nos conjuntos então a divisão deve ser respeitada.

Tabela 11.3: *Corpus* ComParE

Dados		Treino	Validação	Teste
Tosse	Covid positivo	71	48	39
	Covid negativo	215	183	169
	Total	286	231	208
Fala	Covid positivo	72	142	94
	Covid negativo	243	153	189
	Total	315	295	283

Fonte: (Schuller et al., 2021)



11.2.3 Corpora para detecção de outros problemas respiratórios

Trataremos brevemente de *corpora* para a detecção de problemas respiratórios como asma e tabagismo. Na literatura internacional, é comum usar o *corpus* Coswara para a detecção de asma (Looi et al., 2024). Além do Coswara, temos *corpora* privados construídos para detecção de asma (Balamurali et al., 2021; Xu et al., 2025) e infecções das vias aéreas (Balamurali et al., 2021).

No âmbito do Português brasileiro, o projeto SPIRA-BM estuda as condições de insuficiência respiratória, asma e tabagismo, como mencionado anteriormente. O *corpus* coletado para as três condições será similar ao SPIRA 2.0. Adicionalmente, são coletadas comorbidades de pacientes, bem como diversas informações demográficas, como idade, gênero. Também são coletados fatores como frequência respiratória, saturação do oxigênio no sangue (usualmente, via oxímetro). No caso de asma, os participantes respondem um questionário típico para avaliar a situação atual da doença. No caso do tabagismo, é coletado o nível de monóxido de carbono exalado. Para mais informações, veja (Finger et al., 2025).

11.3 Métodos de detecção de problemas respiratórios através de áudios

Nesta seção, apresentaremos as principais técnicas usadas na construção de modelos de aprendizado de máquina para a detecção de problemas respiratórios através do sinal de áudio (voz e fala) de pacientes. Na Seção 11.3.1, discorreremos sobre as etapas de pré-processamento necessárias para evitar a introdução de vieses no modelo (como tratamento de ruído, uniformização da duração dos áudios, divisão em treino, validação e teste) e para preparação de *features* que podem ser passados aos modelos (como o espectrograma e MFCC-grama). Na Seção 11.3.2, apresentaremos os principais modelos de classificação de áudio existentes, destacando as técnicas usadas para tratar com poucos dados (Seção 11.3.2.1) e as arquiteturas usualmente consideradas baseadas em redes convolucionais e redes do tipo Transformer (Seção 11.3.2.2).

11.3.1 Pré-processamento de áudios

Para o treinamento de modelos de aprendizado de máquina em geral, e, em particular, no caso de áudio, é essencial um pré-processamento adequado dos dados. Em termos das *features* inseridas nos modelos, a maioria dos modelos (em especial, os de aprendizado profundo) realiza um treinamento *end-to-end*, ou seja, o modelo recebe como entrada o áudio ‘cru’ e tem como saída a classe daquele áudio recebido. Mesmo assim, existem variações nos formatos de entrada dos áudios que são passados para os modelos. Um meio usado por alguns modelos é receber o áudio em forma de onda. Um exemplo de modelo com essa característica é o Wav2Vec e suas variantes (Baevski et al., 2020). Como alternativa à forma de onda, podemos passar ao modelo, como entrada, o espectrograma do áudio. O espectrograma é obtido através de uma transformada (rápida) de Fourier (Brigham; Morrow, 1967) que para cada janela de tempo, calcula um vetor de frequências relativas ao áudio. Ao realizar este cálculo para cada uma das janelas, obtemos uma representação do áudio no espaço de frequências, chamada de espectrograma. O espectrograma se assemelha a uma imagem (bidimensional) e os modelos de classificação de áudio que usam espectrograma como entrada são muito similares a modelos de processamento de



imagem (Baade et al., 2022; He et al., 2022; Huang et al., 2022; Kong et al., 2020). Adicionalmente, podemos fazer uma transformada de Fourier adicional sobre o espectrograma para obter o ‘espectro’ do espectrograma, chamado de MFCC-grama. O MFCC-grama é muito usado como entrada para os modelos de classificação de áudio (Gauy; Finger, 2021; Gourisaria et al., 2024; Vimal et al., 2021), apresentando muitas vezes desempenho superior ao espectrograma, em particular, nos casos em que o tamanho do *corpus* de áudios analisado seja pequeno.

Além do processamento do áudio para tornar a entrada compatível ao modelo de classificação, seja para espectrograma, MFCC-grama ou mesmo em forma de onda, devemos evitar que vieses sejam introduzidos nos modelos por características do processo de coleta de dados usado (entre outros motivos). Temos três fatores principais que introduzem viés nos áudios: ruído, duração dos áudios e distribuição demográfica. Eles serão descritos nas próximas subseções.

11.3.1.1 Formas de viés: ruído

O ruído é uma das principais fontes de viés. Em particular, como os *corpora* são usualmente pequenos e coletados em poucos locais, é bem possível que eles possuam ruído característico daqueles locais de coleta, o que provavelmente vai gerar vieses nos modelos de classificação. Um exemplo claro em que o ruído é determinante é o *corpus* SPIRA apresentado na Seção 11.2.1.1. No *corpus* SPIRA, os dados de controle e pacientes foram coletados em ambientes distintos e com ruídos claramente distintos. Assim, um modelo treinado diretamente naquele *corpus* para distinguir pacientes de controles, provavelmente, não faria mais do que classificar o tipo de ruído presente (Casanova et al., 2021b). Há duas técnicas gerais para tratar a presença de ruído. A primeira é eliminar o ruído presente nos áudios, por exemplo, através de um filtro que elimina frequências abaixo de um certo limiar. Embora esta técnica seja muito eficaz em diversos problemas, para a detecção de problemas respiratórios, ela pode resultar na remoção de parte do sinal relevante para a classificação dos áudios (como a respiração do paciente). Assim, uma técnica alternativa pode ser considerada: coletar o ruído presente nos ambientes e artificialmente acrescentá-lo nos áudios antes do modelo recebê-los. Esta foi a técnica usada no *corpus* SPIRA (Casanova et al., 2021b; Gauy; Finger, 2021). Observe que acrescentar ruído no áudios diminui a relação sinal-ruído. Assim, determinar a técnica mais adequada depende da realização de experimentos empíricos com ambas as técnicas.

11.3.1.2 Formas de viés: duração

O ruído não é a única forma de viés comumente presente em *corpus* de áudio para detecção de problemas respiratórios. Também é muito comum que vieses existam na própria duração dos áudios. Por exemplo, no caso de insuficiência respiratória, pacientes usualmente gravam áudios mais longos que controles, visto que fazem pausas (para respiração, provavelmente) mais longas e mais numerosas (Fernandes-Svartman et al., 2022). Assim, se os modelos de classificação receberem os áudios com a duração total, eles tenderão a focar excessivamente na duração dos áudios e não no conteúdo em si. Uma solução simples e eficaz para esse problema é quebrar os áudios em trechos de mesmo tamanho. Inclusive, podemos usar janelas de um tamanho fixo (por exemplo 4 segundos) com um salto menor (por exemplo, de 1 segundo), de forma que um áudio de 8 segundos se transforme em 5 áudios de 4 segundos. Este método é chamado de janelamento (Casanova et al., 2021b; Gauy et al., 2023; Gauy; Finger, 2021) e, além de prevenir o viés pela duração dos áudios,



é uma técnica simples e eficaz de aumento de dados. Em particular, ao lidar com conjuntos pequenos de dados, técnicas de aumentação são bastante importantes. Outra técnica de aumentação de dados para o caso de sinal de áudio é a inserção de ruído de diversos tipos nos áudios. Uma revisão da literatura sobre as muitas técnicas de aumentação de dados para problemas de classificação de áudio pode ser encontrada em (Ferreira-Paiva et al., 2022).

11.3.1.3 Formas de viés: distribuição demográfica

Além do ruído e da duração dos áudios, existe uma outra fonte bastante comum de vieses em modelos de classificação de áudio para detecção de problemas respiratórios. Essa fonte corresponde a diferentes características demográficas (por exemplo, sexo e idade) entre as diferentes classes. Por exemplo, podemos ter um *corpus* em que o número de pessoas do sexo feminino com a doença (asma, por exemplo) seja maior que o número de pessoas do sexo masculino. No conjunto de controle, porém, a relação se inverte. Nestes casos, os modelos de classificação teriam foco excessivo no sexo do paciente ao fazer a classificação. A solução para este viés é realizar uma divisão balanceada do *corpus* em conjuntos de treino, validação e teste, de forma que (pelo menos) a proporção de pacientes e controles nos diferentes grupos demográficos (especialmente gênero) seja similar em cada um dos conjuntos (de treino, validação e teste). A divisão em treino, validação e teste deve ser cuidadosa ao considerar técnicas como janelamento, por exemplo. Deve-se evitar, a todo custo, que áudios do mesmo participante caiam em conjuntos diferentes. Em geral, o conjunto de teste deve ser grande o bastante para permitir que os resultados medidos a partir dele sejam representativos, e o mais independente possível dos outros conjuntos. Sobre o conjunto de treino, existem menos restrições, e podemos usar diversas técnicas de aumentação de dados, por exemplo, para tentar melhorar os resultados na validação e, consequentemente, no teste.

11.3.1.4 Capacidade de generalização dos modelos

Por fim, gostaríamos de ponderar que a construção de modelos de aprendizado de máquina sobre conjuntos pequenos de dados é inerentemente problemática. É muito comum que modelos apresentem desempenho excelente no conjunto de teste mas que não generalizem na prática, quando testados em dados novos coletados por outros meios. Isto tende a ocorrer independentemente do cuidado ao pré-processar os dados e é uma característica de modelos de aprendizado de máquina, em especial, dos modelos de aprendizado profundo, por se tratarem de modelos caixa preta que encontram o caminho mais fácil para aprender a tarefa para a qual foram treinados (Cohen et al., 2021). Um exemplo desse tipo de problema, para o caso de insuficiência respiratória, é o trabalho (Gauy et al., 2023). Nesse trabalho, modelos treinados no *corpus* SPIRA foram testados no *corpus* SPIRA 2.0 e vice versa. Em nenhum dos casos, os modelos foram capazes de apresentar generalizações de forma eficaz. A Seção 11.4 irá expandir esta discussão, apresentando os impactos para a sociedade dos modelos construídos e como, potencialmente, recuperar a capacidade de generalização dos modelos.

11.3.2 Modelos de aprendizado profundo para classificação de áudios

Nesta seção, apresentaremos os principais modelos de aprendizado profundo para classificação de áudio, usados no contexto de detecção de problemas respiratórios. Na Se-



ção 11.3.2.1, discutiremos sobre os cuidados necessários ao serem considerados conjuntos pequenos de dados. Trataremos principalmente da técnica de realizar um pré-treinamento sobre o modelo em um conjunto de áudios relacionado e de grande volume, seguido de um rápido refinamento na tarefa em questão. Também discutiremos brevemente técnicas de aumento de dados e regularização dos modelos. Na Seção 11.3.2.2, apresentaremos as arquiteturas comumente usadas pelos modelos, como as redes convolucionais e redes do tipo Transformer. Discutiremos as características das redes profundas de classificação de áudio, o tipo de entrada que recebem e a forma como são treinadas.

11.3.2.1 Tratamento de conjuntos pequenos de dados

Modelos do estado da arte de aprendizado profundo necessitam de grandes volumes de dados para seu treinamento. Quando tratamos de problemas de saúde, porém, atingir um grande volume de dados é extremamente difícil, visto que a coleta de dados é bastante custosa. Os *corpora* apresentados na Seção 11.2, em geral, não chegam a 100 horas de áudio, mesmo quando usam coleta colaborativa (*crowdsourcing*) e não chegam a 10 horas, quando coletados diretamente em hospitais. O contraste com conjuntos de dados gerais é nítido. O AudioSet, por exemplo, possui 5000 horas de áudios anotados em 527 classes. O *corpus* CommonVoice (Ardila et al., 2020), possui mais de 3000 horas, apenas de inglês, e mais de 30000 horas, incluindo mais de 100 idiomas.

A discrepância observada acima, sugere o uso de técnicas de transferência de aprendizado para permitir o uso de modelos de aprendizado profundo de complexidade maior. No contexto de processamento de linguagens naturais, o conhecido modelo BERT (Devlin et al., 2019) usou de uma técnica de pré-treinamento não supervisionado chamada de *masked language modelling*, que consistia em apagar uma proporção (15%) das palavras em uma frase para (pré-)treinar um modelo de forma não supervisionada para reconstruir a parte apagada. Posteriormente, o modelo foi refinado de maneira rápida e eficaz, mesmo em conjuntos com pequeno volume de dados, para outras tarefas de forma supervisionada. Esse método foi adaptado para o contexto de imagens e áudio alguns anos depois (Baade et al., 2022; He et al., 2022; Huang et al., 2022; Liu et al., 2020a, 2020b). O espectrograma dos áudios (ou as imagens) é visto como blocos de pixels (16×16 por exemplo) e aproximadamente 70% dos blocos são apagados. O modelo é treinado para reconstruir a parte apagada. Note que, como áudios e imagens possuem muito mais informação local, é necessário apagar uma fração muito maior dos mesmos de forma a induzir os modelos a compreenderem a estrutura global (e não apenas local) dos dados. Para permitir esta alteração, é necessário usar um modelo mais complexo (baseado na rede Transformer), tanto para a codificação (*Encoder*) quanto para a decodificação (*Decoder*) (Baade et al., 2022; He et al., 2022; Huang et al., 2022). Além do pré-treinamento não supervisionado, podemos também realizar um pré-treinamento supervisionado para classificação de áudio como feito em (Kong et al., 2020). Neste trabalho, as redes foram pré-treinadas de forma supervisionada no AudioSet e depois refinadas para outras tarefas, muitas vezes com poucos áudios. Qualquer que seja o método de pré-treinamento usado, a ideia é que a transferência de aprendizado permita que modelos complexos de aprendizado profundo sejam usados mesmo em conjuntos com pequeno volume de dados.

Outro cuidado que pode ser benéfico para o tratamento de *corpora* pequenos é a aumento dos dados. Técnicas de aumento dos dados ajudam a evitar que modelos se especializem no conjunto de treino sem serem eficazes para fora dele (Ferreira-Paiva et al., 2022). Isto é chamado de *overfitting* e costuma ocorrer com modelos mais complexos.



No caso de áudio, a inserção de ruído é uma das principais técnicas de aumento dos dados. Ela pode ser feita de diversas formas, como adicionar ruído gaussiano (branco), ou adicionar ruído de fundo comuns e não relacionados à tarefa em questão. Outra técnica é qualquer forma de segmentação dos áudios em áudios menores, como o janelamento, a remoção de partes silenciosas no começo e no final ou a translação do áudio por um fator aleatório. Também podemos alterar a velocidade de reprodução do áudio, a escala melódica dos áudios e ajustar o volume. Diversos tipos de filtros podem ser aplicados, para remover frequências mais baixas ou altas por exemplo. Alternativamente, podemos mascarar algumas frequências ou trechos do áudio aleatoriamente. Uma revisão das técnicas de aumento de dados pode ser encontrada em (Ferreira-Paiva et al., 2022).

Além das técnicas acima, existem dois outros métodos gerais para tratar de conjuntos com pequeno volume de dados. O primeiro consiste em aplicar algum tipo de regularização ao modelo. Uma técnica de regularização comum é o acréscimo na função de perda de um termo que controle o tamanho dos pesos da rede, segundo alguma norma como a euclidiana. Outra técnica de regularização é o uso do *dropout* (Srivastava et al., 2014). Esta consiste em remover uma proporção aleatória dos neurônios artificiais de uma camada da rede, forçando a mesma a construir um modelo mais robusto. O segundo método, em muitos casos imprescindível, é usar redes menos complexas. Redes de maior complexidade, com maior quantidade de parâmetros, possuem uma tendência mais forte ao *overfitting*. Uma solução é diminuir a complexidade do modelo construído, reduzindo o número de parâmetros, por exemplo, e evitando assim que a rede memorize o conjunto de treino.

Existe um problema comumente presente em conjuntos com pequeno volume de dados que não possui solução até o momento. Trata-se da tendência de que a distribuição dos dados do *corpus* seja diferente da que será encontrada na prática no momento de aplicação do modelo (Cohen et al., 2021). Este problema é bastante comum para problemas de saúde, visto que a população e as doenças passam por transformações ao longo do tempo. Além disso, a distribuição das doenças é diferente em localizações diferentes e conjuntos com pequeno volume de dados normalmente foram coletados em poucos locais. Na Seção 11.4, discutiremos as implicações dessa dificuldade prática e como poderíamos tentar contorná-la.

11.3.2.2 Arquiteturas

Em termos de arquiteturas comumente usadas em modelos de aprendizado profundo para detecção de problemas respiratórios, as redes convolucionais e as redes baseadas no Transformer são as mais comuns. As redes recorrentes tiveram diminuída a sua relevância com o advento do Transformer, assim como ocorreu em outros domínios. Em geral, os áudios são passados para as redes em forma de onda, ou mais comumente, como espectrograma. Em alguns casos, é usado o MFCC-grama. Tanto espectrograma quanto MFCC-grama são entradas parecidas com imagens, de forma que os métodos para análise de áudio usados se assemelham aos métodos usados para imagens.

11.3.2.3 Redes convolucionais: *Pretrained Audio Neural Networks*

Dentre as arquiteturas que usam redes convolucionais, as *Pretrained Audio Neural Networks* ou PANNs, propostas em (Kong et al., 2020), se destacam. Mais precisamente, as PANNs são redes pré-treinadas de forma supervisionada em grandes volumes de áudios (usualmente no AudioSet) e não necessariamente tem uma arquitetura definida. No caso



de modelos de detecção de problemas respiratórios as PANNs baseadas em redes convolucionais CNN6, CNN10 e CNN14 foram bastante utilizadas. Além de serem pré-treinadas, elas possuem a vantagem de serem 3 modelos similares de redes convolucionais com grau crescente de complexidade. Isto permite evidenciar o efeito do baixo volume de dados presente nos *corpora* considerados. A CNN6 é uma rede convolucional com 6 camadas, enquanto a CNN10 possui 10 camadas e a CNN14 possui 14 camadas. Todas as três recebem o espectrograma dos áudios como entrada. As camadas convolucionais da CNN6 usam kernels 5×5 , enquanto que a CNN10 e CNN14 usam kernels 3×3 . Cada uma dessas camadas é seguida de uma camada de *Batch normalization* (Ioffe; Szegedy, 2015) e a não linearidade escolhida é a *ReLU* (Nair; Hinton, 2010). Quatro dessas camadas estão presentes na CNN6 e entre duas delas uma camada de *average pooling* (Kong et al., 2019) está presente. Na CNN10 e CNN14, dois blocos convolucionais são aplicados em sequência antes do *average pooling* ser realizado. A CNN10 possui 4 pares desses blocos, enquanto a CNN14 tem 6 pares desses blocos. No último bloco convolucional, é usada uma camada de *Global pooling* (soma do *average* e *max pooling*) no lugar do *average pooling*. Uma penúltima camada totalmente conectada, intermediária entre os blocos convolucionais e a saída sigmoide (também totalmente conectada) com as 527 classes, está presente. Para prevenir o *overfitting*, a técnica de *dropout* é aplicada após cada camada. A CNN6 tem 4.8 milhões de parâmetros, a CNN10 tem 5.2 milhões e a CNN14 tem 80.7 milhões.

11.3.2.4 Redes convolucionais: SpiraNet

A SpiraNet, proposta em (Casanova et al., 2021b), também usa camadas convolucionais em sequência e recebe espectrograma ou MFCC-grama (mais eficaz para a detecção de insuficiência respiratória). Como a SpiraNet foi treinada do zero e não recebeu nenhum tipo de pré-treinamento, seu desempenho costuma ser inferior ao das PANNs. A tendência futura da área é o uso de modelos pré-treinados, de forma a facilitar o aprendizado da rede em *corpus* de poucas horas (ou mesmo minutos) de áudio.

11.3.2.5 Redes Transformer: *Masked Autoencoder*

Diversas arquiteturas recentes eficazes, baseadas no Transformer, foram construídas. Discutiremos principalmente um tipo de arquitetura chamada de *Masked Autoencoder* (Baade et al., 2022; He et al., 2022; Huang et al., 2022; Niizumi et al., 2022). Esta arquitetura consiste em duas redes do tipo Transformer, uma para codificação e outra para decodificação. A construção da rede tem por objetivo permitir o pré-treinamento não supervisionado sobre grandes volumes de áudios (ou imagens). O codificador recebe o espectrograma dos áudios como blocos (usualmente 16×16). O codificador é alguma variação do *Vision Transformer* (Dosovitskiy et al., 2021). O decodificador é um Transformer clássico que vai reconstruir o espectrograma original, a partir da saída do codificador. O pré-treinamento consiste em apagar uma proporção alta (70% por exemplo) do espectrograma do áudio original e transmitir apenas a parte não apagada para o codificador. Como apenas uma fração menor dos dados é transmitida ao codificador, ele usualmente possui mais parâmetros que o decodificador. O decodificador deve reconstruir o espectrograma original a partir da saída. Esta arquitetura permite que o modelo aprenda a estrutura global dos áudios de forma auto-supervisionada. Este modelo pré-treinado é um dos mais eficazes classificadores de áudio existentes atualmente. Naturalmente, para as tarefas de detecção de problemas respiratórios, esta arquitetura também costuma se sobressair.



11.3.2.6 Redes Transformer: MFCC-Transformer

O MFCC-Transformer, usado em (Gauy; Finger, 2021), também usa um Transformer como base, tratando os tokens do Transformer como blocos de 1 frame pelo número de faixas de frequência. Ele recebe o MFCC-grama o que o torna mais eficiente do que se recebesse o espectrograma. Embora MFCC-Transformer possua versões pré-treinadas no português brasileiro (Gauy; Finger, 2023), a técnica de pré-treinamento utilizada é menos eficaz (construída antes do surgimento do *Masked Autoencoder*) e o volume de áudios do português usados no pré-treinamento também é menor.

Um caminho potencial futuro é realizar um pré-treinamento adicional sobre um modelo pré-treinado. Esse pré-treinamento adicional poderia usar dados do português brasileiro, por exemplo, para preparar o modelo para melhor compreender o idioma (caso estejamos tratando de detectar problemas respiratórios de falantes do português). A técnica de pré-treinamento adicional foi proposta em (Niizumi et al., 2023, 2025), no modelo chamado *Masked Modeling Duo* ou M2D. Resumidamente, esse modelo usa duas redes em paralelo, uma que codifica os dados apagados e outra que codifica os dados não apagados e prevê o resultado codificado pelo outro modelo. A rede é treinada para maximizar o acordo entre os dois métodos. Esta técnica permite que o modelo seja pré-treinado de forma não supervisionada em um grande volume de dados e, posteriormente, que seja realizado um pré-treinamento adicional em uma tarefa mais próxima da tarefa alvo desejada. Em geral, esse pré-treinamento adicional leva a uma melhora de performance na tarefa final. No entanto, algumas dificuldades existem ao selecionar os parâmetros para o pré-treinamento adicional.

11.3.2.7 Outras arquiteturas

Além das redes apresentadas acima, existem diversos modelos de processamento de áudio cujo enfoque é o processamento de fala. Dentre essas, se destacam os modelos do tipo Wav2Vec (Baevski et al., 2020), que recebem o áudio em forma de onda e usam tanto camadas convolucionais (para extrair a informação do áudio), como camadas no estilo do Transformer. O Wav2Vec usa um quantizador para os áudios que permite extrair componentes discretas que compõe o mesmo (se assemelham a fones, ou seja, a sons da fala). Isto torna o modelo um transcritor de fala bastante potente. Outros modelos que têm foco no processamento de fala são o HuBERT (Hsu et al., 2021) e o XLS-R (Babu et al., 2021). Um tipo de modelo recente que foca em processamento multimodal e pode se tornar relevante para as tarefas de detecção de problemas respiratórios é o OmniVec (Srivastava; Sharma, 2024).

11.3.3 Avaliação dos modelos de detecção de problemas respiratórios

Em termos de avaliação dos modelos, algumas métricas, como a acurácia, são comumente usadas. Estas são métricas que costumam ser relevantes para qualquer problema de classificação. As seguintes métricas são comuns (VP corresponde a verdadeiro positivo, VN corresponde a verdadeiro negativo, FP corresponde a falso positivo e FN a falso negativo):

$$\text{Acurácia} = \frac{VP + VN}{VP + FP + VN + FN}$$

$$\text{Recall} = \frac{VP}{VP + FN}$$



$$\text{Precisão} = \frac{VP}{VP + FP}$$

$$\text{Especificidade} = \frac{VN}{VN + FP}$$

$$\text{G-médio} = \sqrt{\text{Especificidade} \times \text{Recall}}$$

$$\text{F-measure} = 2 * \frac{\text{Precisão} \times \text{Recall}}{\text{Precisão} + \text{Recall}}$$

$$\text{MCC} = \frac{VP \times VN - FP \times FN}{\sqrt{(VP + FP)(VP + FN)(VN + FP)(VN + FN)}}$$

Em geral, a acurácia é a métrica mais comum. Porém, muitos *corpora* possuem um desbalanceamento apresentando mais controles que pacientes. Isto torna medidas alternativas, que consideram a precisão, o *recall* e a especificidade, como o G-médio ou a *F-measure*, muito mais acuradas para a avaliação dos modelos. No caso de problemas de saúde, poderiam ser consideradas ainda variações dessas medidas que põem pesos maiores para falsos negativos do que falsos positivos por exemplo. Curvas ROC e a área abaixo dela também são medidas comumente usadas para problemas de classificação (Hanley; McNeil, 1982), e não é diferente no caso particular de problemas respiratórios.

A Tabela 11.4 apresenta um resumo geral dos resultados obtidos para os modelos de detecção de insuficiência respiratória no escopo do projeto SPIRA. A Tabela 11.4 foca nos dados em português brasileiro, não apresentando modelos treinados em dados de outros idiomas. Em termos de métricas, ela apresenta apenas a acurácia, a título de comparação entre os modelos. Outras métricas podem ser encontradas em (Matheus Gauy et al., 2026).

Tabela 11.4: Resumo de trabalhos para a detecção de insuficiência respiratória

Referência	Corpus	Features	Modelo	Acurácia	Fraqueza
(Casanova et al., 2021b)	SPIRA	MFCC	SpiraNet	87%	Varia significativamente dependendo do ruído
(Gauy; Finger, 2021)	SPIRA	MFCC	MFCC-Transformer	96.5%	Modelo não pré-treinado
(Gauy et al., 2023)	SPIRA	MFCC	pretrained MFCC-Transformers	97.4%	Técnica de auto-supervisionamento é ineficaz
(Gauy et al., 2023)	SPIRA 2.0	MFCC	pretrained MFCC-Transformers	95%	<i>Corpus</i> muito pequeno para uma análise completa
(Matheus Gauy et al., 2026)	SPIRA	Audio Spectrogram	Audio-MAE	99.9%	Dados de pré-treinamento apenas em inglês
(Matheus Gauy et al., 2026)	SPIRA 2.0	Audio Spectrogram	CNN6	98.6%	Dados de pré-treinamento apenas em inglês. <i>Corpus</i> pequeno como (Gauy et al., 2023)

Fonte: Retirada de (Matheus Gauy et al., 2026).

11.4 Considerações finais e impactos para a sociedade

Os resultados apresentados indicam a viabilidade da detecção de problemas respiratórios através da voz e fala de pacientes. Os modelos analisados, embora apresentem um grau de complexidade elevado para o tamanho do *corpus* de treinamento, se mostraram bastante eficazes nas tarefas. O uso de diversas técnicas para prevenir o *overfitting* é crucial para atingir esse objetivo, especialmente o uso de modelos pré-treinados em grandes volumes de



áudios. A possibilidade de identificar problemas respiratórios de forma automática, com uma simples gravação de um paciente em um celular, é de grande valia para a sociedade e uma grande ajuda para médicos, enfermeiros e outros agentes de saúde.

Naturalmente, cuidados são necessários ao usar modelos caixa preta para detecção de problemas de saúde. Uma dificuldade natural é que esses modelos, por serem caixa preta, não permitem uma interpretação fácil de seus resultados, e não fornecem justificativas para a tomada de decisão. Embora existam técnicas como o Grad-CAM (Selvaraju et al., 2017) para facilitar a interpretação desses modelos caixa preta, e muitos autores as usaram para mostrar que os modelos de saúde em muitos casos identificam fatores relevantes nos dados (Moujahid et al., 2022; Panwar et al., 2020; Silva et al., 2024; Sobahi et al., 2022), é um problema em aberto, em geral, produzir modelos de aprendizado profundo que apresentem resultados interpretáveis. Parte desse problema pode ser evitado, deixando claro que esses modelos devem ser usados para uma análise preliminar, como a triagem de pacientes e a opinião final sempre deve ser de um médico especialista na área. Mesmo assim, seria bastante valioso criar modelos mistos que apresentem, como saída, além da previsão caixa preta do modelo, diversos parâmetros de voz e fala característicos e os resultados da comparação desses parâmetros de fala e da voz de pacientes com os mesmos parâmetros da fala e da voz de pessoas saudáveis. As análises dos linguistas e fonoaudiólogos sobre os áudios de pacientes e controles para determinar fatores de voz e fala que distinguem um grupo do outro são cruciais para uma melhor compreensão do problema (Berti et al., 2023, 2025; Fernandes-Svartman et al., 2022).

Outro cuidado, ainda mais relevante no uso desses modelos, consiste no fato de que é muito difícil gerar modelos que generalizem para fora do contexto original de seu conjunto de treinamento. Em particular, quando temos *corpora* pequenos como os de saúde, esses normalmente apresentam características específicas que permitem que o modelo use fatores específicos dos dados coletados e não generalize para fora daquele conjunto. Um exemplo claro desse efeito para problemas respiratórios pode ser encontrado em (Gauy et al., 2023). Os autores demonstraram que modelos de aprendizado profundo treinados no *corpus* SPIRA (coletado durante a pandemia), apresentam desempenho péssimo se testados no *corpus* SPIRA 2.0 (coletado após a vacinação e envolvendo insuficiência respiratória proveniente de causas além da COVID-19). O mesmo ocorre se o processo inverso for feito (treino no SPIRA 2.0 e teste no SPIRA). Ressalta-se que os modelos considerados em (Gauy et al., 2023) possuem todos excelente desempenho nos conjuntos de teste extraídos do *corpus* no qual foram treinados. Uma vez retirados de seu contexto original, porém, esses modelos passam de acurácias acima de 95% para acurácias menores que a de um chute aleatório. Esta dificuldade prática de implementar modelos de aprendizado profundo para problemas de saúde é documentada de forma mais ampla em (Cohen et al., 2021).

Para o caso de detecção de insuficiência respiratória e, potencialmente, para outros problemas respiratórios, é conhecido que diversos fatores de voz e fala são bastante diferentes entre pacientes e controles. Por exemplo, a distribuição das pausas nos enunciados de pacientes é bastante diferente da distribuição das pausas nos enunciados dos controles. A duração das pausas na fala dos pacientes também é bem maior que a dos controles. Considerando que diversos desses fatores que podem ser extraídos dos áudios devem ser semelhantes não apenas na fala dos controles mas na fala de todos os falantes do português brasileiro (a distribuição das pausas, por exemplo, é prescrita pela gramática do português, na medida em que há lugares na sentença em que sua ocorrência resultaria em agramaticalidade), é um caminho promissor construir modelos que não tratem dos áudios



diretamente mas desses fatores específicos de voz e fala que sabemos serem relevantes e mais estáveis, independentemente da forma de coleta. Naturalmente, esse método estaria abandonando a estratégia *end-to-end* tão comum na construção de modelos de aprendizado profundo. As dificuldades de construir *corpus* de saúde de grande volume e de criar modelos de aprendizado profundo que generalizem para fora de seu contexto podem tornar esse método alternativo, ao tentar controlar o tipo de entrada que é fornecida aos modelos, uma estratégia necessária para criar modelos mais robustos na prática.



Capítulo 12

PLN no Direito

Perspectivas e Desafios com Textos Jurídicos e Legais

Maria José Bocorny Finatto

Aline Macohin

Publicado em: 26/09/2023

Atualizado em: 16/04/2026

12.1 Introdução

Neste capítulo, tratamos de diferentes aspectos associados ao trabalho computacional com textos produzidos na esfera do Direito. As tarefas de PLN envolvidas, em geral, são a análise textual e a representação de conteúdos por meio de diferentes técnicas, mas há várias abordagens e estudos, voltados para diferentes finalidades. O nosso objetivo é apresentar apenas algumas perspectivas e desafios no âmbito de trabalhos que exploram materiais produzidos em português, considerando somente o cenário do Direito Brasileiro. Afinal, o Direito, de país para país, tem especificidades linguísticas e culturais que repercutem muito sobre seus textos, discursos e tipo de vocabulário.

Enquanto este capítulo introduz o gênero jurídico e discute alguns desafios para o PLN, o Capítulo 31 trata da tarefa de Reconhecimento de Entidades Nomeadas nos textos legais. Por sua vez, o Capítulo 32 propõe métodos de geração automática de ementas (resumos de processos judiciais).

Iniciamos este capítulo apresentando alguns aspectos sócio-históricos do Direito Brasileiro, que acabam influenciando suas práticas de escrita e os seus conteúdos textuais. Em seguida, situamos exemplos de reconhecimento e exploração do vocabulário jurídico, dos seus modos de dizer e, especialmente, das suas terminologias. Vamos partir de dois diferentes cenários textuais: as leis e sentenças judiciais. A primeira parte de exemplos tem a ver com um trabalho que denominamos reconhecimento terminológico (**RT**). O **RT** envolve a identificação dos usos e dos significados de um vocabulário especializado ao longo de acervos textuais, os quais formam um corpus de estudo. Esse trabalho, atualmente, é baseado em fontes escritas disponíveis em formato digital e se beneficia muito das técnicas da Linguística de *Corpus* (Sardinha, 2000) e do PLN. Depois dessa parte, mais dedicada ao vocabulário e terminologias, segue um exemplo de estudo em PLN, na área conhecida como **Análise de Sentimentos** (Capítulo 16).

O território de materiais para estudo e de enfoques, em Direito, é extremamente amplo, isso se ficarmos restritos aos trabalhos que lidam com os textos jurídicos da atualidade. Ainda assim, vale mencionar que uma série de estudos históricos sobre a linguagem jurídica brasileira, tratando de seus conceitos e até preconceitos, têm sido muito úteis para uma crítica social e política sobre o Direito. Para esses estudos históricos, os processos sobre crimes no período colonial e do império, reunidos em *corpora* que se exploram com apoio



computacional, têm mostrado a importância de se fazer uma linha de tempo de ações e de entendimentos até os dias de hoje. No Brasil, temos já, por exemplo, diferentes pesquisas filológicas e linguísticas dedicadas a estudos de processos criminais dos séculos 17, 18 e 19.

Entretanto, para se trabalhar com textos antigos em português, há todo o processo de *normalizar* e padronizar a apresentação escrita das “palavras antigas”, para então podermos fazer o seu processamento. A normalização de textos é, assim, um desafio multidisciplinar de uma nova área de estudos denominada *Humanidades Digitais* e que inclui o PLN no tratamento de acervos antigos.

Conforme nos coloca o artigo de Cameron et al. (2023), os desafios são muitos. Afinal, geralmente trabalha-se com textos em forma de arquivo provenientes de manuscritos que foram “decifrados” e transcritos. Isso significa enfrentar muitas questões associadas à variabilidade da escrita. Afinal, uma mesma palavra podia apresentar-se de vários modos, em um mesmo documento, escrito por uma mesma pessoa, como nos casos de ÁGUA/AGUA/AGOA ou UMA/HUA/HUMA. Há exemplos interessantes desses tipo de estudo histórico, no âmbito do Direito Penal e da Medicina Legal, com processos judiciais que envolveram crimes contra mulheres no Brasil do século 19. O artigo de Teixeira et al. (2022) mostra um exemplo de estudo bem interessante nesse tema do histórico dos tratamentos jurídicos para as práticas de violência contra as mulheres.

Mas, voltando à atualidade dos textos e da linguagem do Direito, veremos, mais adiante, como exemplos, alguns textos jurídicos brasileiros, buscando ilustrar suas peculiaridades. Vamos destacar: **a)** o texto do Estatuto da Criança e do Adolescente (ECA), conforme apresentado na Lei 8.069-90, promulgada em 13 de julho de 1990 e atualizada em 2021; **b)** o texto da nossa Constituição do Brasil, de 1988 (CF88); e, **c)** um conjunto de Sentenças Judiciais dos chamados “tribunais de pequenas causas”, os Juizados Especiais Cíveis. Conforme pretendemos deixar claro, esses três tipos de fontes, em suas características linguísticas e textuais, podem estar associados a diferentes tarefas de PLN, desde a descrição do português até a pontos específicos de *Recuperação da Informação*, área conhecida como *Information Retrieval* (Capítulo 2).

Por isso, um outro exemplo que trazemos neste capítulo é o da análise de conteúdos em sentenças judiciais via *Análise de Sentimentos*. Este tipo de técnica pode ser muito útil para identificar, por exemplo, padrões de sentenças judiciais favoráveis ou desfavoráveis a um determinado assunto. A utilidade dessa técnica para os profissionais do Direito é grande, pois um profissional geralmente faz buscas para entender como um determinado tribunal já vem decidindo sobre um assunto específico. Nesse trabalho de pesquisa, são buscadas retornadas inúmeras sentenças e documentos. Vale destacar que existem tribunais no Brasil inteiro, que lidam diversos assuntos (trabalhistas, penais, civis, eleitorais, entre outros). Nesses órgãos são protocolados milhares de novos processos diariamente e neles existe uma base de milhões de processos já julgados, muitos já em formato digital. A criação de um método que possa filtrar, por exemplo, as causas que foram consideradas favoráveis, em um dado tema, tende a reduzir o trabalho de leitura individual de cada sentença, ajudando o profissional a buscar e encontrar a informação que precisa.

12.2 O Direito – uma moldura para a significação

Conforme já mostraram os estudos (Motta, 2021, 2022), o Direito se manifesta através da língua, pois são as palavras que emprega e os enunciados que produz que conferem e confirmam a sua existência peculiar (Maciel, 2001) como uma prática social e área de conhecimento. Assim, temos uma relação intensa entre o Direito e a língua em uso



pelas pessoas que nele atuam. Isto é, pelo emprego de certas palavras¹, com um sentido particular e pela forma como suas proposições e teses são enunciadas vemos todo um cenário de valores. Isso é tão importante que temos, há bastante tempo, uma área de estudos específica conhecida como *jurilinguística* (veja mais em Cornu (1990)).

Estudiosos dessa área da linguística (Montoro, 1998, p. 1998) explicam que a linguagem jurídica, sempre com destaque para escrita, compreende diversas “espécies” de práticas que se subdividem, conforme uma dada finalidade e foco. Vejamos um detalhamento dessas espécies ou modos de se apresentar conforme seus propósitos (Petri, 2017, p. 47):

1. linguagem legislativa – a linguagem dos códigos - como o Código Penal ou Civil, das normas; sua finalidade: criar o Direito;
2. linguagem judiciária, forense ou processual - é a linguagem dos processos e sentenças; sua finalidade: aplicar o Direito;
3. linguagem convencional ou contratual – é a linguagem dos contratos, por meio dos quais se criam direitos e obrigações entre as partes;
4. linguagem doutrinária – é a linguagem dos mestres, dos doutrinadores, cuja finalidade é explicar os institutos jurídicos, é ensinar o Direito;
5. linguagem cartorária ou notarial – a linguagem jurídica que tem por finalidade registrar os atos de Direito.

Assim, embora se possa pensar em uma linguagem jurídica em geral, quando lidamos com sentenças produzidas em processos judiciais, temos linguagem judiciária, forense ou processual. Quando trabalhamos com os textos de leis, decretos e portarias, temos a linguagem legislativa. Cada tipo de suporte e/ou instrumento jurídico tende a adotar usos diferenciados e um vocabulário diferenciado. E esses elementos podem ser importantes quando se trabalha com o processamento em larga ou pequena escala desses textos. Como há uma especificidade de discursos envolvida, considerar os seus elementos linguísticos e modos de dizer próprios poderá nos ajudar a desempenhar tarefas de um modo mais produtivo. Afinal, o “Direito é, por excelência, entre as que mais o sejam, a ciência da palavra. Mais precisamente: do uso dinâmico da palavra” (Xavier, 2002, p. 1).

Conforme já mencionamos no início deste capítulo, em cada país, a linguagem jurídica tende a realizar um uso particular da língua nacional. Por isso, a linguagem do Direito de um país se diferencia da de um outro – como acontece com a linguagem jurídica dos diferentes países de Língua Portuguesa. Embora haja elos comuns, o Direito brasileiro é bastante distinto do de países como Angola ou Portugal. E, mesmo os textos jurídicos, em suas diversas formas (no Brasil conhecidos como petições, recursos, decisões judiciais etc.) podem adotar nomes e modelos de apresentação diversos. Conforme a cultura jurídica e as tradições de cada país, os produtores dos textos jurídicos serão também “autorias” diferentes, conforme o que é estabelecido no ordenamento legal de cada país.

O Direito no Brasil é regido pelo sistema da *civil law*, isso significa que uma a lei escrita tem preponderância sobre a jurisprudência – que são as decisões dos juízes – lembrando que os juízes são encarregados de verificar e direcionar a aplicação das leis. No Brasil, quem produz as leis são os membros do poder legislativo, eleitos, democraticamente, pelo povo. Os textos das leis são discutidos e votados, e então aprovados para entrarem em vigor. Os membros do poder executivo, também eleitos pelo povo, devem executar as leis

¹Veja mais sobre a delimitação de “palavras” e de unidades de processamento no Capítulo [Sequência de caracteres e palavras](#). Neste capítulo sobre Direito e PLN, vamos acentuar a noção de palavra como uma unidade da língua escrita, situada entre dois espaços em branco, ou entre espaço em branco e sinal de pontuação.

aprovadas. Vejamos um resumo sobre como se organiza o Direito brasileiro, atualmente, em suas hierarquias:

- **Constituição Federal:** é a lei maior do país e define os direitos e deveres dos cidadãos, além de estabelecer a organização e o funcionamento dos poderes Executivo, Legislativo e Judiciário.
- **Legislação infraconstitucional:** são as leis, decretos e normas que regulamentam assuntos específicos, como por exemplo, o Código Civil, Código Penal, Código de Defesa do Consumidor, entre outros.
- **Poder Executivo:** é composto pelo Presidente da República, Vice-Presidente e ministros. É responsável pela administração do país e pela implementação das leis.
- **Poder Legislativo:** é formado pelo Congresso Nacional (Senado Federal e Câmara dos Deputados). É responsável por criar, modificar e aprovar as leis.
- **Poder Judiciário:** é composto pelos tribunais e juízes. É responsável por aplicar a lei em casos concretos, solucionar conflitos e garantir os direitos fundamentais dos cidadãos.

Apesar de o sistema jurídico brasileiro ser o *civil law*, há grande influência da jurisprudência nas decisões judiciais, principalmente quando agrupadas pelos tribunais e transformadas em *súmulas*. A súmula é um tipo de documento que consiste em um verbete que registra a interpretação pacífica ou majoritária adotada por um Tribunal a respeito de um tema específico. Portanto, quando textos legislativos e documentos processuais tornam-se objetos do PLN, com vistas a obter conhecimento para os profissionais do Direito, será preciso compreender esses elementos e valores diferenciados. Sem isso, há o risco de “misturar alhos com bugalhos”.

12.3 Entre as terminologias e as palavras no Direito do Brasil

Grosso modo, um RT equivale à identificação e à sistematização de denominações associadas a conceitos conforme utilizadas em um dado campo ou área do conhecimento. Geralmente, o RT envolve a produção de uma “lista” de nomes (termos) vinculados aos seus significados (conceitos). Além disso, junto de cada item dessa “lista”, tem-se um conjunto de informações que ajudam a contextualizar e a entender o seu uso ao longo de um conjunto de documentos escritos.

Desse modo, vamos pensar nesse processo ao longo de um conjunto de documentos jurídicos - em um dado tipo - tendo em mente a situação particular do uso de suas palavras. A Terminologia descritiva, para nós um ramo da Linguística Aplicada brasileira, e os terminólogos dedicam-se a estudar – descrever e compreender - os diferentes fenômenos linguísticos da comunicação técnico-científica, o que se estende ao Direito, em seus variados cenários.

O que diferencia uma terminologia ou um “termo técnico” de uma palavra “comum” é, em primeiro plano, o seu ambiente comunicativo. E, repetindo a ideia de uma das maiores autoridades da nossa área da Terminologia descritiva internacional (Cabré, 2005), podemos dizer: uma palavra não é um termo técnico-científico, ela está nessa condição em determinados contextos, que conferem a ela um significado “especial”. Esse significado ou modo de compreensão especial, chamaremos, grosso modo, de conceito.

Vejamos um exemplo, com a palavra/item CRIANÇA, muito corriqueira no nosso dia a dia. Como seu significado básico, geralmente, entendemos algo como “pessoa não adulta”.



Mas, quando empregada e “significada” em um dado ambiente comunicativo de especialidade, como é o caso do Direito brasileiro, essa palavra “comum” assume contornos semânticos diferenciados.

No contexto do nosso Estatuto da Criança e do Adolescente, documento brasileiro conhecido como ECA, que corresponde à Lei 8.069-90, atualizada em 2021, que podemos enquadrar no domínio do Direito Civil do Brasil, temos o seguinte:

“Art. 2º Considera-se **criança**, para os efeitos desta Lei, a pessoa até doze anos de idade incompletos, e adolescente aquela entre doze e dezoito anos de idade” (BRASIL, 1990, grifo nosso).

Como se percebe, há um significado “especializado”, jurídico, uma delimitação em termos de anos de idade, que se soma ao nosso entendimento mais comum de **criança**. E você deve estar se perguntando: no que isso importará ou como poderia repercutir em um trabalho computacional sobre o tema das crianças em leis e documentos em português? A resposta é: isso importa muito! Se você comparar esse conceito com os que estabelece a OMS, Organização Mundial da Saúde, a faixa etária de uma pessoa considerada como **criança** é outra, pois compreende pessoas até 19 anos de idade. Isto é, os traços/valores de uso da palavra, que adquire estatuto terminológico, são variáveis. Além disso, temos uma conceituação jurídica específica e bem particular associada a um dado termo que, à primeira vista, não pareceria ser um termo, e sim uma palavra comum.

No caso do segmento de lei acima, o ECA, podemos considerar que há uma definição específica para CRIANÇA, que se opõe à de ADOLESCENTE. Além disso, essa definição é circunscrita, isto é, ela vale apenas em um dado contexto ou “*frame* de significação”. Assim, teríamos um problema, para aquelas pessoas que se interessassem pelo Direito das Crianças, seja em sistemas jurídicos específicos, como o do Brasil, ou que busquem um mapeamento sobre esse tema no âmbito do Direito Internacional. Afinal, há variação no modo como se delimita o conceito em foco e isso pode ser um problema também para o tratamento automático dessas informações, não é mesmo?

Vamos supor uma aplicação de PLN que pudesse nos ajudar a dar conta de uma busca de informações sistematizadas sobre esse tema, o Direito das Crianças, mas restrita ao cenário brasileiro. Como vimos, em Direito, temos uma definição que tende a ser circunscrita, isto é, ela vale apenas em um dado contexto, correspondendo a um valor que estabelece frente a todo um CONJUNTO DE OUTROS TERMOS E CONCEITOS com ela relacionados. Isso é o que chamamos de sistema conceitual, que tem a ver como uma rede de conceitos e de terminologias que se entrelaçam. Como vimos, o ECA está subordinado à Constituição do Brasil, e ainda podemos ter, por exemplo, leis estaduais ou municipais – ou mesmo códigos e portarias – que “valem como leis locais” sobre, por exemplo, o tratamento de crianças em estabelecimentos de Saúde em diferentes estados do Brasil. Além das normas, também pode haver interpretações jurídicas unânimes ou diversas sobre assuntos relacionados à criança e disponibilizadas em sentenças judiciais.

12.4 Um caso concreto: em pequeníssima escala

Para realizar um ensaio de um reconhecimento terminológico (RT), podemos explorar um conjunto de textos que servem de referência ou espelhamento em uma dada área de conhecimento (veja um passo a passo detalhado de um RT com a *Constituição do Brasil* em Finatto et al. (2022)). Lidando com textos jurídicos, como vimos, será importante



levar em conta suas naturezas e tipologias. Assim, vamos supor um RT associado, por hipótese, ao tema “Direitos das Crianças no Brasil”. Esse RT poderia envolver identificar, em diferentes documentos relevantes previamente selecionados, os seguintes elementos:

- a) TERMOS e seus respectivos CONCEITOS
- b) TERMOS e seus respectivos FORMATOS LINGÜÍSTICOS
- c) TERMOS, CONCEITOS e respectivos TERMOS E CONCEITOS RELACIONADOS.

Nos itens a) e b), acima, entra em jogo uma questão muito importante: a variação terminológica. Essa variação que um TERMO pode ter tem a ver com as diferentes formas das denominações e das conceituações, dentro de uma dada especialidade ou subárea. Você poderá perguntar sobre o escopo do RT: vamos explorar esse tema nos âmbitos do Direito Civil até o Direito Criminal? Ou vamos ficar apenas em um dado recorte legislativo? Essa decisão é muito importante e deve ser bem ponderada antes do trabalho começar.

Para administrar a variabilidade de termos e conceitos, sem a ideia de condená-la, pois o enfoque linguístico e conceitual em um reconhecimento terminológico (RT) é sempre descritivo, teremos, para nos socorrer, os vocabulários controlados e/ou padronizados. Esses vocabulários mostram padrões de denominações que geralmente são colocados pela autoridade de órgãos profissionais associados a uma dada especialidade. Nesses vocabulários, encontramos as “terminologias padronizadas” e também as “normas técnicas” de uma área. Assim, uma forma de denominar um respectivo conceito/significado pode ser formalmente estabelecida em um dado contexto, de modo a se garantir precisão e boa correlação com outros termos e conceitos relacionados. Essa padronização será importante especialmente em situações de trocas de conhecimento e de trocas em geral, como no caso da comunicação jurídica internacional. Ainda assim, todos os terminólogos sabem que as variabilidades dos usos da língua, apesar dos esforços normativos, sempre se apresentam e que elas devem ser descritas e administradas.

Guardadas as devidas diferenças, é semelhante o caso, por exemplo, do conceito de CRIANÇA frente ao conceito de ADOLESCENTE no nosso Estatuto da Criança e do Adolescente, o ECA. Ainda que pareça um “detalhe”, as crianças não poderão ser confundidas, em um cenário legal e jurídico, com adolescentes, muito menos com pessoas adultas, salvo condições especiais definidas naquele texto, que funciona como uma moldura de significação para suas terminologias e conceituações.

O mesmo vemos nos casos dos nomes “oficiais” para algumas doenças, que inclusive correspondem a um código numérico, conhecido como CID ou Classificação Internacional de Doenças. A ideia, nesse contexto de padronização das terminologias da área da Saúde, é evitar confusões de nomes e tentar garantir que todos possam ter um mesmo entendimento – ou conceito uniforme – de um dado TERMO + CONCEITO/DESCRIÇÃO DE SEU SIGNIFICADO. Abaixo, alguns exemplos dessa padronização da CID para o termo SARAMPO e seus tipos – uma doença, no Brasil, geralmente associada a **crianças**.

- CID 10 – B05 – Sarampo
- CID 10 – B05.0 – Sarampo complicado por encefalite
- CID 10 – B05.1 – Sarampo complicado por meningite

Dada a relevância e necessidade de tratar esse assunto, alguns tribunais como o Supremo Tribunal Federal e Superior Tribunal de Justiça criaram um site denominado “Tesauro” como forma de ferramenta para controle terminológico que tem por objetivo a padronização da informação. Nesta ferramenta, o tesauro, são apresentados os termos, conceitos,



termos relacionados, mas também categorias, termos genéricos e termos específicos. A partir deste mapeamento do tesouro, é possível orientar que os servidores públicos redijam os documentos judiciais com uma terminologia uniforme, para auxiliar na pesquisa e recuperação da informação posteriormente. Para saber mais sobre o tema dos *tesauros e sua interface com as terminologias*, recomendamos consultar o trabalho de Vargas; Van der Lann (2011).

Em função dos diversos contextos a que os tribunais se vinculam, há uma variedade maior de termos relacionados em comparação à legislação e cada tribunal pode apresentar informações diversas nos *tesauros* para o mesmo termo. *Tesauros* são listas de assuntos, palavras-chave e de terminologias de uma dada área de conhecimento. Essas listagens, que também se estudam há bastante tempo na Terminologia descritiva do Brasil (Van der Laan, 2002), dão suporte à indexação e catalogação de documentos em bibliotecas e em diferentes acervos físicos, como também nas bases de dados. Geralmente, quem produz esses tesauros são os bibliotecários, documentalistas e cientistas da informação que lidam com a catalogação de informações técnicas e científicas. Veja este exemplo de como funciona um tesouro, quando se busca pelo item CRIANÇA no *tesouro* do Supremo Tribunal Federal. Nessa busca, temos o seguinte resultado:

- **Termo Genérico:** Menor
- **Termos relacionados:** Adulto, Castigo Físico, Direito da Criança e do Adolescente, Educação, Educação Infantil, Estatuto da Criança e do Adolescente (ECA), Guarda de Menor, Investigação de Paternidade, Poder Familiar.
- **Categoria:** DCT Direito Constitucional, ECA Estatuto da Criança e do Adolescente.

O que se informa aqui não é um conceito para CRIANÇA, mas se aponta que ele tem um correspondente ou equivalente genérico nesse âmbito. Isto é, CRIANÇA = MENOR (genérico). Em seguida, nos termos relacionados, vemos assuntos em que se inclui esse item.

Feitas essas explicações sobre peculiaridades das terminologias jurídicas, em suas diferentes circunstâncias e variabilidades de uso e de significações, um reconhecimento terminológico (RT) pode ser visto como um tipo de trabalho de mediação de comunicação, realizado por profissionais de uma área, terminólogos, linguistas, informatas, entre outros. Salienta-se, assim, a ideia de uma mediação terminológica (Conceição; Zanola, 2020).

Além disso, um RT pode ser um trabalho multidocumento, por exemplo, com diferentes tipos de leis e normas, e multitemático, quando se tenta cobrir diferentes temas ou assuntos que se entrelaçam. Pode, ainda, apontar ligações entre documentos de diferentes naturezas, extrapolando-se o reconhecimento de um dado tópico para diferentes fronteiras. Um exemplo seriam os materiais sobre temas e políticas de Saúde Pública voltadas para crianças e os documentos jurídicos que estabelecem seus direitos. Outro exemplo de trabalho seria verificar como um determinado tribunal interpreta e aplica a legislação sobre crianças, diante de alguns problemas específicos, em âmbitos regionais ou locais.

12.5 Outros casos/exemplos: Direito Ambiental

Um reconhecimento terminológico (RT) legislativo também poderia servir de apoio para um recurso didático voltado para o cidadão comum, sem formação em Direito, ou mesmo para diferentes estudantes universitários interessados na legislação ambiental do Brasil.



Nesse caso, vamos imaginar um conjunto composto, por exemplo, por 800 leis, as quais versam sobre diferentes aspectos ambientais. Vamos supor que estamos trabalhando em um RT para uso de jornalistas novatos que lidam com temas ambientais. Como explorar essas 800 leis para chegar, por exemplo, a um conjunto de seus termos e conceitos conforme sejam mais comumente empregados nessas leis? Como apresentar a informação de forma a melhor atender o nosso suposto usuário jornalista, que, sem ter formação em Direito ou Biologia, precisaria ler e entender a legislação? Bastaria perguntar ao ChatGPT?

Naturalmente, hoje, dada a larga prática de digitalização desse tipo de documento e a garantia de seu acesso a todo o cidadão, parece ser fácil encontrar e percorrer uma base de dados com leis ambientais. O Senado Federal do Brasil, por exemplo, oferece todo um banco de leis, decretos e outros documentos afins para acesso público. Basta a pessoa acessar um site determinado e salvar os documentos no seu computador. Feito isso, “bastaria” a pessoa – o jornalista novato que imaginamos – ler, calmamente e com cuidado, todas as 800 leis do nosso caso imaginário e ir fazendo um registro, em um arquivo de texto, de suas terminologias e conceituações à medida que avance com a leitura.

Outra opção seria o “nosso” jornalista consultar um dicionário especializado sobre esse tema. A constituição de um dicionário especializado também serve como um exemplo de um ponto final possível, entre outros, quando se produz um reconhecimento terminológico (RT). Dicionários terminológicos são um ponto final ou um produto possível, entre outros, quando se elabora um RT Assim, ao empreender um RT específico, podemos nos beneficiar de RTs anteriores ao nosso.

Como a legislação, em alguns aspectos, pode ser principiológica ou apenas fornecer diretrizes, ou ainda possuir conflitos de termos entre normas diversas, pode ser necessário associar um outro reconhecimento terminológico (RT) para identificar o entendimento prevalente. Uma legislação principiológica é um tipo de norma jurídica que se baseia mais em princípios do que em regras detalhadas. Em vez de definir condutas específicas de forma minuciosa, ela apresenta valores, diretrizes gerais e objetivos fundamentais que devem orientar a interpretação e aplicação do Direito. Neste caso, podemos associar a um RT Judicial, alguns *tesauros* de tribunais, como o mencionado anteriormente.

Naturalmente, além dessas fontes padronizadas, há dicionários especializados em terminologias e glossários descritivos sobre o tema do Direito Ambiental do Brasil. Um exemplo é o *Dicionário de Direito Ambiental* do Grupo Termisul da UFRGS, publicado em segunda edição em 2008. A produção desse dicionário demandou construir, desde 1994, toda uma base legislativa sobre temas do meio ambiente associada à obra, a Base Legis². Essa base, que começa com o texto do Código de Águas do Brasil, de 1930³.

Mas, voltando aos tesauros, segundo o *tesauro* do Supremo Tribunal Federal, temos as seguintes informações:

- **Termo Genérico:** Direito (Ciência Jurídica)
- **Termo Relacionado:** Lei da Biossegurança

²A Base Legis Termisul-UFRGS é composta de textos da Legislação Ambiental do Brasil, Alemanha, Argentina, Estados Unidos, França, Paraguai e Uruguai. Também inclui códigos brasileiros, constituições dos países anteriormente mencionados e dos demais países que empregam a língua portuguesa (Angola, Cabo Verde, Guiné Bissau, Moçambique, Portugal, São Tomé e Príncipe e Timor Leste), Atos Internacionais relativos ao meio ambiente (Agenda 21, Convenção de Estocolmo, Declaração do Rio e Protocolo de Kyoto). Todos os textos possuem uma descrição de conteúdos e pontos principais, podendo ser baixados em formato TXT.

³Você pode acessar todos os textos legislativos ambientais reunidos pelo grupo Termisul- UFRGS em <http://www.ufrgs.br/termisul>, na aba “Recursos” , “Bases textuais” e “Base legis”.



- **Categoria:** DAM Direito Ambiental

Como termo, há poucas informações sobre Direito Ambiental, mas ao acessar a categoria “DAM Direito Ambiental”, temos cerca de 200 termos relacionados. Vale destacar que os termos apresentados fazem parte do contexto ao qual o tribunal pertence e os processos judiciais que julga. Um RT legislativo pode ter mais termos que os apresentados no tesauro do Supremo Tribunal Federal, mas em contrapartida, o *tesauro* do tribunal, pode ter um nível de detalhamento maior.

Já para o Superior Tribunal de Justiça do Brasil, ao se consultar o termo/assunto **Direito Ambiental**, temos os seguintes resultados:

- **Termos relacionados:** Bioética, Meio Ambiente, Órgão público ambiental, Princípio da Precaução, Princípio da Prevenção, Princípio do In Dubio Pro Natura, Teoria do Risco Integral.

A vantagem do tesauro elaborado por alguns tribunais é que esse instrumento reduz esforços na aplicação do PLN ao Direito, uma vez que já associa os termos mais frequentes nas decisões judiciais e associa esses termos às decisões judiciais existentes. Dado que muitos dos tesauros não associam conceitos aos termos, ao trabalhar com RT de outras fontes, como o legislativo, podem ser obtidas estas informações.

12.6 Aplicação: Análise de Sentimentos em Direito: desafios e exemplos

A análise de sentimento em textos jurídicos envolve a aplicação de técnicas de PLN para determinar o tom emocional, qualificativo ou opinativo presente nos documentos legais.

Dentre diferentes possibilidades de aplicação da técnica de análise de sentimento em textos jurídicos, podemos trabalhar com sentenças judiciais. Assim, analisa-se o contexto das sentenças judiciais e identifica-se se o juiz foi favorável ou desfavorável ao pedido de cada parte. Para esse tipo de trabalho, alguns passos devem ser seguidos como: 1) Coleta de Dados; 2) Pré-Processamento do Texto; 3) Rotulação de Dados; 4) Escolha da técnica de análise de sentimento; 5) Execução da técnica e 6) Avaliação e Validação.

Na coleta de dados, é necessário escolher um repositório que possua os textos das decisões judiciais, sejam elas de forma resumida ou na íntegra. Dentre as opções públicas e gratuitas, estão os diários de justiça dos tribunais, uso de APIs (*Application Programming Interface*) públicas como o DataJud do Conselho Nacional de Justiça e decisões disponibilizadas nos sistemas de busca dos tribunais. Para automatização desta coleta, é preciso o conhecimento de técnicas de *web crawling* e *web scraping* (Macohin; Carneiro, 2020). Essas técnicas consistem na automatização do download das páginas e arquivos que possuem decisões judiciais e posterior filtragem da informação que se deseja usar), respectivamente.

Um exemplo de informação que pode ser obtida de um tribunal pode ser verificado abaixo. Neste exemplo foram suprimidas algumas informações que pudessem identificar os envolvidos. Veja que lidamos com um tipo de texto que contém *uma parte* denominada *ementa* e outra parte que é o *acórdão*, com o resultado do processo.

A partir do download desta página, o objetivo é extrair a informação do *acórdão*, último parágrafo da imagem, onde consta se foi dado ou negado provimento ao pedido do autor. Neste caso, foi negado provimento ao autor do recurso, como se verifica através do uso das palavras “negou provimento”.



Figura 12.1: Exemplo de resumo de acórdão judicial de um tribunal superior brasileiro.

PROCESSO			
REsp	/ SC		
RECURSO ESPECIAL			
2019/	-0		
RELATOR	ÓRGÃO JULGADOR	DATA DO JULGAMENTO	DATA DA PUBLICAÇÃO/FONTE
Ministro	T2 - SEGUNDA TURMA	/11/2022	DJe /12/2022
RELATOR PARA ACÓRDÃO			
Ministro			
EMENTA			
ADMINISTRATIVO. AÇÃO CIVIL PÚBLICA AJUIZADA PELA UNIÃO COM VISTA A OBTEN O RESSARCIMENTO POR DANO PATRIMONIAL DECORRENTE DE EXPLORAÇÃO DE MINÉRIO (BASALTO) SEM AUTORIZAÇÃO. PRETENSÃO SUJEITA À INCIDÊNCIA DA PRESCRIÇÃO QUINQUENAL. RECURSO ESPECIAL A QUE SE NEGA PROVIMENTO.			
1. No caso dos autos, a União ajuizou ação civil pública com o objetivo de obter ressarcimento pela lavra ilegal de basalto.			
2. O Tribunal Regional Federal manteve a sentença de improcedência do pedido, pois, "Em se tratando de ação civil pública movida pelo Poder Público em face de particular (não abrangido pelo conceito de agente público), objetivando a reparação de dano decorrente da extração ilegal de recursos minerais, deve ser aplicado o prazo prescricional quinquenal delineado na Lei da Ação Popular (Lei n. 4.717/1965)".			
Inconformada, a União recorrente defendendo o afastamento da prescrição.			
3. O entendimento adotado pelo Tribunal Regional Federal da 4ª Região não merece reparos, pois a imprescritibilidade da pretensão de ressarcimento ao erário se aplica somente em casos excepcionais, como é o do ato doloso de improbidade administrativa; e a incidência da prescrição, como regra, consagra o princípio da segurança jurídica (e até mesmo o da ampla defesa), não sendo cabível o sacrifício de direito fundamental do particular como medida de compensação da ineficiência da máquina pública.			
4. Recurso especial a que se nega provimento.			
ACÓRDÃO			
Vistos, relatados e discutidos esses autos em que são partes as acima indicadas, acordam os Ministros da SEGUNDA TURMA do Superior Tribunal de Justiça, na conformidade dos votos e das notas taquigráficas, o seguinte resultado de julgamento:			
"Proseguindo-se no julgamento, após o voto-vista do Sr. Ministro Francisco Falcão, acompanhando o Sr. Ministro-Relator, dando provimento ao recurso especial, a Turma, por maioria, negou provimento ao recurso especial, nos termos do voto do Sr. Ministro Og Fernandes, vencidos os Srs. Ministros Herman Benjamin e Francisco Falcão. Lavrará o acórdão o Sr. Ministro Mauro Campbell Marques."			
Os Srs. Ministros Mauro Campbell Marques e Assusete Magalhães votaram com o Sr. Ministro Og Fernandes.			

Em posse dos trechos das decisões judiciais que se deseja analisar, de forma automatizada, se o desfecho foi favorável ou desfavorável, pode-se iniciar a fase de pré-processamento do texto. A fase de pré-processamento pode contemplar diversas subtécnicas. Mas, para fins de exemplificação, vamos citar apenas a tokenização, a remoção de pontuações, conversão de todas as letras para minúsculas e remoção de *stop words*.

A tokenização consiste em dividir o texto em palavras ou unidades menores, chamadas de *tokens* (veja mais detalhes no Capítulo [Sequência de caracteres e palavras](#)), que são conjuntos de caracteres separados por um espaço em branco. Um critério para separação dos *tokens* pode ser o espaço entre as palavras. Já a remoção de pontuações visa eliminar pontuação e caracteres especiais que não são relevantes para a análise de sentimento.

Em seguida, a conversão para minúsculas consiste em transformar todas as palavras em minúsculas para garantir consistência nas comparações. Por fim, a remoção de *stop words* consiste na remoção de palavras que são comuns - as palavras gramaticais ou instrumentais - e não contribuem significativamente para uma análise de sentimento, como "a", "o", "em", "por" etc. Vale destacar que a lista de *stop words* deve ser na mesma língua do texto analisado. Veja o Quadro [12.1](#) abaixo.

Quadro 12.1: Exemplo de pré-processamento de um trecho de decisão judicial.



Texto Original: “Vistos, relatados e discutidos esses autos em que são partes as acima indicadas, acordam os Ministros da SEGUNDA TURMA do Superior Tribunal de Justiça, na conformidade dos votos e das notas taquigráficas, o seguinte resultado de julgamento: ‘Prosseguindo-se no julgamento, após o voto-vista do Sr. Ministro Francisco Falcão, acompanhando o Sr. Ministro-Relator, dando provimento ao recurso especial, a Turma, por maioria, negou provimento ao recurso especial, nos termos do voto do Sr. Ministro Og Fernandes, vencidos os Srs. Ministros Herman Benjamin e Francisco Falcão. Lavrará o acórdão o Sr. Ministro Mauro Campbell Marques.’ Os Srs. Ministros Mauro Campbell Marques e Assusete Magalhães votaram com o Sr. Ministro Og Fernandes.”

Após Pré-processamento: “vistos relatados discutidos autos partes acima indicadas acordam ministros segunda turma superior tribunal justiça conformidade votos notas taquigráficas seguinte resultado julgamento prosseguindo julgamento após voto-vista sr ministro francisco falcão acompanhando sr ministro-relator dando provimento recurso especial turma maioria negou provimento recurso especial termos voto sr ministro og fernandes vencidos srs ministros herman benjamin francisco falcão lavrará acórdão sr ministro mauro campbell marques srs ministros mauro campbell marques assusete magalhães votaram sr ministro og fernandes”

Ainda na fase de pré-processamento, é possível aperfeiçoar a tarefa e incluir novas *stop words*, com o objetivo de limpar mais ainda o texto e facilitar futuramente a identificação das palavras positivas ou negativas na sentença. No Quadro 12.1, verifica-se que as palavras “srs”, “sr”, “vistos”, não influenciam na interpretação da decisão judicial e podem ser removidas.

A próxima fase, rotulamento de dados, consiste em classificar manualmente algumas decisões como positivas ou negativas, para fins de validação futura se a classificação automatizada está desempenhando um bom resultado. A partir do Quadro 12.1, facilmente esta decisão seria classificada como “NEGATIVA”. Outras opções de rótulo seriam “POSITIVA” e “NEUTRA”. Os casos de neutro poderiam ser utilizados, por exemplo, quando o juiz decidiu parcialmente pelo provimento (aceitação do pedido).

Já a fase da escolha da técnica de análise de sentimento, consiste em selecionar qual abordagem será utilizada, se baseada em regras ou baseada em aprendizado de máquina. Na abordagem baseada em regras, é criado um conjunto de regras e heurísticas que determinam o sentimento ou polaridade com base em palavras-chave, padrões gramaticais e outras características linguísticas. Por exemplo, certas palavras negativas podem indicar um sentimento negativo. Já, na abordagem baseada em aprendizado de máquina, é treinado um modelo de aprendizado de máquina usando-se os dados linguísticos rotulados. Algoritmos como *Naïve Bayes*, *Support Vector Machines* (SVM) ou redes neurais podem ser usados para construir um modelo.

Para dar continuidade ao exemplo mencionado, utilizaremos a abordagem baseada em regras. Nesse caso, utilizamos um dicionário prévio com palavras positivas e negativas. Quando são usados dicionários, deve ser considerada a língua do texto. Como exemplo de dicionário de palavras positivas, negativas e neutras em português, temos o SentiLex-PT⁴ (Carvalho; Silva, 2017).

Na fase de execução da técnica e a partir do texto pré-processado anteriormente, será preciso verificar se cada palavra do texto consta no dicionário como palavra positiva, nega-

⁴O SentiLex-PT está disponível em: <https://b2share.eudat.eu/records/93ab120efdaa4662baec6adee8e7585f>.

tiva ou neutra. Segundo o SentiLex-PT, foi encontrado o seguinte resultado apresentado no Quadro 12.2.

Quadro 12.2: Palavras positivas e negativas segundo o SentiLex-PT.

Dicionário:
 Positivas: [“acordam”, “conformidade”, “aprovar”]
 Negativas: [“negou”, “vencidos”]
 Sentimento:
 Positivas: 3 (acordam, conformidade, aprovar)
 Negativas: 2 (negou, vencidos)
 Sentimento Geral: 1 (POSITIVO) (Positivas - Negativas = 3 - 2)

O sentimento geral é calculado realizando-se uma subtração do número de palavras positivas com o número de palavras negativas. Um resultado com valor positivo indica um sentimento positivo, já um resultado com valor negativo indica um sentimento negativo e um valor próximo de zero indica um sentimento neutro. Este cálculo pode ser aperfeiçoado ao dividir o número encontrado pelo total de palavras (*tokens*) existentes no texto ((palavras positivas - palavras negativas) / total de palavras). Ou seja, se há 10 palavras no texto, 3 são positivas e 0 negativas, indica uma maior “probabilidade” que o texto realmente seja positivo. Por outro lado, se há 50 palavras no texto, apenas 1 negativa e nenhuma positiva, há probabilidade de ser um falso negativo. Essa divisão pode indicar que novos aperfeiçoamentos no dicionário podem ser necessários.

Neste exemplo acima, verifica-se que o resultado não reflete a resposta correta (NEGATIVO) e ajustes devem ser feitos. O dicionário SentiLex-PT pode ser adaptado para a linguagem jurídica, uma vez que “acordam” “conformidade” e “aprovar”, não indicam necessariamente que o juiz está dando provimento (julgando como positivo) a decisão judicial, logo, devem ser desassociados do sentimento “POSITIVO”. Outro ajuste que pode ser feito também é não associar a palavra “vencidos” ao sentimento “NEGATIVO”, uma vez que é comum, quando há divergência entre o grupo de juízes votantes, aparecer a palavra “vencidos”. Outro ajuste que também pode ser feito no dicionário é associar palavras frequentemente encontradas juntas e que reflitam a intenção da decisão judicial, por exemplo “negou provimento”, “negado provimento”, “não provido o recurso”, “ não dado provimento”, entre outros.

Feitos estes ajustes, teremos o seguinte resultado apresentado no Quadro 12.3.

Quadro 12.3: Resultado após os ajustes.

Dicionário:
 Positivas: [“ ”]
 Negativas: [“negou provimento”]
 Sentimento:
 Positivas: 0 ()
 Negativas: 1 (negou provimento)
 Sentimento Geral: -1 (NEGATIVO) (Positivas - Negativas = 0 - 1)

Lembrando de que essa abordagem é uma simplificação e pode não capturar todas as nuances de sentimento e/ou as polaridades em textos jurídicos complexos. Principalmente quando há uma variedade de pedidos sendo julgados com decisões diferentes para cada pedido. O contexto legal específico também pode influenciar a interpretação das palavras. Portanto, ajustes e validações são sempre necessários.



Por fim, com relação à fase de avaliação e validação, se foi utilizada a abordagem baseada em regras, como a acima exemplificada, é mais simples validar com a amostra rotulada previamente e comparar a taxa de acertos e erros. Já no caso da abordagem baseada em aprendizado de máquina, é possível utilizar parâmetros estatísticos para demonstrar a precisão e desempenho do modelo.

Como mencionado anteriormente, é necessário fazer ajustes em cada fase da execução da análise de sentimentos devido às peculiaridades discursivas dos textos jurídicos que nem sempre constam nos dicionários de referência existentes. A partir das informações obtidas na última fase, de avaliação e validação, pode-se sugerir que novos aperfeiçoamentos sejam feitos nas fases anteriores, para que o algoritmo tenha um desempenho similar e até superior ao de uma atividade humana.

Dentre os desafios para aplicar este tipo de técnica em decisões judiciais, está principalmente na coleta dos dados. Os tribunais, no geral, não possuem repositórios com estas informações prontas, estruturadas, segmentadas e públicas. Isso, por si só, já dificulta iniciar qualquer trabalho de Processamento de Linguagem Natural. Apesar do problema poder ser contornado com a criação de algoritmos de *web crawling* e *web scraping*, alguns tribunais fazem uso de *captchas* que impedem o acesso automatizado e massivo às informações. Embora haja iniciativas do Conselho Nacional de Justiça, como o DataJud, para centralizar e fornecer informações estruturadas por meio de uma API, ainda não há a íntegra das decisões disponibilizadas. Entretanto, como o DataJud continua em constante evolução, é possível que, futuramente, esse recurso seja disponibilizado. A limpeza e seleção das informações contidas em uma página HTML ou em um arquivo PDF também é bastante custosa e trabalhosa. Por isso, somente a partir destes esforços se torna possível dar seguimento à aplicação, em melhores condições, da técnica de análise de sentimentos em Direito.

12.7 Considerações finais

Este capítulo tentou situar o “mundo textual” oferecido e formatado pelo Direito do Brasil. Com isso, o objetivo foi trazer alguns exemplos de trabalhos e estudos com a sua linguagem e as suas práticas de escrita. Comentamos um pouco dos estudos históricos com a linguagem empregada em processos brasileiros antigos sobre crimes contra mulheres, mencionamos que a tarefa de reconhecimento terminológico (RT) pode ser uma demanda importante da sistematização de termos e de conceitos dos Direitos da Criança e do Direito Ambiental. Por fim, trouxemos um breve exemplo de estudo de sentimentos ou de polaridades, em PLN, que serve para ajudar a detectar os tipos de decisões que os juízes brasileiros tomam em determinados tipos de processos federais.

Buscamos salientar que é importante considerar as características e elementos de diferentes tipos de documentos, sejam leis, códigos, processos ou sentenças, produzidos por diferentes instâncias jurídicas. E, visto esse panorama, caso você possa se interessar especificamente por sentenças de tribunais administrativamente menores, como os Juizados Especiais Cíveis (JECs) - conhecidos como “tribunais de pequenas causas”, vale conhecer o trabalho de doutorado de Motta (2022). A autora estudou elementos da complexidade das sentenças dos JECs, quanto ao vocabulário, terminologias e sintaxe, frente aos preceitos da legislação que estabelece que tais sentenças devem ser escritas em linguagem simples, que possibilitem fácil compreensão sobre o que se decide em uma causa. Afinal, o cidadão comum recorre aos JECs geralmente sem advogados, em meio a causas de valor limitado. Além de ampla análise, Motta (2022) oferece, no seu trabalho, acesso a todo



um *corpus* de sentenças por ela reunido e analisado com a ferramenta NILC-METRIX ⁵ (Leal et al., 2023). Também os *corpora* que ela usou como contraponto para ponderar a complexidade/facilidade de linguagem dessas sentenças estão disponíveis nos anexos da sua tese.

O Direito é um mundo feito de palavras e modos de dizer, o que oferece um terreno fértil para os nossos trabalhos de análise linguístico-textual, em geral, e, em especial, para diferentes tarefas do PLN. Os resultados desses trabalhos tendem a auxiliar tanto os profissionais quanto o cidadão e a sociedade, que são os principais focos e beneficiários das ações do Direito.

⁵<http://fw.nilc.icmc.usp.br:23380/nilcmatrix>



Capítulo 13

Reconhecimento de Entidades Nomeadas no Domínio Legal

Um Panorama para a Língua Portuguesa

*Ellen Souza
Hidemberg O. Albuquerque
Nádia F. F. Silva
Matheus Cerqueira
Rafael O. Nunes
André C. P. L. F. de Carvalho
Adriano L. I. Oliveira*

Publicado em: 20/11/2024

Atualizado em: 16/04/2026

 <https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/>

13.1 Introdução

Neste capítulo, são apresentadas as principais iniciativas na área de reconhecimento de entidades nomeadas no domínio legal, com foco na língua portuguesa. São descritos os principais *corpora*, modelos de linguagem neurais e de larga escala, além de exemplo de aplicação de entidades nomeadas na recuperação de documentos legislativos.

O Capítulo 3 apresenta conceitos relacionados à Extração de Informação (EI), que é, normalmente, dividida em diversas tarefas de interesse, com foco no tipo de informação a ser extraída do texto, entre elas, o Reconhecimento de Entidades Nomeadas (REN), a Extração de Relações (ER) e a Extração de Eventos (EE). Com relação ao REN, são apresentadas informações históricas, conceituação formal, abordagens para rotulação, métricas de avaliação, principais entidades e *corpora*, e estado da arte dos modelos de REN para a língua portuguesa.

O Capítulo 12, por sua vez, apresenta diferentes aspectos associados ao Processamento de Linguagem Natural (PLN) na esfera jurídica. As tarefas de PLN envolvidas, em geral, são a análise textual e a representação de conteúdos por meio de diferentes técnicas, mas há várias abordagens e estudos voltados para diferentes finalidades. São descritos desafios e perspectivas no âmbito de trabalhos que exploram materiais produzidos em português, considerando somente o cenário do Direito brasileiro. Também, é exemplificada uma aplicação de análise de sentimentos em Direito.

13.2 Contextualização

Nos últimos anos, houve um grande crescimento no uso de técnicas de PLN para a área jurídica, que produz uma grande quantidade de dados no formato de texto (Zhong et



al., 2020). Em várias atividades do meio jurídico, é necessário extrair informações desses textos. No entanto, a crescente quantidade de documentos legais, associada ao tamanho desses documentos, torna muitas vezes impraticável sua análise manual. A Inteligência Artificial (IA), em particular PLN, provê técnicas que podem automatizar a extração de conhecimento de textos. Por conta disso, PLN é cada vez mais utilizada para lidar com a grande quantidade de documentos produzidos pelas organizações legislativas e judiciais. Em muitos países, há um considerável acúmulo de casos jurídicos a serem processados e o número de documentos gerados é enorme (Kapoor et al., 2022). A Câmara dos Deputados brasileira, desde sua fundação, já processou mais de 144 mil projetos de lei e tem processado aproximadamente 30 mil projetos de lei a cada ano (Brandt, 2020). Para cada projeto, diversos documentos são produzidos e agregados, em diferentes etapas, até sua discussão e votação.

O levantamento anual do Conselho Nacional de Justiça (CNJ) revelou um aumento significativo no número de projetos de IA no Poder Judiciário Brasileiro em 2023¹, como parte do Programa Justiça 4.0. Em comparação com 2022, houve um crescimento de 17% no número de tribunais com projetos de IA. Das 140 soluções tecnológicas mapeadas, 63 já estão em uso ou aptas a serem utilizadas e 46 estão em fase final de desenvolvimento. Já em 2023, 63% dos tribunais participaram no desenvolvimento de soluções baseadas em IA, em um crescimento de 27% em relação à 2022². Os principais motivadores para o uso da IA pelos tribunais incluem o aumento da produtividade, a busca por inovação, a melhoria na qualidade dos serviços e a redução dos custos. Em 2018, em parceria com a Universidade de Brasília (UnB), o Supremo Tribunal Federal (STF) desenvolveu o Victor (Hartmann Peixoto, 2020), uma IA que separa e classifica as peças processuais mais usadas nas atividades do STF e identifica os temas de repercussão geral de maior incidência. Além do projeto Victor, o STF desenvolveu a VitorIA que identifica, no acervo do Tribunal, os processos que tratam do mesmo assunto e os agrupa automaticamente, e a RAFA 2030 (Redes Artificiais Focadas na Agenda 2030), uma IA, lançada em 2022, para apoiar a classificação de processos de acordo com os Objetivos de Desenvolvimento Sustentável da Agenda 2030 da Organização das Nações Unidas³. Em 2026, o CNJ apresentou o Promptus⁴, um repositório nacional para padronização de instruções (*prompts*) de IA no ecossistema da Plataforma Digital do Poder Judiciário, visando aumentar, entre outras, a segurança jurídica na Justiça brasileira.

Em 2019, a Câmara dos Deputados lançou o Ulysses, um conjunto institucional de iniciativas de IA com o propósito de aumentar a transparência, melhorar a relação da Câmara com os cidadãos e apoiar a atividade legislativa com análises complexas (Almeida, 2021). Inicialmente, o Ulysses analisava, classificava e distribuía os pedidos dos parlamentares entre as 22 áreas de conhecimento da Consultoria Legislativa da Câmara. Posteriormente, em parceria com pesquisadores do Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (ICMC-USP), novos algoritmos⁵ foram desenvolvidos, permitindo a busca por documentos similares (Albuquerque et al., 2024b; Souza et al., 2021a; Vitório et al., 2022, 2024), análise do posicionamento de cidadãos sobre um projeto de lei

¹<https://www.cnj.jus.br/programa-justica-4-0-divulga-resultados-de-pesquisa-sobre-ia-no-judiciario-brasileiro/>

²<www.cnj.jus.br/wp-content/uploads/2025/09/relatorio-entregas-do-programa-justica-4-0-versao-interativa.pdf>

³<https://portal.stf.jus.br>

⁴<https://www.cnj.jus.br/tecnologia-da-informacao-e-comunicacao/justica-4-0/conheca-o-conecta/promptus/>

⁵<https://github.com/ulysses-camara>



em tramitação (Maia et al., 2022; Silva et al., 2021), reconhecimento de entidades relevantes para o domínio legislativo (Albuquerque et al., 2022, 2023a; Costa et al., 2022), além da construção de modelos de linguagem específicos para o domínio legislativo (Garcia et al., 2024a) e um grande *corpus*, o Ulysses Tesemão (Siqueira et al., 2024), composto por mais de 3,5 milhões de arquivos, totalizando 30,7 GiB de texto bruto, coletados de 159 fontes que abrangem dados judiciais, legislativos, acadêmicos, notícias e outros documentos relacionados.

O domínio legal abrange uma grande variedade de textos jurídicos, incluindo legislação, jurisprudência e trabalhos acadêmicos (Maxwell; Schafer, 2008). A natureza dos documentos legais também acrescenta um novo desafio às aplicações de PLN que pode não estar presente em outros domínios, pois esses documentos são, tipicamente, muito longos, desestruturados, contêm ruídos, e são escritos usando jargão e uma linguagem específica do domínio (Kapoor et al., 2022). Há iniciativas em várias tarefas de PLN e IA, como: Recuperação de Informação Legal (Chalkidis et al., 2021; Souza et al., 2021a), Sumarização de Texto (Kornilova; Eidelman, 2019), Previsão de Julgamentos (Chalkidis et al., 2019), Segmentação Semântica (Malik et al., 2021) e Reconhecimento de Entidades Nomeadas (Albuquerque et al., 2022; Alles, 2018; Araujo et al., 2018; Castro, 2019; Costa et al., 2022).

Legal Named Entity Recognition ou reconhecimento de entidades nomeadas no domínio legal (REN-Legal), tem como objetivo detectar e rotular todas as instâncias de entidades nomeadas específicas, juridicamente relevantes, dentro de textos legais (Bonifacio et al., 2020; Cabrera-Diego; Gheewala, 2023; Cardellino et al., 2017). Considerando que os relatórios jurídicos contêm um grande número de termos e terminologias complexas, identificar entidades é um aspecto crucial na organização de informações e conhecimento dentro do domínio. É uma tarefa com grande importância, pois consiste no primeiro passo da análise semântica do texto, com potencial aplicação em diversas tarefas. Contudo, estudos sobre REN-Legal para a língua portuguesa evidenciam que os modelos utilizados para este idioma enfrentam desafios não encontrados para outros idiomas, o que pode ser explicado pelo baixo volume de *corpora*, ferramentas e modelos pré-treinados desenvolvidos para a língua portuguesa (Albuquerque et al., 2023b; Castro, 2019), necessitando de um esforço maior, principalmente no desenvolvimento de recursos e abordagens, como ocorre com a língua inglesa (Pirovani, 2019).

13.3 Iniciativas de REN-Legal

A revisão sistemática apresentada por Firdaus Solihin; Makarim (2021) mapeou 4.798 pesquisas sobre Extração de Informação no domínio legal, entre janeiro de 1990 e dezembro de 2020, extraíndo dados de 107 estudos selecionados. Os autores relatam que 53% das pesquisas utilizaram *corpora* compostos de decisões judiciais, 23% de leis ou regulações, 21% utilizaram outros documentos legais, e os demais 4% utilizaram *corpora* abertos pré-existentes. Os estudos mostraram que as pesquisas utilizando REN-Legal tiveram maior destaque (acima de 60 estudos), ressaltando sua importância para a EI.

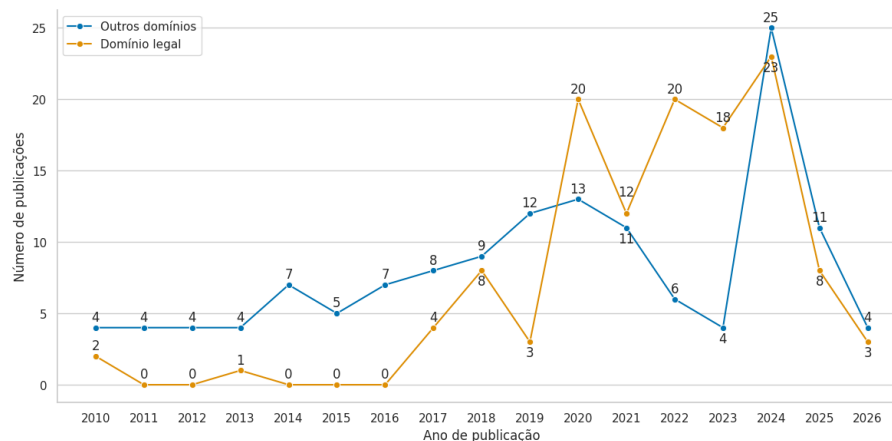
Para se ter uma ideia do avanço do estado da arte, para a edição deste capítulo foi executada uma pesquisa amostral das publicações sobre REN-Legal utilizando o *Google Scholar*⁶. Foram utilizadas duas *strings* de busca sobre REN, uma sem informação de

⁶A escolha pelo motor de busca foi baseada na sua praticidade, amplitude (e.g., artigos de eventos, revistas e trabalhos acadêmicos), e facilidade de replicação.



domínio, e a outra com o domínio legal⁷. Para nossos propósitos, foram selecionados arbitrariamente os 100 primeiros artigos de cada resultado, filtrados pelo título e pelo resumo, no período de 2010 a 2026⁸. A Figura 13.1 apresenta o resultado amostral do avanço das pesquisas sobre REN-Legal, principalmente a partir de 2017, ressaltando a importância dada a este domínio especialmente entre 2020 e 2024, período em que o número de publicações em REN-Legal se manteve próxima ou superior às demais pesquisas na maior parte dos anos.

Figura 13.1: Amostra da evolução do estado da arte de REN-Legal



Além da natureza global e complexidade documental próprias do domínio, alguns fatores como diferenças culturais, idioma, linguagem, padronização de documentos, entre outros, podem ser determinantes para o sucesso das aplicações desenvolvidas. Também é preciso considerar seus subdomínios jurídicos existentes (e.g., processual, penal, trabalhista, previdenciário etc.), e suas esferas de atuação (e.g., tipos de tribunais, jurisdição etc.) (Akbik et al., 2016; Cabrera-Diego; Gheewala, 2023; Firdaus Solihin; Makarim, 2021). Neste sentido, alguns trabalhos se destacam por sua relevância para as tarefas de REN-Legal.

Dale; Mazur (2007) exploram o uso de aprendizado de máquina supervisionado para desambiguação de entidades em documentos legais, utilizando 13.460 documentos em inglês da Bolsa de Valores Australiana, aplicados na estimação de frequência de conjunções de entidades candidatas. Foram extraídos 16 tipos de entidades utilizando anotação automática, com desambiguações resolvidas manualmente. Dos nove algoritmos de aprendizagem avaliados, o classificador IBk⁹ (baseado no algoritmo *K-nearest neighbors* (KNN) (Aha et al., 1991)) obteve o melhor resultado, com 84% de medida F1.

Dozier et al. (2010) utilizaram REN para identificação de informações legais em 43.936 documentos de processos legais norte-americanos. Os documentos foram anotados manualmente utilizando cinco entidades legais, classificadas através de abordagens híbridas baseadas em pesquisa (utilizando listas de entidades de interesse), regras contextuais (com regras dedutivas) e modelos estatísticos (com indicadores ponderados). O classificador *Support Vector Machine* (SVM) (Cortes; Vapnik, 1995) obteve o melhor resultado com 95% de medida F1.

⁷Dado que a maioria dos trabalhos sobre REN são escritos em inglês, foram utilizados os termos neste idioma: “*named entity recognition*” e (“*named entity recognition*”+*legal*), respectivamente.

⁸Pesquisas executadas em 24/07/2024, e atualizadas em 28/03/2026 sem o uso de filtros da ferramenta. Para 2026, os resultados representam apenas as publicações recuperadas até março deste ano.

⁹<https://weka.sourceforge.io/doc.stable-3-8/weka/classifiers/lazy/IBk.html>



Chalkidis et al. (2017) exploram a automação da extração de elementos contratuais, construindo um *corpus* de ~3.500 contratos em inglês, manualmente anotados com 11 entidades, das quais três são do domínio legal. Utilizando regras manuais e classificadores lineares como Regressão Logística (Tolles; Meurer, 2016) e SVM, os melhores resultados foram alcançados através de um método híbrido de aprendizado de máquina e regras pós-processamento, com medida F1 acima de 84%. Posteriormente, o *corpus* desenvolvido foi utilizado para gerar vetores de palavras utilizando ~750.000 contratos (Chalkidis; Androutsopoulos, 2017), com melhorias utilizando os modelos *Long Short-term Memory* (LSTM) (Hochreiter; Schmidhuber, 1997), *Bidirectional LSTM* (BiLSTM) (Schuster; Paliwal, 1997) e *Conditional Random Field* (CRF) (Lafferty et al., 2001). Os resultados obtidos alcançaram medida F1 de 87% na classificação por *tokens*, e 88% na identificação de elementos nos contratos.

Chalkidis et al. (2020) exploram a adaptação do modelo *Bidirectional Encoder Representations from Transformers* (BERT) (Devlin et al., 2019) para o domínio legal, através de estratégias de pré-treinamento e ajuste fino do modelo em um *corpus* de 354.621 textos legais em inglês, composto por leis, decisões de tribunais e contratos de diversas fontes. Para as tarefas de REN, foram utilizados dados de cerca de 2.000 contratos, chamado de Contracts-NER. As entidades utilizadas foram as mesmas dos trabalhos anteriores dos autores. O estudo conclui que o modelo LEGAL-BERT, especialmente uma versão menor (LEGAL-BERT-small), foi eficaz para tarefas específicas do domínio legal, apresentando desempenho comparável a modelos maiores.

Pilán et al. (2022) apresentam o *corpus* TAB (*Text Anonymization Benchmark*), desenvolvido para a anonimização de dados pessoais utilizando REN. Composto por 1.268 casos judiciais em inglês do Tribunal Europeu dos Direitos Humanos, o *corpus* foi anotado manualmente utilizando oito entidades universais. No processo de anonimização, entidades marcadas como sensíveis foram categorizadas de acordo com informações como crença filosófica, opinião política, entre outras. Juntamente com o *corpus*, a pesquisa introduz três métricas específicas de avaliação de anonimização: identificadores diretos, identificadores quase-indiretos, e identificadores diretos e quase-indiretos. Um modelo BERT ajustado ao *corpus* alcançou o resultado de 99,9% de revocação para identificadores diretos.

Aejas et al. (2023) apresentam um método para automação de contratos escritos em linguagem natural no contexto de cadeia de suprimentos, utilizando *blockchain*, REN e ER. Foram utilizados 200 contratos públicos em inglês, anotados manualmente com nove entidades do domínio. Os experimentos foram executados com modelos de *deep learning* na avaliação do *corpus* e na criação de um meta-modelo comum para as diferentes estruturas de REN e ER, utilizadas na representação de regras de negócio padronizadas. A acurácia encontrada nestes modelos foi de 99,4% e 99,6%, utilizando BERT e BiLSTM, respectivamente.

Pesquisas relevantes sobre REN-Legal também foram executadas em outras línguas, além da língua inglesa. Na língua alemã, Leitner et al. (2019) apresentam uma metodologia detalhada para REN em documentos legais, com 19 classes de entidades semânticas específicas, generalizadas em sete tipos de classes, que foram avaliados utilizando os modelos CRF e BiLSTM. Mais recentemente, Darji et al. (2023) adapta o modelo GermanBERT para o domínio legal, utilizando as mesmas entidades (Leitner et al., 2019), com resultados melhores que os modelos originais em 14 tipos de entidades específicas. No grego, Angelidis et al. (2018) combinam os modelos BiLSTM, CRF e Regressão Logística para REN em um *corpus* governamental, utilizando seis tipos de entidades. A pesquisa desenvolveu um vocabulário de referências para integrar dados legais com outras fontes na internet,



criando assim uma rede de informações legais acessíveis. Na língua chinesa, Chen et al. (2020) desenvolveram um sistema de extração de entidades e suas relações semânticas em documentos criminais relacionados a drogas, utilizando os modelos RoBERTa (Liu et al., 2019c), BiLSTM e a técnica Seq2Seq (Sutskever et al., 2014). A pesquisa de Çetindağ et al. (2022) apresenta o primeiro estudo sobre REN-Legal para o idioma turco. Foi utilizado um *corpus* de casos de tribunais com oito tipos de entidades (04 universais e 04 legais), os modelos BiLSTM e CRF, além de seis diferentes combinações de vetores de palavras. Na língua espanhola, Samy (2021) apresenta uma metodologia para anotação de entidades em textos legais escritos em espanhol peninsular, utilizando uma abordagem híbrida com expressões regulares, listas de fontes externas e ferramentas para REN. Cinco tipos de entidades foram utilizadas.

O REN também foi explorado em outros idiomas, como árabe (Shamma et al., 2020), finlandês (Oksanen et al., 2019a, 2022), francês (Andrew; Tannier, 2018; Barriere; Fouret, 2019; Sleimi et al., 2021), holandês (Heusden et al., 2023; Schraagen et al., 2017), italiano (Agnoloni et al., 2022; Lorè et al., 2023), romeno (Pais et al., 2021, 2024), além dos idiomas russo (Averina et al., 2022), urdu (Iftikhar et al., 2019), vietnamita (Bach et al., 2019), japonês (Onaga et al., 2023), entre outros, incluindo abordagens multilíngues (Kulkarni et al., 2023; Moreno Schneider et al., 2022; Quaresma; Gonçalves, 2010).

13.4 Iniciativas de REN-Legal em língua portuguesa

Albuquerque et al. (2023b) realizaram um mapeamento das técnicas, métodos e recursos de REN para a língua portuguesa. Foram realizadas buscas manual e automática no período de janeiro de 2010 a junho de 2022, recuperando 447 estudos primários, com 45 estudos incluídos na revisão. Os autores observaram um crescente interesse no desenvolvimento de pesquisas em REN a partir de 2019. Vinte e um estudos primários focaram na análise comparativa entre técnicas e ferramentas, tendo o modelo BiLSTM como o mais utilizado, seguido por métodos mais tradicionais que utilizam apenas CRF. Trabalhos mais recentes utilizaram o modelo BERT. Além disso, 24 estudos descreviam *corpora* de REN para vários domínios, entre eles o legal, com 8 estudos primários mapeados. A Figura 13.2 apresenta a evolução de trabalhos mapeados¹⁰, agrupados de acordo com a variante da língua portuguesa, sendo PT-EU para textos escritos exclusivamente com a variante Europeia e PT-BR com a variante Brasileira.

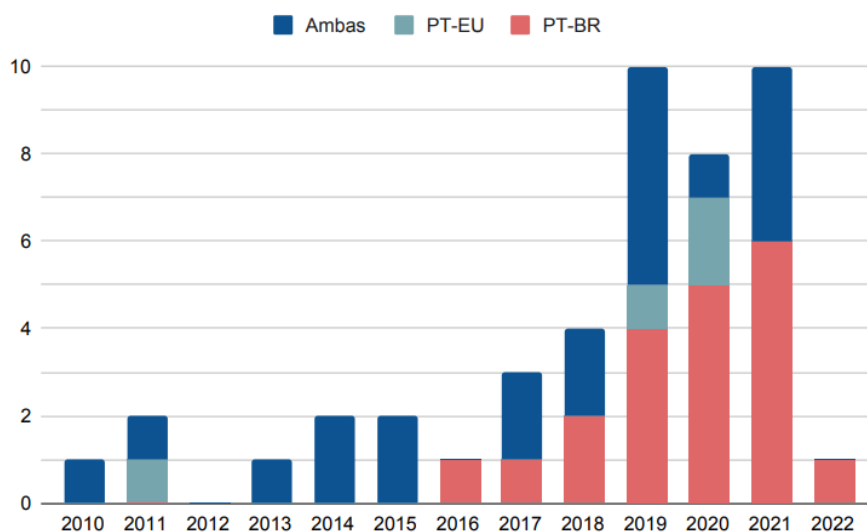
13.4.1 *Corpora* para REN-Legal

Esforços para a criação e compartilhamento de recursos, como *corpora/datasets* e modelos, vêm sendo feitos pela comunidade de PLN e EI (Firdaus Solihin; Makarim, 2021; Oksanen et al., 2019b). Estes recursos tornam-se cruciais para impulsionar novas pesquisas e o aprimoramento de pesquisas em curso (Bonifacio et al., 2020; Bordino et al., 2021). O Quadro 13.1 apresenta informações sobre alguns *corpora* para REN-Legal em língua portuguesa encontrados na literatura, destacadas a seguir.

¹⁰Segundo a pesquisa, o baixo número de pesquisas no ano de 2022 é justificado pelo período de execução do mapeamento, finalizado em junho de 2022.



Figura 13.2: Distribuição temporal dos artigos mapeados de REN para a língua portuguesa, entre janeiro/2010 e junho/2022 (Albuquerque et al., 2023b)

Quadro 13.1: *Corpora* REN-Legais em língua portuguesa com acesso público (PB) ou privado (PV)

<i>Corpora</i> / (Autores)	Entidades	#Docs/Anotação	Arquitetura	Acesso
LeNER-Br / Araujo et al. (2018)	Legislação, Local, Jurisprudência, Organização, Pessoa, Tempo	70 / manual	BiLSTM, CRF, Glove	PB ¹¹
DOU-corpus / Alles (2018)	Cargo, Data, Evento, Lei, Lugar, Número, Organização, Pessoa, Processo, Valor-monetário	470 / híbrida	OpenNLP	PV
Castro (2019)	Fundamento, Função, Local, Organização, Pessoa, Tribunal, Valor_Acordo, Valor_Causa, Valor_Condenação, Valor_Custas, Vara	1.305 / híbrida	Elmo, GloVe	PV
DrugSeizures-Br / Bonifacio et al. (2020) Bonifacio (2020)	Droga, Local, Pessoa, Organização, Outro, Tempo	6.218 / automática	BERT, ELMo	PV
KauaneJunior / Fernandes et al. (2020b)	DecMorDm, IncMorDm, InitIntArr, InitMon, ValLegFee, ValMorDm	3.022 / manual	BiLSTM, BIGRU, CRF	PV
Fernandes et al. (2020a)	Local, Organização, Pessoa	63.380 / automática	CRF, L-BFGS	PB ¹²
Rodríguez; Bezerra (2020)	Pessoa	10.000 / manual	NLTK	PB ¹³
Aposentadoria / Neto; Faleiros (2021)	Ato, Nome_Ato, Classe, Cod_Matricula_Ato, Empresa_Ato, Fund_Legal, Cargo, Quadro, Padrão, Processo	5.515 sentenças / híbrido	CNN, LSTM, CRF, GloVe	PB ¹⁴
Batista et al. (2021)	CPF_CNPJ, Estado_civil, Nacionalidade, Nome, OAB, RG	7.966 / híbrida	BiLSTM, CRF, Glove, Flair	PV
Mota et al. (2021)	Jurisprudência, Legislação	74 / manual	BiLSTM, CRF, Flair, Spacy	PV
Correia et al. (2022)	Citação acadêmica, Precedente, Pessoa, Referência legislativa	594 / manual	BiLSTM, CRF	PV
Lima; Borges (2022)	Aditamento Contratual, Aviso de Licitação, Aviso de Revogação/Anulação de Licitação, Aviso de Suspensão de Licitação, Extrato de Contrato, Extrato de Convênio	3.850 / manual	CRF, CNN, LSTM, Word2Vec	PB ¹⁵

UlyssesNER-Br / Albuquerque et al. (2022), Costa et al. (2022) CDJUR-BR / Brito et al. (2023) PersoSEG / Guimarães et al. (2024)	Data, Evento, Fundamento, Local, Organização, Pessoa, Produto de Lei Endereço, Norma, Pena, Pessoa, Prova, Sentença Obs.: iniciam com “Ato” Abono Permanencia, Cessao, Exoneracao_Comissionado, Exoneracao_Efetivo, Nomeacao_Comissionado, Nomeacao_Efetivo, Retificacao_Comissionado, Retificacao_Efetivo, Reversao, Substituicao, Tornado_Sem_Efeito_Apo, Tor- nado_Sem_Efeito_Exo_Nom	950 + 1000 comentários / manual 1.216 / manual 127 / manual	BERT, BiLSTM, CRF, Glove BERT, BiLSTM, CRF, Spacy CNN, CRF, LSTM	PB ¹⁶ (PL e C-corpus) PB ¹⁷ PB ¹⁸
UlyssesLegalNER-Br / Albuquerque et al. (2026)	Data, Evento, Fundamento, Jurisprudência, Local, Organização, Pessoa, Produto de Lei, Valor Monetário	560 / híbrida	BERT, BiLSTM, CRF, Glove	PB ¹⁹

O *corpus* LeNER-Br (Araujo et al., 2018) é composto por 70 documentos anotados manualmente para a tarefa de REN, sendo 66 provenientes de tribunais judiciais brasileiros, tais como: STF, Superior Tribunal de Justiça, Tribunal de Justiça de Minas Gerais e Tribunal de Contas da União. Outros quatro documentos são legislativos, como a Lei Maria da Penha. Além das entidades Pessoa, Local, Tempo e Organização, o *corpus* contém etiquetas específicas para as entidades Legislação, específica para leis, e Jurisprudência, para decisões relacionadas a casos legais. Um modelo BiLSTM-CRF (Lample et al., 2016) com arquitetura Glove (Pennington et al., 2014) foi retreinado com o *corpus*, alcançando as medidas F1 de 97,04%, 88,82% e 92,53% para entidade Legislação, para Jurisprudência e para todas as entidades, respectivamente.

O DOU-Corpus (Alles, 2018) contém entidades existentes nos documentos do Diário Oficial da União (DOU), um boletim oficial do Governo Federal brasileiro. Para o domínio, foram definidas novas entidades como Cargo, Lei, Número, Organização, Processo e Valor-monetário. A entidade Lei é definida pelos seguintes exemplos: Lei n 5548/97 e LEI 1865/89. A entidade Número descreve todo numerário explorado nos textos do DOU. A entidade Processo contém uma expressão que possui a palavra “processo” com um sequencial de números de uma certa formatação. O *corpus* é composto por 60 mil trechos extraídos de diários dos meses de agosto de 2015 à outubro de 2017, anotados utilizando um processo de anotação híbrida (manual e automática). As diferentes seções que compõem o diário foram avaliadas individualmente, alcançando medida F1 de 44,50% com a ferramenta OpenNLP²⁰.

O *corpus* da Justiça Trabalhista (Castro, 2019) é composto por 1.305 documentos obtidos do Processo judicial eletrônico (PJ-e), distribuídos entre atas de audiência, sentenças e acórdãos, e selecionados a partir de processos distribuídos em todas as 24 regiões da justiça do trabalho brasileira, entre os anos de 2008 e 2018. Utilizando uma abordagem de

¹⁹<https://github.com/peluz/lener-br>

¹⁹<https://journals.sagepub.com/doi/10.1177/0010414019858964#supplementary-materials>

¹⁹https://github.com/marciamrodriguez/AtosAdministrativos_RodriguezMarcia2018

¹⁹<https://avio11.github.io/resources/aposentadoria/aposentadoria.html>

¹⁹<https://github.com/alice7lima/PLN/tree/main/Projeto>

¹⁹<https://github.com/ulysses-camara/ulysses-ner-br>

¹⁹<https://github.com/mauriciobritojr/CDJUR-BR>

¹⁹<https://github.com/UnB-KnEDLe/persoseg-corpus>

¹⁹<https://github.com/ulysses-camara/ulysses-ner-br>

²⁰<https://opennlp.apache.org/>



anotação híbrida, foram anotadas as entidades Função, Fundamento, Local, Organização, Pessoa, Tribunal, Vara e Valores (Acordo, Causa, Condenação e Custas). A categoria Função corresponde ao papel das pessoas mencionadas nos documentos, tais como advogado, juiz, reclamado, entre outras funções. Fundamento refere-se a qualquer dispositivo jurídico que possa ser referenciado no documento. Tribunal e Vara visam a identificação de um órgão específico. Foi treinado um modelo de REN-Legal para a Justiça do Trabalho do Brasil, alcançando a medida F1 de 93,81%, utilizando as arquiteturas ELMo (Peters et al., 2018) e GloVe.

A pesquisa desenvolvida por Fernandes et al. (2020a) analisa o impacto de discursos legislativos na captação de votos em sistemas eleitorais portugueses por localidade. Foi utilizado um *corpus* composto por 60.000 discursos parlamentares escritos em português europeu, no período de 1999 a 2015. REN-Legal foi utilizada na análise dos discursos, identificando as entidades Local, Organização e Pessoa, que serviram para medir a frequência e o contexto em que os parlamentares mencionavam seus distritos eleitorais ou outros locais relevantes. Foi utilizado o modelo CRF para identificação automática de entidades, ajustado com o algoritmo *Limited-memory Broyden-Fletcher-Goldfarb-Shanno* (L-BFGS) (Malouf, 2002), a partir de dados sequenciais anotados manualmente. Além disso, para resolver ambiguidades de entidades geográficas, foi utilizada a técnica de Resolução de Topônimos com um dicionário de localidades e regras heurísticas. Para fundamentar a hipótese da pesquisa, foi utilizada uma avaliação qualitativa baseada em entrevistas. Os resultados indicaram que o localismo nos discursos é mais presente conforme a necessidade eleitoral, independente da magnitude do distrito eleitoral, e que, quanto maior o número de candidatos por distrito, maior a quantidade de localismo presente nos discursos.

DrugSeizures-Br (Bonifacio et al., 2020) é um *corpus* contendo 6.218 petições relacionadas a procedimentos de apreensão de drogas. Estas petições foram produzidas pelo Ministério Público do Estado de Mato Grosso do Sul (MPMS) entre os anos de 2015 e 2019. O *corpus* inclui uma tabela de metadados relacionados às petições que compreendem entidades de interesse do MPMS. As entidades anotadas estão divididas em 25 tipos diferentes envolvendo, por exemplo, nome e quantidade da droga apreendida, dados da pessoa denunciada (nome, documentos etc.), local da apreensão, entre outros. Os tipos estão agrupados em 6 categorias: Local, Pessoa, Tempo, Organização, Droga e Outro. Para anotação, foi aplicado um algoritmo simples de busca de *string*: a partir de uma entidade atrelada a uma denúncia, o algoritmo busca a *string* da entidade no texto da denúncia. O *corpus* foi avaliado utilizando a medida F1, com os modelos BERT Multilingual e BERT Acórdãos, alcançando 73,83% e 72,78%, respectivamente, utilizando todas as categorias. Considerando um cenário seletivo, no qual a categoria Outros foi removida do *corpus*, os modelos alcançaram medida F1 de 80,94% e 81,04%.

O *corpus* KauaneJunior (Fernandes et al., 2020b) é composto por 3.022 decisões judiciais sobre a matéria “negativação indevida”, do Tribunal de Justiça do Estado do Rio de Janeiro no ano de 2016. Os documentos foram anotados manualmente, utilizando diretrizes específicas de anotação baseadas em exemplos. Foram utilizados seis tipos de entidades legais: três relacionadas à definição de valor, aumento ou diminuição do dano moral (ValMorDm, IncMorDm e DecMorDm, respectivamente), uma para valor de honorários advocatícios (ValLegFee), além de duas entidades relacionadas a prazos iniciais, tanto para correção monetária (InitIntArr), como para juros de mora (InitMon). Foram utilizados cinco modelos de aprendizagem na avaliação da tarefa: *Bidirectional Gated Recurrent Units network* (BiGRU) (Graves et al., 2013), BiLSTM, CRF, e duas combinações destes modelos. Tanto na análise por entidades quanto no resultado geral, o modelo

BiLSTM-CRF obteve o melhor desempenho, alcançado a medida F1 de 94,79%.

O *corpus* Aposentadoria (Neto; Faleiros, 2021) contém entidades nomeadas de 10 classes associadas a atos de aposentadoria de servidores públicos do Diário Oficial do Distrito Federal (DODF), validadas pelo Tribunal de Contas do Distrito Federal e Territórios. O *corpus* faz parte de um grande conjunto de dados criado por colaboradores do projeto *KnEDLe*²¹ da Universidade de Brasília. As entidades mapeadas foram Ato, Nome_Ato, Classe, Cod_Matricula_Ato, Empresa_Ato, Fund_Legal, Cargo, Quadro, Padrão, Processo e estão anotadas no formato CoNLL-2003 (Sang; De Meulder, 2003). A divisão entre treinamento, validação e teste foi feita em ordem cronológica dos DODFs, para não dividir entidades do mesmo documento do DODF em diferentes conjuntos. O conjunto de treinamento foi formado por 3.860 sentenças, o de validação por 828 sentenças e o de teste por 827 sentenças.

O *corpus* UlyssesNER-Br (Albuquerque et al., 2022) é composto por 150 projetos de lei (PL-corpus) e 800 consultas legislativas (ST-corpus) da Câmara dos Deputados brasileira. Anotados manualmente em três fases e por três grupos de anotadores, o *corpus* possui sete categorias semânticas, algumas contendo tipos específicos. A avaliação da concordância entre anotadores alcançou coeficiente Kappa de 91% na média. Além de cinco categorias baseadas no HAREM (Santos; Cardoso, 2007b), Pessoa, Localização, Organização, Evento e Data, o *corpus* possui duas categorias específicas para o domínio legislativo: Fundamento e Produto de Lei. Fundamento refere-se a entidades relacionadas a leis (FUNDlei e FUNDapelido), resoluções, decretos, bem como a entidades específicas do domínio, como projetos de lei (FUNDprojetodelei), que são propostas de lei em discussão pelo parlamento, e consultas legislativas (FUNDsolicitacaotrabalho), realizadas pelos parlamentares. Produto de Lei refere-se a sistemas (PRODUTOsistema), programas (PRODUTOprograma) e outros produtos (PRODUTOoutros) criados a partir da legislação. O *corpus* foi avaliado com os modelos HMM e CRF, obtendo melhor resultado com o CRF, alcançando a medida F1 de 81,04%.

Posteriormente, o *corpus* UlyssesNER-Br foi expandindo com texto informal gerado pelo usuário (Costa et al., 2022). Para a expansão, foram utilizados, aproximadamente, 1000 comentários públicos de cidadãos (C-corpus), relacionados a projetos de lei em tramitação, coletados de enquetes do portal da Câmara dos Deputados. Foram anotadas manualmente as mesmas entidades da primeira versão (Albuquerque et al., 2022), com exceção da FUNDsolicitacaotrabalho, a qual só está presente nos documentos de consultas legislativas. O coeficiente de anotação Kappa alcançou a média de ~80%. O modelo BERT com ajuste fino, treinado com a junção do C-corpus e PL-corpus, melhorou o resultado individual do treinamento para C-corpus, alcançando medida F1 de 73,90%.

Correia et al. (2022) disponibilizaram um *corpus* anotado manualmente para REN em português no contexto do Judiciário brasileiro. O *corpus* contém 594 decisões publicadas pelo STF e anotadas por 76 estudantes de Direito. O processo de anotação foi composto por três fases, duas de treinamento e uma fase final. Os autores utilizaram um processo de anotação detalhada de decisões jurídicas, na qual foram anotados dois níveis de entidades jurídicas aninhadas. No primeiro nível, foram anotadas quatro entidades mais abrangentes (categorias): Precedente, Citação acadêmica, Referência legislativa e Pessoa. No segundo nível, foram anotadas entidades específicas de cada uma das quatro categorias. No total, foram identificadas 24 entidades aninhadas (tipos). Dois grupos de estudantes, ao longo de quase um ano, realizaram a tarefa de anotação sob rigorosa supervisão. Ambos os grupos

²¹<https://unb-knedle.github.io/>



realizaram anotação na mesma coleção de 1.363 trechos, com um coeficiente de concordância Kappa médio de 70%. Os modelos CRF e BiLSTM-CRF construídos obtiveram a mesma medida F1 de 90%.

O artigo de Brito et al. (2023) apresenta uma metodologia de anotação manual de documentos jurídicos que serviu como base para construção do CDJUR-BR, um *corpus* de REN-Legal com 1.216 documentos de processos judiciais brasileiros, contendo 21 entidades específicas das categorias Pessoa, Prova, Pena, Endereço, Sentença e Norma. Um comitê com professores da área do Direito e da Computação, juntamente com os especialistas do domínio, definiu os principais parâmetros do CDJUR-BR, bem como garantiu a adequada aplicação da metodologia. Foram selecionados documentos das classes-CNJ mais representativas dos processos de primeiro grau, encerrados em 2019 no Tribunal de Justiça do Estado do Ceará (85% dos processos de primeiro grau são das classes selecionadas). O processo de anotação foi realizado em duas fases, sendo a primeira de treinamento. Foram definidos três grupos de anotadores formados por especialista. A avaliação da concordância alcançou coeficiente Kappa de 0,69 para 73% dos documentos, e os demais documentos passaram por revisões por especialistas e etapas de refinamento. Com a CDJUR-BR, os resultados apontaram superioridade do modelo BERT com medida F1 média de 58%. Testes comparativos entre CDJUR-BR e LeNER-BR indicaram que o modelo REN treinado com a CDJUR-BR é superior em reconhecer entidades de outros *corpora*.

O *corpus* PersoSEG (Guimarães et al., 2024) é composto por 127 documentos legais em português, no período de 2001 a 2015, obtidos do DODF. A pesquisa utiliza técnicas de REN aplicadas às tarefas de segmentação e rotulação em documentos legais. Foram utilizadas 12 tipos de entidades manualmente anotadas, representando atos da administração pública, como por exemplo, abono de permanência, (Ato_Abono_Permanencia), nomeação ou exoneração de cargos comissionados (Ato_Nomeacao_Comissionado e Ato_Exoneracao_Comissionado) ou efetivos (Ato_Nomeacao_Efetivo e Ato_Exoneracao_Efetivo), entre outras. O *corpus* foi avaliado utilizando combinações dos modelos CRF, CNN e LSTM, com melhor medida F1 média de 75,65% para REN com o modelo CRF.

A pesquisa desenvolvida por Albuquerque et al. (2026) estende o *corpus* UlyssesNER-Br originalmente voltado para o domínio legislativo, para um *corpus* jurídico mais abrangente, unificado e de dados abertos, cobrindo três subdomínios: projetos de lei, jurisprudência e leis. O *corpus* UlyssesLegalNER-Br é composto por 560 documentos públicos, anotados por meio de uma abordagem híbrida, que combina sugestões automáticas com curadoria humana especializada. Com o objetivo de preservar a compatibilidade com a versão anterior, os autores utilizaram as entidades do UlyssesNER-Br, e acrescentaram duas novas categorias, Jurisprudência e Valor Monetário, totalizando 9 categorias semânticas e 23 tipos específicos. A pesquisa revisou sistematicamente as anotações anteriores a partir da versão original do PL-corpus (público), atualizando o esquema de anotação, corrigindo inconsistências e incorporando as novas entidades. O *corpus* ampliado passou então a ser formado por três *subcorpora*: o PL-corpus_v3, voltado para o domínio legislativo, ampliado para 300 documentos e com cobertura temática mais ampla, incluindo textos do Senado Federal; o Juris-corpus, com 160 decisões de diferentes cortes superiores e órgãos do sistema jurídico brasileiro; e o Leis-corpus, o primeiro *corpus* de REN composto exclusivamente por leis brasileiras, num total de 100 documentos, contendo, entre outros, a Constituição Federal de 1988, três constituições estaduais, além de outros códigos. Os *corpora* foram avaliados utilizando um protocolo experimental sistemático, tanto individualmente como de forma unificada, com modelos de aprendizado de máquina consolidados



no estado da arte (CRF, BiLSTM, Glove, BERTimbau). O estudo incluiu ainda uma análise complementar de eficiência computacional destes modelos. O BERTimbau apresentou os melhores resultados em todos os *corpora*, alcançando a maior medida F1 de 96,18% no *corpus* unificado. A análise dos resultados também permitiu observar como diferentes arquiteturas se comportam em *corpora* jurídicos distintos, reforçando o papel do *corpus* desenvolvido como um recurso multigênero que reduz a fragmentação de *corpora* jurídicos em português, oferecendo uma base consistente para estudos futuros em diferentes granularidades e tipos documentais. Todos os recursos do projeto (dados e modelos) foram publicados abertamente.

13.4.2 Modelos de linguagem neurais para REN-Legal

O Capítulo [Modelos de linguagem](#) descreve os conceitos relacionados aos modelos de linguagem, detalhando o funcionamento dos modelos probabilísticos, neurais e de grande escala (*Large Language Models* ou LLMs). Nesta subseção, são apresentados estudos e modelos neurais específicos para o REN-Legal em língua portuguesa.

A pesquisa de Bonifacio et al. (2020) avalia o uso de modelos de linguagem neural para as tarefas de REN-Legal, utilizando os modelos pré-treinados para a língua portuguesa ELMo e BERT, ajustados para o domínio legal. Os ajustes foram feitos utilizando o *corpus* Acordãos-TCU, composto de 298 mil documentos do Tribunal de Contas da União (TCU)²². Os modelos foram avaliados em diferentes cenários, utilizando os *corpora* HAREM, LeNER-Br e DrugSeizures-Br. Os resultados demonstraram que os modelos treinados no domínio foram melhores. O modelo BERT obteve o melhor desempenho, alcançando a medida F1 de 89,39%, superando o LeNER-Br em cerca de quatro pontos na média.

O trabalho de Zanuz; Rigo (2022) apresenta os primeiros modelos BERT ajustados para REN-Legal em língua portuguesa. Dois modelos pré-treinados, `bert-base-portuguese-cased` e `bert-large-portuguese-cased`, foram utilizados no treinamento. O ajuste fino foi realizado com 10 épocas sobre os dados de treinamento e o melhor modelo, considerando a medida F1, foi selecionado para avaliar os resultados no conjunto de teste. Os modelos propostos foram comparados com resultados anteriormente obtidos para o *corpus*: LSTM+CRF, BERT-Multi, BERT-Acordaos e ELMo-Acordaos. Para o *corpus* LeNER-Br, os modelos BERT pré-treinados para língua portuguesa (Souza et al., 2020a) ajustados para o domínio alcançaram medida F1 de 91,14% no BERTimbau-Large e 90,62% no BERTimbau-Base.

O estudo de Albuquerque et al. (2023a) avaliou duas arquiteturas de *deep learning*, BiLSTM-CRF e BERT+Fine-Tuning, aplicadas a tarefa de REN-Legal em português brasileiro, utilizando os *corpora* UlyssesNER-Br e LeNER-Br. Para o UlyssesNER-Br, foi utilizado o conjunto de documentos que contém apenas projetos de lei (PL-corpus) visto que estes estão disponíveis publicamente. O modelo BERT+Fine-Tuning alcançou resultados superiores, com medida F1 de 88,27% para o LeNER-Br e 83,17% para o UlyssesNER-Br/PL-corpus, com um aumento de cerca de 7% neste último em relação à primeira versão.

O trabalho de Nunes et al. (2024b) explora como o REN pode ser aprimorado usando técnicas semi-supervisionadas. A abordagem consiste em uma estratégia de autoaprendizagem para ajustar um modelo BERT projetado para REN, usando documentos legislativos escritos em português brasileiro como estudo de caso. O UlyssesNER-Br foi utilizado para treinamento e avaliação. O BERTimbau com autoaprendizagem obteve a medida F1 de $86,70 \pm 2,28$.

²²<https://www.kaggle.com/datasets/ferraz/acordaos-tcu>



A pesquisa de Garcia et al. (2024a) investiga a aplicação de métodos de PLN na adaptação de modelos para o domínio legal em língua portuguesa. O modelo RoBERTaLexPT foi desenvolvido aplicando técnicas de deduplicação de dados (identificação e eliminação de cópias duplicadas de dados) ao modelo RoBERTa, utilizando os *corpora* LegalPT (composto de seis *corpora* legais), o *corpus* CrawlPT (composto por três *corpora* do domínio geral) e o Tesemô (Siqueira et al., 2024). Os experimentos incluíram pré-treinamento e *fine-tuning* dos modelos em tarefas de REN-Legal e classificação com *benchmarks*, comparando os resultados dos modelos genéricos com específicos. Os resultados mostraram que o RoBERTaLexPT treinado com a combinação de *corpora* legal e geral, superou outros modelos legais, incluindo o RoBERTaTimbau, alcançando medida F1 média de 90,73% para LeNER-Br e 88,56% UlyssesNER-Br (categoria).

A Tabela 13.1 apresenta os resultados obtidos para a tarefa de REN-legal em língua portuguesa utilizando modelos de linguagem neural. É possível verificar que, a partir de 2020, com o trabalho de Bonifacio et al. (2020), os trabalhos assumiram o estado da arte utilizando o modelo BERT.

Tabela 13.1: Modelos de linguagem neural estado da arte para REN-Legal em português

<i>Corpus</i>	Modelo	Medida F1
Justiça do Trabalho	ELMO+GLOVE (Castro, 2019)	93,81%
DrugSeizures-Br	BERT (Bonifacio et al., 2020)	89,39%
LeNER-Br	BERTimbau-Large+Fine-Tuning (Zanuz; Rigo, 2022)	91,14%
CDJUR-BR	BERT (Brito et al., 2023)	58,00%
UlyssesNER-Br (PL-Corpus)	RoBERTaLexPTbase (Garcia et al., 2024a)	88,56% (categoria)

As avaliações recentes de REN-Legal em português mostram que ganhos de desempenho precisam ser interpretados com cautela, sobretudo quando há risco de contaminação entre treino, validação e teste ou inconsistências de anotação nos *corpora*. Essa preocupação acompanha discussões mais amplas em PLN sobre sensibilidade a particionamentos e estabilidade de métricas (Gorman; Bedrick, 2019; Sainz et al., 2023), e foi evidenciada por Nunes (2025), ao analisar os *corpora* LeNER-Br e UlyssesNER-Br/PL-corpus em um *benchmark* com 22 modelos de diferentes arquiteturas, a partir de um processo de decontaminação que envolveu remoção de sentenças duplicadas entre partições e revisão de anotações parciais e ambíguas. No LeNER-Br, a limpeza reduziu em 11,2% o total de entidades anotadas, enquanto no UlyssesNER-Br/PL-corpus também foram removidas instâncias tanto no nível de categorias quanto no de tipos. Os efeitos foram expressivos: no LeNER-Br, 18 dos 22 modelos apresentaram diferença estatisticamente significativa entre as versões contaminada e decontaminada, com perdas de micro-F1 de até 10,14 pontos; no PL-corpus, as oscilações chegaram a variar de $-17,71$ a $+16,23$ pontos em categorias e de $-10,00$ a $+14,18$ em tipos. Em outras palavras, a contaminação não apenas pode inflar resultados absolutos, mas também modificar a leitura comparativa entre os próprios modelos.

Na versão decontaminada, os melhores resultados de micro-F1 foram obtidos por RoBERTaLexPT-base no LeNER-Br, por Legal-portuguese-R-large no UlyssesNER-Br/PL-corpus em categorias e por Albertina-100m no nível de tipos, o que sugere que o tamanho do modelo, por si só, não explica o desempenho observado. Ao mesmo tempo, a análise de erros mostrou que a REN-Legal continua marcada por dificuldades



típicas do domínio jurídico, como ambiguidades entre classes semanticamente próximas (e.g., LOCAL e ORGANIZAÇÃO, no LeNER-Br), problemas no reconhecimento de entidades compostas (principalmente em referências legais), e forte sensibilidade à qualidade das anotações (e.g., PRODUTODELEI, FUNDAMENTO e ORGANIZAÇÃO, no UlyssesNER-Br/PL-corpus). Esse quadro reforça duas implicações metodológicas centrais: a curadoria do *corpus* deve integrar o protocolo experimental desde o início, com deduplicação, revisão de anotações e tratamento de ambiguidades; e a comparação entre modelos não deve se apoiar apenas em médias de medida F1. Nesse sentido, a análise estatística com teste de Wilcoxon pareado e teste de Friedman com pós-teste de Nemenyi indica que pequenas diferenças numéricas nem sempre correspondem a vantagens consistentes, o que torna mais adequada uma avaliação que combine desempenho, robustez estatística e controle rigoroso dos dados.

13.5 Modelos de linguagem de grande escala (LLM) para REN

Conforme descrito na Seção **Tendências**, os LLMs se diferenciam dos demais modelos de linguagem neural pré-treinados devido a (i) quantidade enorme de parâmetros, o (ii) enquadramento na categoria de métodos de IA Gerativa, e (iii) habilidades emergentes, as quais não costumam ser observadas em modelos menores. Neste sentido, destacamos aqui algumas aplicações de LLMs em tarefas de PLN, independente de língua. Por fim, na Seção 13.5.1, detalhamos as iniciativas de REN-Legal em língua portuguesa.

O trabalho de Wang et al. (2021) foi um dos primeiros a investigar como o modelo de linguagem GPT-3 (Brown et al., 2020c) pode ser utilizado como uma ferramenta de rotulagem de dados de baixo custo para treinar outros modelos de PLN. Os autores conduziram um estudo empírico em nove tarefas de PLN, como sumarização, análise de sentimento e classificação de tópico. A pesquisa demonstrou que (i) o uso do GPT-3 para rotulagem de dados pode reduzir os custos de rotulagem em até 96% em comparação com a rotulagem humana, (ii) dados rotulados pelo GPT-3 podem superar o desempenho do próprio GPT-3 em configurações de *few-shot learning*, (iii) a combinação de rótulos gerados pelo GPT-3 com rótulos humanos pode melhorar ainda mais o desempenho dos modelos e, por fim, (iv) a estratégia de rotulagem ativa proposta no artigo, na qual os rótulos de baixa confiança do GPT-3 são revisados por humanos, melhora a qualidade dos dados anotados.

Especificamente para REN, Wang et al. (2023) apresentaram o GPT-NER para adaptar modelos de linguagem de grande escala à tarefa de REN, transformando-a em uma tarefa de geração de texto com *tokens* especiais para entidades. Foi introduzida uma estratégia de auto-verificação para mitigar alucinações do modelo. Foram realizados experimentos em cinco conjuntos de dados de REN para língua inglesa em domínios variados. O GPT-NER alcançou desempenho comparável aos *baselines* supervisionados utilizando *few-shot learning*, superando modelos supervisionados com dados de treinamento escassos. De maneira similar, o trabalho de Wei et al. (2024) apresenta o ChatIE, um framework de extração de informações com *few-shot learning* através de conversas com o ChatGPT²³. O ChatIE foi avaliado em três tarefas de EI, entre elas a tarefa de REN, a qual foi transformada em um problema de pergunta e resposta com duas etapas. Na primeira etapa, são identificados os tipos de elementos em uma sentença. Na segunda etapa, é realizada a EI para cada tipo de elemento identificado. Para a tarefa de REN, foram usados dois conjuntos de dados relacionados a notícias, um em inglês e outro em chinês.

²³<https://chatgpt.com/>



O ChatIE alcançou excelente desempenho, superando alguns modelos em vários conjuntos de dados.

No domínio legal, o trabalho de Hussain; Thomas (2024) investiga a aplicação de modelos de LLM na extração de entidades específicas do domínio jurídico, em documentos de casos legais indianos. O estudo avalia a eficácia de várias arquiteturas de LLMs de última geração na identificação de entidades jurídicas. O conjunto de dados utilizado para avaliação foi o InLegalNER, um *corpus* contendo 14 entidades, sendo 11 entidades específicas do domínio legal, tais como Petição, Vítima, Número do Caso, Juiz, Advogado, entre outras. Foram avaliados quatro modelos de larga escala: LLaMA 3 (AI@Meta, 2024), Gemma (Team et al., 2024), Mistral (Jiang et al., 2023) e Phi-3 (Abdin et al., 2024), utilizando a técnica de *few-shot learning*, na qual o *prompt* foi elaborado instruindo o LLM a gerar respostas em formato JSON, incluindo o texto extraído e os rótulos de entidade correspondentes. Os modelos foram avaliados utilizando precisão, revocação e medida F1 como métricas. O LLaMA 3 alcançou a medida F1 de 59,17%, Gemma de 63,53%, Mistral de 63,76%. Phi-3 de 54,40%. Os principais resultados indicam que os modelos Mistral e Gemma se destacaram em termos de equilíbrio entre precisão e revocação, essenciais para a identificação precisa de entidades.

13.5.1 LLM para REN-Legal em língua portuguesa

Visando encontrar alternativas para os métodos tradicionais de anotação manual, reduzindo os custos e esforços de anotação, o trabalho de Oliveira et al. (2024c) investiga o uso de LLM baseados em *prompt* e supervisão fraca para a tarefa de REN-Legal em português. O estudo utilizou o DODFCorpus I, disponibilizado pelo projeto KnEDLe da Universidade de Brasília. Foi utilizado o *corpus* “Atos de Contratos e Licitações”, que contém documentos públicos diários que relatam as ações do governo do Distrito Federal do Brasil, particionado em 783 atos de treinamento, 379 atos de validação e 380 atos de teste. Foram comparadas três abordagens: (i) baseada em *prompt*, com *few-shot learning*, apresentando três exemplos de dados rotulados ao GPT-3, (ii) supervisão fraca, com sistema baseado em regras e heurísticas para gerar conjuntos de dados rotulados rapidamente e (iii) anotação humana. As 783 instâncias de treinamento foram rotuladas pelo GPT-3, realizando a predição de 1.565.108 *tokens*, com um custo final de 31,30 dólares utilizando o modelo Davinci. Quatro modelos de linguagem neural pré-treinados, com ajuste fino, foram usados para a construção do modelo de REN: BERTimbau, Lener-BR, RoBERTa e DistilBERT-PT²⁴. Para as onze entidades do *corpus*, considerando a média dos quatro modelos e a métrica de medida F1, os resultados foram: 52,8% (GPT-3), 67,9% (supervisão fraca), 71,3% (anotação humana) e 64,6% (GPT-3 + supervisão fraca). Considerando a métrica de pontuação de preservação, que preserva o desempenho apresentado pelos modelos treinados por humanos, os resultados foram: 74,0% (GPT-3), 95,6% (supervisão fraca), 90,7% (GPT-3 + supervisão fraca) e 83,9% (GPT-3 + 30% de anotação humana). O BERTimbau superou o LenerBR em alguns casos, mesmo este último tendo sido adaptado para documentos legislativos brasileiros.

O trabalho de Nunes et al. (2024a) explora a aplicação do paradigma de Aprendizagem em Contexto (em inglês, *In-Context Learning* ou ICL), para melhorar o desempenho dos LLMs para o REN-Legal em língua portuguesa. O Sabiá, um LLM em português (Pires et al., 2023), foi utilizado para extrair entidades nomeadas no domínio. Dois *corpora* de REN para o domínio legal (LeNER-Br e UlyssesNER-Br) foram utilizados como entrada para a

²⁴<https://huggingface.co/adalbertojunior/distilbert-portuguese-cased>



tarefa de *prompt*. Para a seleção das amostras, foram utilizadas três heurísticas: (i) top K exemplos similares, (ii) top K exemplos similares por tipo de entidade, e (iii) K amostras aleatórias. O modelo generativo foi ensinado a catalogar as entidades e o resultado foi comparado com a anotação original. No *corpus* LeNER-Br, a heurística K similares por categorias apresentou medida F1 de 51% com oito exemplos. No resultado por tipos, três classes de entidades superaram 50% de medida F1: Pessoa com 68%, Organização com 53%, e Jurisprudência com 53%. No *corpus* UlyssesNER-Br, a mesma heurística atingiu medida F1 de 48% com oito exemplos, necessitando de onze exemplos para alcançar 51%. No resultado por tipo, três classes também obtiveram mais de 50% de medida F1: Data com 81%, Local com 69%, e Pessoa com 60%. De acordo com os autores, as classificações incorretas são resultados de uma formatação similar de termos que se referem a conceitos diferentes (Jurisprudência e Legislação, por exemplo), ambiguidade semântica (o “jornal diário catarinense” pode ser um Local ou uma Organização) e estruturas e jargões específicos do domínio (Plenário, junto com referências a jurisprudência ou legislação, separadas por um traço, confunde o modelo). A amostragem aleatória apresentou o pior desempenho.

O trabalho de Coelho et al. (2024) aborda a extração de informações no domínio jurídico brasileiro, especificamente extraindo características estruturadas de pareceres jurídicos relacionados a reclamações de consumidores. O estudo compara dois métodos: técnicas tradicionais de aprendizado supervisionado e o uso de LLM, especificamente o ChatGPT, nas tarefas de classificação de texto e de REN. O *corpus* utilizado nos experimentos contém 959 pareceres jurídicos manualmente anotados, redigidos por juízes de primeira instância no Tribunal de Justiça do Rio de Janeiro. As entidades anotadas foram: Dano Moral, Dano Material, Honorários Advocatícios e Restituição. Utilizando a versão GPT-3.5-turbo, a técnica de *prompt* envolveu três principais itens: detalhes das informações a serem extraídas, formato de saída e exemplos opcionais (*few-shot*). Os resultados obtidos foram em termos de acurácia e raiz quadrática média dos erros (em inglês, *Root Mean Squared Errors* ou RMSEs): BERT (*Fine-tuning*) 96,2% e 825,8, BERT-CRF (*Fine-tuning*) 98,2% e 639,8, BERT-LSTM (*Feature-based*) 95,4% e 738,6, BERT-LSTM-CRF (*Feature-based*) 98,9% e 277,2, ChatGPT Entity Extractor 98,4% e 1230,9. Os resultados dos experimentos mostraram que ambas as abordagens (aprendizado supervisionado tradicional e ChatGPT) alcançaram desempenhos semelhantes em termos de métricas de avaliação tradicionais. No entanto, o uso do ChatGPT reduziu substancialmente a complexidade e o tempo necessário para o processo de EI.

A Tabela 13.2 apresenta os resultados obtidos para a tarefa de REN-legal em língua portuguesa utilizando, exclusivamente, LLMs.

Tabela 13.2: LLMs estado da arte para REN-Legal em português.

<i>Corpus</i>	Modelo	Métrica
DODFCorpus I	GPT-3 Davinci + <i>few-shot learning</i> com 3 exemplos (Oliveira et al., 2024c)	52,8% (F1)
LeNER-Br	Sabiá + <i>in-context learning</i> com 8 exemplos (Nunes et al., 2024a)	51,0% (F1)
UlyssesNER-Br	Sabiá + <i>in-context learning</i> com 11 exemplos (Nunes et al., 2024a)	51,0% (F1)
Reclamações de Consumidores	GPT-3.5-turbo + <i>few-shot learning</i> (ChatGPT Entity Extractor) (Coelho et al., 2024)	98,4% (Acurácia)

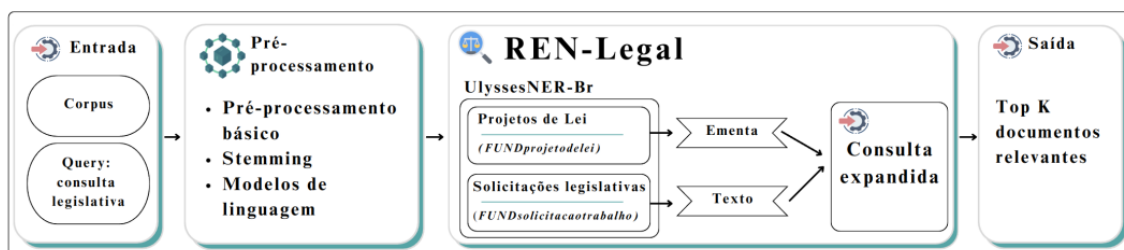


13.6 Aplicação de REN em sistema de recuperação de informação legislativa

O Capítulo 2 destaca a importância da área da Recuperação de Informação (RI) para o PLN. Pesquisas na área de *Legal Information Retrieval* ou recuperação de informação no domínio legal (RI-Legal) focam principalmente em aplicações que utilizam a linguagem judiciária, forense ou processual (Sansone; Sperlí, 2022). O Capítulo 12 destaca que a esfera de aplicação do subdomínio legislativo é a criação de textos legais, como leis e estatutos, entre outros. Sistemas RI-Legal têm sido criados ou adaptados para diversas tarefas legislativas (Heusden et al., 2023; Smywiński-Pohl et al., 2021).

Como mencionado anteriormente, o projeto Ulysses (Seção 13.2) é um conjunto de iniciativas de IA aplicados no processo legislativo brasileiro, criado pela Câmara dos Deputados, o qual recebeu recentemente novos recursos em várias frentes de trabalho²⁵. Entre estes recursos, foi utilizado o *pipeline* para um sistema de RI-Legal proposto por Souza et al. (2021a), visando automatizar tarefas de busca de documentos relevantes. REN-Legal foi inserido neste contexto.

Figura 13.3: *Pipeline* do UlyssesNERQ, adaptado de Albuquerque et al. (2024b)



O UlyssesNERQ (Albuquerque et al., 2024b) atualiza o *pipeline* do sistema de RI-Legal original, utilizando NER-Legal e a técnica de expansão de consulta. Através desta técnica, o sistema de RI modifica a *string* de consulta original, inserindo informações que possam maximizar seus resultados (Seção 2.5). A Figura 13.3 detalha o processo utilizado. O sistema recebe como entrada a *query* com solicitação de consulta legislativa e o *corpus* UlyssesNER-Br. Após a aplicação de tarefas de pré-processamento de texto na consulta, é feita a identificação de duas entidades legislativas, FUNDprojetoidelei e FUNDSolicitacaotrabalho. Em caso de identificação destas entidades, a consulta é expandida com os conteúdos dos documentos correlacionados: (i) para a entidade FUNDprojetoidelei, a consulta é expandida com a ementa dos projetos de lei encontrados; (ii) para FUNDSolicitacaotrabalho, a expansão é feita com todo o conteúdo da solicitação legislativa encontrada. Em caso de não ter sido encontrada nenhuma entidade, a consulta não é modificada. O Quadro 13.2 apresenta exemplos fictícios de consultas legislativas, com e sem expansão utilizando entidades. Ao final, são retornados os *top K* documentos mais relevantes encontrados.

²⁵Recursos disponíveis publicamente em <https://github.com/ulysses-camara>

Quadro 13.2: Exemplos de expansão de consultas legislativas utilizando REN, adaptado de (Albuquerque et al., 2024b). Expansões de consulta em negrito.

Consulta legislativa (<i>query</i>)	Entidades encontradas	Consulta expandida
Solicito revisão dos projetos de lei PLP 438/2014 e PLP 137/2015.	PLP 438/2014 PLP 137/2015	Solicito revisão dos projetos de lei PLP 438/2014 e PLP 137/2015. Dispõe sobre o procedimento para a criação, a incorporação, a fusão e o desmembramento de Municípios, nos termos do § 4º do art. 18 da Constituição Federal, altera a Lei nº 5.172, de 25 de outubro de 1966, e dá outras providências. Dispõe sobre o procedimento para a criação, a incorporação, a fusão e o desmembramento de Municípios, nos termos do § 4º do art. 18 da Constituição Federal, e dá outras providências.
Solicito complementação de parecer em função da apensação do PL 5165/2019 ao mesmo.	PL 5165/2019	Solicito complementação de parecer em função da apensação do PL 5165/2019 ao mesmo Altera o art. 3º da Lei n. 13.855, de 8 de julho de 2019, para modificar a sua cláusula de vigência
Parlamentar solicita aprovação	–	Parlamentar solicita aprovação

Para validação do UlyssesNERQ, os experimentos utilizaram 32 configurações de pré-processamento de texto nas *strings* de busca com as técnicas de *stemming*, contagem de frequência, modelos de linguagem *n-gram*, e diversas combinações. A expansão de consulta foi então avaliada aplicando três modelos de REN-Legal ajustados para o domínio: CRF (Albuquerque et al., 2022), BERT (Albuquerque et al., 2023a), e uma adaptação do modelo Bertikal (Polo et al., 2021) ajustado para as entidades legislativas. Além destes modelos, foram utilizadas duas técnicas de expansão de consulta, a primeira utilizando sinônimos, e a segunda utilizando termos relacionados ou representativos ao conteúdo, através do algoritmo RM3 (Nogueira et al., 2019). Os resultados foram verificados isoladamente e através de análises estatísticas. A métrica utilizada foi a medida de Revocação para 20 documentos (Recall@20). O melhor resultado foi alcançado pelo modelo BERT, alcançado 74,58% de revocação na análise individual, e média de $65,19\% \pm 0,0840$, ultrapassando os resultados do *pipeline* original. Este modelo foi novamente combinado com as técnicas utilizadas, alcançando o mesmo resultado individual anterior, mas com média um pouco melhor utilizando a combinação de RM3 + BERT, com $65,35\% \pm 0,0810$ de revocação. Comparando com o resultados do pipeline original, houve uma melhoria de cerca de 1,94% para os melhores resultados, e de 8,58% no resultado geral.

13.7 Considerações finais

Neste capítulo, foram detalhadas as principais iniciativas na área de reconhecimento de entidades nomeadas no domínio legal, com foco na língua portuguesa. Destacou-se o crescimento do uso de técnicas de Processamento de Linguagem Natural nesse campo, especialmente no Brasil. Foram discutidos os desafios inerentes à extração de informações de documentos jurídicos, considerando a grande quantidade e a complexidade desses textos. Além disso, foram mencionadas aplicações relevantes de inteligência artificial nas áreas jurídica e legislativa, como os projetos Victor e Ulysses, respectivamente. Apresentou-se um panorama abrangente das iniciativas em diversas línguas relacionadas ao REN no domínio legal, com suas abordagens e métodos, como aprendizado supervisionado e modelos neurais. Foram descritos *corpora* importantes para o REN-Legal em língua portuguesa e avaliação de modelos treinados especificamente para esse domínio, destacando seu desempenho em *corpora* brasileiros. Por fim, explorou-se a aplicação de REN em sistemas de recuperação de informação legislativa, como o UlyssesNERQ, que aprimora as consultas



por meio da identificação de entidades nomeadas, aumentando a eficiência na busca de documentos relevantes.

No campo de pesquisa sobre REN, diversas áreas têm se destacado com foco na melhoria contínua dos modelos. Investigações com LLMs têm se mostrado promissoras, permitindo avanços no processamento de grandes volumes de dados e na adaptação a tarefas específicas. Contudo, o desbalanceamento entre classes de entidades, no qual certas categorias apresentam uma representação desproporcionalmente menor nos dados, continua sendo um desafio. Esse fenômeno, aliado à existência de entidades pouco representativas, podem impactar negativamente o desempenho dos modelos. Estratégias como *Data augmentation*, que visa aumentar a quantidade de dados de treinamento de forma sintética, e *Active learning*, que melhora a eficiência do aprendizado ao selecionar amostras mais informativas, têm sido exploradas para mitigar esses problemas. Por fim, a dependência de domínio, dada a especificidade dos textos legais, impõe a necessidade de adaptar os modelos a contextos muito particulares, o que exige uma personalização contínua dos modelos para cada subdomínio do Direito.

Agradecimentos

Gostaríamos de expressar nossa gratidão à Câmara dos Deputados Brasileira, ao Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (ICMC/USP) e ao Instituto Nacional de Inteligência Artificial (IAIA) por possibilitarem o acesso a informações cruciais para o desenvolvimento da pesquisa, assim como a todos os colaboradores do projeto Ulysses. Além disso, agradecemos também à colaboração científica dos autores dos trabalhos citados neste capítulo, que gentilmente nos responderam com dados e recursos de suas pesquisas.



Capítulo 14

Geração Automática de Ementas Jurídicas

*Kenzo Sakiyama
Roseval Malaquias Junior
Raphael Montanari
Rodrigo Nogueira
Roseli A. F. Romero*

Publicado em: 20/11/2024

 <https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/>

14.1 Introdução

14.1.1 Contextualização

Segundo dados do Conselho Nacional de Justiça (CNJ) (Justiça - CNJ, 24DC), ao final de 2022, havia 81,4 milhões de processos em trânsito no judiciário brasileiro, um aumento de 7,5% em relação ao ano anterior. O relatório também destaca que, nos casos mais extremos, o tempo de análise pode levar vários anos. Além do esforço humano necessário para a análise e elaboração dos documentos, é preciso alocar recursos financeiros para sua manutenção ao longo de todo o processo. Neste cenário, o aprimoramento de sistemas computacionais jurídicos, com a automação do fluxo de documentos, torna-se essencial para agilizar a tramitação no judiciário brasileiro.

Quadro 14.1: Exemplo de ementa. A verbetação está destacada em negrito.

DIREITO CONSTITUCIONAL, TRIBUTÁRIO, PREVIDENCIÁRIO E PROCESSUAL CIVIL. CONTRIBUIÇÃO SOCIAL. ART. 2º DA LEI Nº 7.856, DE 25 DE OUTUBRO DE 1989. MAJORAÇÃO DE 8% PARA 10%, DA ALÍQUOTA DA CONTRIBUIÇÃO SOCIAL. SUCUMBÊNCIA. HONORÁRIOS ADVOCATÍCIOS. CUSTAS PROCESSUAIS.

1. Firmou-se em Plenário do Supremo Tribunal Federal o entendimento no sentido de que, em se tratando de “lei de conversão da Medida Provisória nº 86, de 25 de setembro de 1989, da data da edição desta é que flui o prazo de noventa dias previsto no art. 195, § 6º, da CF, o qual, no caso, teve por termo final o dia 24 de dezembro do mesmo ano, possibilitando o cálculo do tributo, pela nova alíquota, sobre o lucro da recorrente, apurado no balanço do próprio exercício de 1989” (RE 197.790-6-MG, Rel. Min. ILMAR GALVÃO).
2. Adotados os fundamentos desse precedente, o RE, no caso, é conhecido e provido.
3. As autoras ficaram vencedoras, quanto a outros pedidos, nas instâncias ordinárias, sem que o R.E. abordasse esses pontos. Sendo maior a sucumbência da ré, pagará honorários advocatícios às autoras.
4. Custas em proporção.

Este Capítulo trata da geração automática de **ementas**, documentos amplamente utilizados em tribunais brasileiros, cuja função é fornecer um resumo de processos judiciais. As



ementas oferecem uma representação sucinta e simplificada das decisões, permitindo que profissionais do Direito compreendam rapidamente o conteúdo sem a necessidade de ler o acórdão completo. Conforme demonstra o exemplo no Quadro 14.1, estes documentos usualmente contêm dois elementos principais: **verbetações** e **parágrafos enumerados**.

As **verbetações** são cabeçalhos constituídos por frases curtas, em caixa-alta, que descrevem os assuntos mais relevantes da decisão. Além de resumir o conteúdo, elas auxiliam na tarefa de busca e recuperação de jurisprudências (precedentes) (Guimarães; Santos, 2016). Por outro lado, os **parágrafos enumerados** expõem as teses da decisão, sendo que o último apresenta geralmente a conclusão do tribunal. Essa estrutura contribui para a agilidade no acesso à informação jurídica, facilitando a consulta à jurisprudência.

Vale destacar que, devido à sua estrutura e estilo, as verbetações assemelham-se a palavras-chave e a textos sumarizados. No entanto, diferentemente das palavras-chave, elas podem consistir em frases curtas. Além disto, enquanto resumos são redigidos de maneira fluida, as verbetações seguem um estilo linguístico único e rígido em comparação aos resumos. Um desafio adicional na redação das verbetações é que muitos dos termos utilizados não aparecem no restante da ementa, tornando sua produção uma tarefa que vai além da simples extração de informação.

Considerando as características do sistema judiciário brasileiro, os componentes das ementas e sua grande disponibilidade no Brasil, o objetivo deste Capítulo é apresentar métodos de PLN que propiciam a automatização da escrita de verbetações empregando os parágrafos enumerados como entradas para um modelo supervisionado generativo. Em específico, como Transformers (Brown et al., 2020c; Carmo et al., 2020b; Scao et al., 2022; Zhang et al., 2022) se mostraram efetivos em tarefas de geração de texto, neste capítulo os mesmos são avaliados na tarefa de geração de verbetações. O Quadro 14.2 ilustra exemplos de verbetações geradas por Transformers, nos quais fica evidente a capacidade desse modelo generativo, obtendo resultados semelhantes à verbetações escritas por especialistas.

Quadro 14.2: Exemplos de verbetações geradas por Transformers.

Original: PROCESSUAL CIVIL. EMBARGOS DE DECLARAÇÃO NO RECURSO ESPECIAL. ALEGAÇÃO DE OMISSÃO. RESSALVAS DE FUNDAMENTAÇÃO CONTIDAS NO VOTO-VISTA. OMISSÃO NÃO CARACTERIZADA.

Gerada: PROCESSUAL CIVIL. EMBARGOS DE DECLARAÇÃO. OMISSÃO. VÍCIO INEXISTENTE. REJEIÇÃO.

Original: AGRAVO REGIMENTAL NO HABEAS CORPUS. TRÁFICO DE DROGAS. PRISÃO PREVENTIVA. PERICULUM LIBERTATIS. CRISE MUNDIAL DE COVID-19. NÃO DEMONSTRADA A EXISTÊNCIA DE RISCO À INTEGRIDADE FÍSICA DO ENCARCERADO. AGRAVO REGIMENTAL NÃO PROVIDO.

Gerada: AGRAVO REGIMENTAL NO HABEAS CORPUS. TRÁFICO DE DROGAS. PRISÃO PREVENTIVA. PERICULUM LIBERTATIS. AGRAVO REGIMENTAL NÃO PROVIDO.

Este Capítulo apresenta mais uma aplicação de PLN na área do Direito, assim como as apresentadas no Capítulo 12. São utilizados modelos de linguagem (descritos no Capítulo Modelos de linguagem) em uma tarefa de geração de texto, similar à tarefa de sumarização (Capítulo 5). A metodologia proposta descreve como a coleta de dados foi realizada, apresenta o modelo transformer utilizado e também a técnica de validação e métrica adotadas. Além disto, emprega-se conceitos de Recuperação de Informação (Capítulo 2) para avaliação final das verbetações geradas.



14.1.2 Organização do Capítulo

Este Capítulo está organizado da seguinte forma. A Seção 14.2.1 introduz conceitos preliminares relacionados à aplicação. Em seguida, a Seção 14.3 descreve a metodologia utilizada para a geração de verbetações. A Seção 14.4 detalha a metodologia adotada. Na sequência, os resultados obtidos são apresentados e discutidos na Seção 14.5. Por fim, conclusão é apresentada na Seção 14.6.

14.2 Conceitos Preliminares

14.2.1 Transferência de Aprendizado e Ajuste Fino

Uma propriedade importante para a popularidade de Transformers em PLN é a **Transferência de Aprendizado** (TA). Como o próprio nome sugere, a técnica visa transferir conhecimento de um domínio para outro. Em outras palavras, busca-se generalizar o conhecimento adquirido em um domínio A para um domínio B . É importante notar que essa transferência é mais vantajosa se os domínios estiverem de alguma forma relacionados (Zhuang et al., 2020).

Esta propriedade é útil em casos em que há uma grande diferença na quantidade de dados disponíveis no domínio e na tarefa alvo. Por exemplo, pensando no contexto de PLN, há uma abundância de textos disponíveis publicamente na web, porém, apenas uma fração é relacionada ao Direito (uma fração ainda menor, se considerarmos textos em português). Neste caso, a TA permite que um modelo seja treinado em um domínio com dados abundantes e, depois, treinado novamente no conjunto de dados destinado à tarefa alvo. Este segundo treinamento, por sua vez, é chamado de **Ajuste Fino**.

Um exemplo de TA no contexto de PLN utilizando Transformers é encontrado no modelo BERTimbau (Souza et al., 2020a) (Transformer codificador). Inicialmente, ele foi treinado no *corpus* genérico brWaC (Wagner Filho et al., 2018). Ambos, modelo e *corpus*, foram apresentados no Capítulo Modelos de linguagem. O treinamento inicial foi feito em um grande volume de texto, visando aprender estruturas linguísticas comuns da língua portuguesa. Este treinamento é baseado em modelagem de linguagem (também discutido no Capítulo Modelos de linguagem). Similar ao modelo BERT, o modelo BERTimbau, gera representações para os *tokens* que podem ser utilizadas em diferentes tarefas de PLN, como classificação, reconhecimento de entidades nomeadas, agrupamentos etc.

Vale notar que o treinamento inicial não necessariamente corresponde à tarefa final na qual o modelo de aprendizado de máquina será empregado. Além disto, o pré-treinamento possibilita um bom desempenho na tarefa final, mesmo que esta não possua uma quantidade abundante de dados para o ajuste fino. Além disso, após essa etapa inicial de treinamento, os modelos pré-treinados podem ser disponibilizados publicamente¹². Este fato é interessante, pois possibilita que outros pesquisadores tenham à sua disposição modelos pré-treinados, prontos para serem utilizados em suas tarefas alvo de interesse.

14.2.2 Bilingual Evaluation Under-study (BLEU)

Para avaliar quantitativamente sistemas de geração de texto, muitas métricas foram propostas ao longo dos anos, como por exemplo a medida BLEU.

¹<https://pytorch.org/hub/>

²<https://huggingface.co/models>



A métrica BLEU (Papineni et al., 2002) foi proposta originalmente para avaliar sistemas de tradução. De maneira simplificada, seu funcionamento envolve analisar a sobreposição entre as palavras (ou *tokens*) no texto gerado e aquelas nos textos de referência humanos. Na prática, as palavras são representadas por n-gramas (usualmente de uma a quatro palavras), após a tokenização dos textos. Os valores da métrica estão em uma faixa de 0 a 1, na qual uma pontuação mais próxima de 1 indica uma melhor qualidade no texto gerado.

Outra característica interessante é que a métrica considera todo o *corpus* de avaliação para seu cálculo. Assim, as estatísticas são obtidas a partir de todos os textos gerados e referências, sem realizar qualquer tipo de agregação, como uma média, para determinar a pontuação final.

Formalmente, o BLEU está definido da seguinte forma:

$$BLEU = Penalidade \times \left(\prod_{i=1}^4 Precisão_i \right)^{1/4} \quad (14.1)$$

$$Penalidade = \min \left(1, \exp \left(1 - \frac{\text{tamanho da referência}}{\text{tamanho do texto gerado}} \right) \right) \quad (14.2)$$

$$Precisão_i = \frac{\sum_{tg \in \text{Gerados}} \sum_{i \in tg} \min(c_{gerado}^i, c_{ref}^i)}{C_t^i = \sum_{tg' \in \text{Gerados}} \sum_{i' \in tg'} c_{gerado}^{i'}} \quad (14.3)$$

na qual i define o i -ésimo n-grama, tg corresponde a um texto gerado, c_{gerado}^i corresponde ao número de ocorrências simultâneas do i -grama na referência e no texto gerado, c_{ref}^i corresponde ao número de ocorrências da i -grama no texto de referência, $c_{tg}^{i'}$ denota a quantidade de n-gramas no texto gerado e C_t^i informa o número total de i -gramas dos textos gerados.

Na fórmula apresentada, o componente de Penalidade (Equação 14.2) penaliza a produção de textos curtos, já que seu valor diminui à medida que os textos gerados são menores. Já que o número de termos no texto gerado influencia o denominador da Equação 14.3, quando esse valor é pequeno, facilita-se maximizar a precisão. Logo, a penalidade atua para incentivar a geração de textos mais longos. Note que os tamanhos são calculados considerando todos os exemplos, tanto os gerados quanto os de referência.

Em sequência, a Equação 14.3 ilustra um cálculo de precisão no qual verifica-se qual porcentagem dos n-gramas ocorre tanto no texto gerado quanto no texto de referência.

A Equação 14.1 define o BLEU, combinando componentes mencionados anteriormente. Como se pode observar, a métrica BLEU é a média harmônica das precisões dos n-gramas (tipicamente, entre 1 e 4), com uma penalização pelo tamanho do texto gerado.

Em suma, o BLEU destaca-se por sua simplicidade e rapidez de cálculo, além de estar correlacionado positivamente com as avaliações humanas (Papineni et al., 2002). No entanto, é importante reforçar que sua principal limitação está na falta de consideração dos aspectos semânticos dos textos comparados, já que sua metodologia se baseia no casamento de n-gramas. Além disso, a métrica é sensível ao método de tokenização utilizado.

14.3 Geração de verbetações

Neste Capítulo, serão apresentados os componentes da metodologia proposta para o desenvolvimento desta aplicação. Iniciando pelos dados utilizados, a Seção 14.3.1 discute a coleta dos documentos. Em seguida, a metodologia empregada para a geração automática



de verbetações e sua avaliação é discutida na Seção 14.3.2. Mais detalhes a respeito da metodologia podem ser encontrados em (Sakiyama, 2023).

14.3.1 Aquisição de dados

Em 2022, como parte de um movimento em direção à transparência administrativa, o Supremo Tribunal de Justiça (STJ) lançou a plataforma Dados Abertos³. Este domínio público destina-se ao compartilhamento de decisões judiciais de diversos tribunais do Brasil, sendo estas julgadas por ministros do STJ. Além de promover a transparência, a plataforma foi criada visando reduzir o uso de *web scrappers* no portal oficial do STJ e estimular a pesquisa e o desenvolvimento de ferramentas baseadas em inteligência artificial que dependem da disponibilidade de dados.

As ementas coletadas estão classificadas em seções. Estas representam as seções do STJ e se diferenciam pelos temas jurídicos julgados. Por exemplo: impostos (primeira seção), comércio (segunda seção) e crimes no geral (terceira seção) (STJ, 24DC). Desta forma, os documentos obtidos abrangem uma ampla variedade de temas do contexto jurídico brasileiro. Para o desenvolvimento desta pesquisa foram coletados um total de 726.384 documentos da plataforma (a coleta ocorreu em agosto de 2022) e foram analisadas as ementas dos processos obtidos.

14.3.2 Geração textual

Nesta Seção, é apresentada a metodologia utilizada para geração das verbetações e sua avaliação.

14.3.2.1 Preparação de exemplos para geração de texto

Para a preparação de entradas para o modelo generativo escolhido, verbetações e parágrafos enumerados foram separados, por meio da identificação de frases iniciadas por letras maiúsculas. Assim, as verbetações originais extraídas passam a compor o conjunto de referência (conjunto ouro ou *gold standard*) a ser usado para avaliar a qualidade das verbetações geradas pelas estratégias estudadas.

Feito o processamento adicional, o *corpus* final foi dividido em três conjuntos: treino (70%), validação (10%) e teste (20%). Tal divisão foi feita de forma estratificada, para preservar as proporções das origens (seções de julgamento) dos documentos. Este procedimento foi feito visando uma avaliação utilizando a técnica de validação cruzada (mais detalhes no Capítulo Avaliação de tecnologias de linguagem).

A fim de ilustrar aos leitores os tamanhos dos parágrafos enumerados e das verbetações, na Tabela 14.1 é sumarizada a análise estatística descritiva da variação dos tamanhos dos *tokens* presentes no conjunto de treinamento. Dentre as medidas foram calculadas a média, o desvio padrão e os 3 quartis: Q1, Q2 e Q3. O conjunto em questão possui cerca de 1,6 milhão de *tokens*. Os parágrafos enumerados têm, em média, 203,26 *tokens*, enquanto as verbetações têm, em média, 55,84 *tokens*. Os valores do desvio padrão revelam que os tamanhos dos parágrafos enumerados e das verbetações variam consideravelmente.

³<https://dadosabertos.web.stj.jus.br/>



Tabela 14.1: Variação dos tamanhos dos parágrafos enumerados e das verbetações com base em medidas da estatística descritiva para *tokens* do conjunto de treino (78.375 exemplos).

	Média	Desvio Padrão	Q1	Q2	Q3
Parágrafos enumerados	203,262	183,332	92	155	253
Verbetação	55,842	32,136	34	49	69

14.3.2.2 Transformers para geração de verbetações

Como mencionado anteriormente, as verbetações podem conter termos que não estão presentes no corpo da ementa (parágrafos enumerados). De fato, ao analisar os documentos do conjunto de validação, observou-se que apenas cerca de 10% dos termos das verbetações estão contidos no restante do texto. Sendo assim, optou-se por abordar a geração de verbetações como sendo uma geração de sequência-a-sequência (ou texto-para-texto). Utilizou-se o conteúdo dos parágrafos enumerados como entrada para os modelos Transformers, enquanto que as verbetações originais foram consideradas como as saídas desejadas.

Estudos anteriores (Carmo et al., 2020b; Rosa et al., 2021a; Souza et al., 2020a) apontam que modelos pré-treinados para a linguagem da tarefa tendem a superar os modelos multilíngues nas mesmas tarefas. Em experimentos detalhados em (Sakiyama, 2023), o modelo PTT5 (Carmo et al., 2020b)⁴, ajustado para português brasileiro, destacou-se como o mais eficaz, e seus resultados serão reportados aqui. Este modelo foi pré-treinado no *corpus* brWaC e foi utilizada sua versão base (com 220M de parâmetros). Tendo treinado o modelo PTT5, a geração de verbetações foi feita usando uma simples decodificação gulosa (que consiste sempre em escolher o *token* mais provável na geração).

14.3.2.3 Avaliação dos textos gerados

A avaliação dos textos gerados pelo modelo de língua PTT5 foi feita empregando-se a metodologia descrita no Capítulo [Avaliação de tecnologias de linguagem](#). Utilizou-se a técnica de amostragem *holdout* para validação e a métrica BLEU (*Bilingual Evaluation Under-study*) para quantificar a qualidade dos textos gerados.

A fim de aprofundar a avaliação, foram realizadas outras análises comparativas entre o texto gerado e as entradas originais. Por exemplo, os tamanhos das entradas geradas pelo melhor modelo foram comparados com os das entradas originais, bem como a quantidade de palavras (*tokens*) copiadas e geradas. Palavras ‘novas’ (ou geradas) são aquelas que não aparecem nos parágrafos enumerados, utilizados na geração de verbetações.

14.4 Avaliação utilizando RI

Para a avaliação final das verbetações geradas, elas foram concatenadas aos seus documentos originais para simular um caso de uso real em uma tarefa de Recuperação de Informação (RI). Foram empregadas tanto a decodificação gulosa quanto a decodificação por amostragem na geração das verbetações nesta avaliação. Mais detalhes a respeito da avaliação experimental podem ser encontrados em (Sakiyama, 2023).

⁴<https://huggingface.co/unicamp-dl/ptt5-base-portuguese-vocab>



14.4.1 Formulação da tarefa

Os documentos apresentados na Seção 14.3.1, juntamente com o texto da ementa, contêm metadados úteis, entre eles, o tema da decisão. Tais temas, ou temas de recursos repetitivos, consistem em identificadores únicos que correspondem a questões jurídicas comuns. Cada documento pode estar associado a mais de um tema, e há mais de mil temas distintos listados no STJ⁵. Exemplos desses temas estão ilustrados no Quadro 14.3.

Quadro 14.3: Exemplos de temas de recursos repetitivos, listados pelo STJ.

Tema 192: Incidência da pensão alimentícia sobre o décimo terceiro salário e o terço constitucional de férias .

Tema 312: Prazo para restituição ao consorciado desistente das parcelas pagas ao grupo de consórcio .

Tema 1166: Natureza material do crime de apropriação indébita previdenciária e sua consumação com a constituição definitiva do crédito tributário na via administrativa.

Com base nos temas, foi utilizada a definição binária de relevância para formular a seguinte tarefa de RI: dado um documento de consulta Q , os documentos relevantes R devem pertencer ao mesmo tema de Q . Esta formulação simula o caso de uso em que um advogado busca processos semelhantes a um caso base, por exemplo. A formulação apresentada é semelhante à utilizada por Ostendorff et al. (2021), na qual foram criados pares de relevância (documento de consulta e documento relevante) usando decisões da Suprema Corte dos Estados Unidos. As decisões foram categorizadas por livros de casos ou categorias pré-definidas.

Em resumo, na avaliação proposta, ementas são usadas para buscar outras ementas. Para imitar o cenário real nos experimentos, foram empregadas ementas completas (incluindo verbetes e parágrafos enumerados). Foram utilizadas ementas completas em todos os experimentos, exceto o primeiro, que será discutido na próxima Seção.

Entre as ementas coletadas, apenas 801 possuíam anotações de tema, com 99,3% apresentando valores ausentes. Estas ementas foram removidas do conjunto de treinamento dos geradores de texto e utilizadas para compor pares de consulta e documentos relevantes nos experimentos com RI. Foram filtrados temas que apareceram pelo menos duas vezes, resultando em 482 consultas.

A Figura 14.1 mostra uma visualização do número de documentos relevantes (com o mesmo tema) por documento de consulta. Observa-se que a maioria das consultas possui entre 1 e 3 documentos relevantes. No caso mais extremo, uma consulta pode ter até 10 documentos relevantes.

Para criar o *corpus* final de busca para a tarefa mencionada, foram combinados os 801 documentos com tema ao conjunto de teste discutido na Seção 14.3.2.1, totalizando 23.194 documentos. Esta combinação foi feita visando tornar a tarefa de RI mais desafiadora, pois a mesma pode incluir falsos negativos ao *corpus* de busca, que acabam prejudicando as métricas de RI. Falsos negativos são documentos que têm o mesmo tema da consulta, mas são considerados irrelevantes por não possuírem anotação de tema (dado ausente).

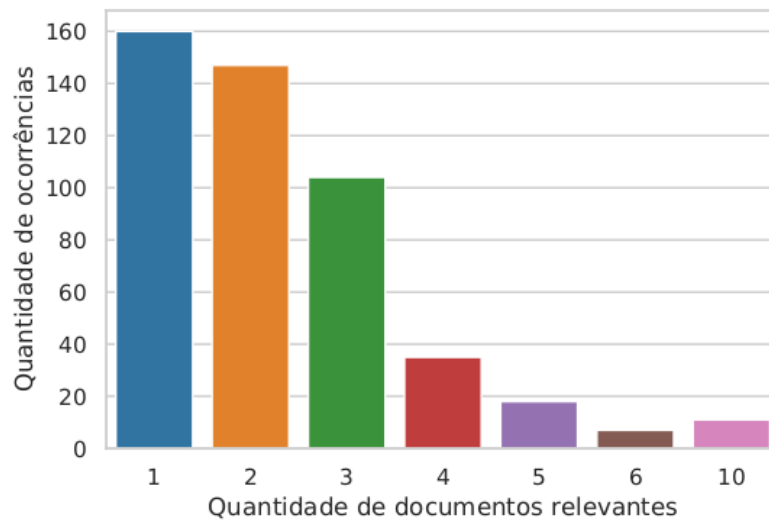
14.4.2 Experimentos realizados

A avaliação baseada em RI foi conduzida por meio de dois experimentos distintos. Todos os experimentos utilizaram métodos de decodificação simples (decodificação gulosa) e mé-

⁵<https://scon.stj.jus.br/SCON/recrep/>



Figura 14.1: Visualização do número de documentos relevantes (de mesmo tema) por documento de consulta.



todos não determinísticos (decodificação com amostragem). Seus detalhes serão descritos a seguir.

1. **Comparação de documentos com e sem verbetações:** a influência das verbetações na tarefa de RI foi avaliada, analisando-se métricas de RI para buscas, usando documentos com e sem suas verbetações originais, tanto para documentos de consulta quanto para documentos do *corpus* de busca. Este experimento teve como objetivo validar o papel das verbetações no contexto de uma tarefa de RI.
2. **Avaliação das verbetações geradas:** neste experimento, foi investigado o efeito de usar verbetações geradas no lugar das originais. O texto gerado foi concatenado ao início de ambos os documentos de consulta e dos documentos do *corpus* de busca. O objetivo deste experimento é verificar se as verbetações artificiais melhoram as métricas de RI em comparação com a ausência de verbetações. Para este experimento, foram usadas verbetações geradas com decodificação gulosa.

14.4.3 Métodos de RI avaliados

Dois métodos tradicionais de RI para a tarefa descrita anteriormente: TF-IDF e BM25 (Robertson; Walker, 1999) foram considerados para a avaliação. Tais métodos são amplamente utilizados em sistemas de busca populares (como Lucene⁶), e são frequentemente adotados como *base* em comparações de desempenho de algoritmos (Pradeep et al., 2020; Rosa et al., 2021b). Além disto, conforme investigado por estudos anteriores (Lima et al., 2021; Mandal et al., 2021), métodos de representação esparsa tendem a apresentar melhor desempenho em tarefas similares no domínio jurídico. Tendo apresentado os modelos investigados, a seguir serão apresentados mais detalhes sobre suas execuções.

⁶<https://lucene.apache.org/>



14.4.4 Ordenação dos documentos

Ao utilizar o método TF-IDF para realizar as buscas, calculou-se a similaridade de cosseno entre consultas e documentos e os documentos do *corpus* foram ordenados a partir de suas semelhanças em ordem decrescente. Já o método BM25 utiliza o Princípio de Classificação por Probabilidade (*Probability Ranking Principle*) para estimar a relevância de um documento para uma consulta (Crestani et al., 1998). Logo, os documentos foram ordenados com base em suas relevâncias estimadas.

14.4.5 Métricas avaliadas

Para avaliar de forma quantitativa o desempenho da busca e, em adição, a qualidade das verbetações geradas, foram computadas métricas tradicionais de RI. As métricas utilizadas foram: *Mean Reciprocal Rank* (MRR), *Normalized Discounted Cumulative Gain* (NDCG), Revocação, Precisão e *Mean Average Precision* (MAP).

Utilizou-se os 10 primeiros documentos ordenados para o cálculo das métricas mencionadas. Este limite foi escolhido para simular um cenário restritivo, em que o usuário investiga somente até 10 documentos. Conforme (Russell-Rose et al., 2018), profissionais do Direito geralmente aceitam analisar até 50 documentos em suas consultas. Portanto, este estudo avalia um cenário ainda mais desafiador do que o mencionado pelos autores.

14.5 Resultados e discussões

Esta Seção apresenta e discute os resultados obtidos na geração de verbetações e sua avaliação usando RI.

14.5.1 Comparação entre verbetações originais e geradas

Como apontado na Seção 14.3.2.2, as verbetações foram geradas usando o modelo PTT5, o qual obteve uma pontuação BLEU de 38,607 no conjunto de teste preparado. Nesta Seção, serão mostrados exemplos (Quadros 14.4, 14.5 e 14.6) dos textos gerados e os mesmos também serão comparados com as verbetações originais.

Ao avaliá-las qualitativamente (Quadro 14.4), observa-se que, com pontuações BLEU próximas a 40%, as verbetações geradas não apresentam erros ortográficos ou léxicos. Elas simulam o estilo de escrita das verbetações originais e são bastante similares a elas. Além disto, fica evidente que as verbetações geradas não são cópias perfeitas das originais, uma vez que nem todos os termos das verbetações originais estão nas geradas. Assim, os resultados da geração textual com Transformers foram muito positivos, se consideramos o tamanho do conjunto de treinamento utilizado neste caso (menos de 100 mil exemplos).

Quadro 14.4: Exemplos de verbetações geradas, destacando-se termos em comum (idênticos) entre as verbetações.

Original: AGRAVO REGIMENTAL NO HABEAS CORPUS . TRÁFICO DE DROGAS. PRISÃO PREVENTIVA. PERICULUM LIBERTATIS. CRISE MUNDIAL DE COVID-19. NÃO DEMONSTRADA A EXISTÊNCIA DE RISCO À INTEGRIDADE FÍSICA DO ENCARCERADO. AGRAVO REGIMENTAL NÃO PROVIDO.

Gerada: AGRAVO REGIMENTAL NO HABEAS CORPUS. TRÁFICO DE DROGAS. PRISÃO PREVENTIVA. PERICULUM LIBERTATIS. AGRAVO REGIMENTAL NÃO PROVIDO.

Original: ADMINISTRATIVO. RECURSO ESPECIAL REPETITIVO. SERVIDOR PÚBLICO.



ARTIGO 46, CAPUT, DA LEI N. 8.112/1990. TESE DEFINIDA NO TEMA 531/STJ. AUSÊNCIA DE ALCANCE NOS CASOS DE PAGAMENTO INDEVIDO DECORRENTE DE ERRO DE CÁLCULO OU OPERACIONAL DA ADMINISTRAÇÃO PÚBLICA. POSSIBILIDADE DE DEVOLUÇÃO. SALVO INEQUÍVOCA PRESENÇA DA BOA-FÉ OBJETIVA.

Gerada: ADMINISTRATIVO. RECURSO ESPECIAL REPETITIVO. TEMA 531/STJ. ERRO DE CÁLCULO OU OPERACIONAL. PAGAMENTO INDEVIDO. BOA-FÉ OBJETIVA. DEVOLUÇÃO. DESCABIMENTO.

Quadro 14.5: Mais exemplos de verbetações geradas. Termos em comum foram destacados em verde. Parafrases ou sinônimos foram destacados em amarelo.

Original: CONFLITO NEGATIVO DE COMPETÊNCIA. JUSTIÇA COMUM ESTADUAL E JUSTIÇA DO TRABALHO. AÇÃO DE OBRIGAÇÃO DE FAZER C/C COBRANÇA. CONTRIBUIÇÃO SINDICAL DE SERVIDORES PÚBLICOS MUNICIPAIS. ART. 114, III, DA CF/1988 COM REDAÇÃO DADA PELA EC 45/2004. COMPETÊNCIA DA JUSTIÇA DO TRABALHO. PRECEDENTES.

Gerada: CONFLITO NEGATIVO DE COMPETÊNCIA CONTRIBUIÇÃO SINDICAL. AGENTE PÚBLICO ESTATUTÁRIO. ART. 114, III, DA CONSTITUIÇÃO FEDERAL DE 1988. EMENDA CONSTITUCIONAL 45/2004. COMPETÊNCIA DA JUSTIÇA DO TRABALHO.

Quadro 14.6: Exemplo de verbetação em que são introduzidos termos não presentes na verbetação original.

Original: TRIBUTÁRIO E PROCESSUAL CIVIL. RECURSO ESPECIAL. INTERPOSIÇÃO DE RECURSO, TEMPESTIVAMENTE, POR MEIO DE FAC-SIMILE. AUSÊNCIA DE APRESENTAÇÃO DA PETIÇÃO ORIGINAL, NO PRAZO PREVISTO NO ART. 2º DA LEI 9.800/99. RECURSO ESPECIAL INTERPOSTO VIA E-MAIL. INADMISSIBILIDADE. NÃO EQUIPARAÇÃO AO FAC-SIMILE. PRECEDENTES DO STJ. RECURSO ESPECIAL NÃO CONHECIDO.

Gerada: : TRIBUTÁRIO E PROCESSUAL CIVIL. RECURSO ESPECIAL, POR FAC-SIMILE. INTERPOSIÇÃO, NA ORIGEM, DE AGRAVO DE INSTRUMENTO, POR INTERPOSIÇÃO DE RECURSO POR MEIO DE FAC-SIMILE, ORIGINAIS APRESENTADOS DENTRO DO PRAZO LEGAL. IMPOSSIBILIDADE. ART. 2º DA LEI 9.800/99. RECURSO INTENESTIVO. PEÇA INCOMPLETA. NÃO APRESENTAÇÃO. PRESCRIÇÃO. PRAZO PARA INTERPOSIÇÃO DE RECURSO. PRECEDENTES DO STJ. ART. 543-C DO CPC/73. TERMO INICIAL. SÚMULA 83/STJ. AGRAVO INTERNO IMPROVIDO.

Dando prosseguimento às comparações, ao analisarmos o Quadro 14.5, é possível observar uma característica interessante. Houve casos em que o PTT5 gerou paráfrases ou expandiu siglas presentes na verbetação original. Tal fato é facilmente perceptível no Quadro 14.5, observando os termos “SERVIDOR PÚBLICO” e “AGENTE PÚBLICO”. Por outro lado, siglas como “CF/1998” foram expandidas para “CONSTITUIÇÃO FEDERAL DE 1998”, assim como “EC” foi expandida para “EMENDA CONSTITUCIONAL”. Os comportamentos observados são justificados pelo funcionamento de modelos de linguagem baseados em Transformers, uma vez que, durante a geração de texto, os mesmos tendem a gerar *tokens* que aparecem em contextos semelhantes.

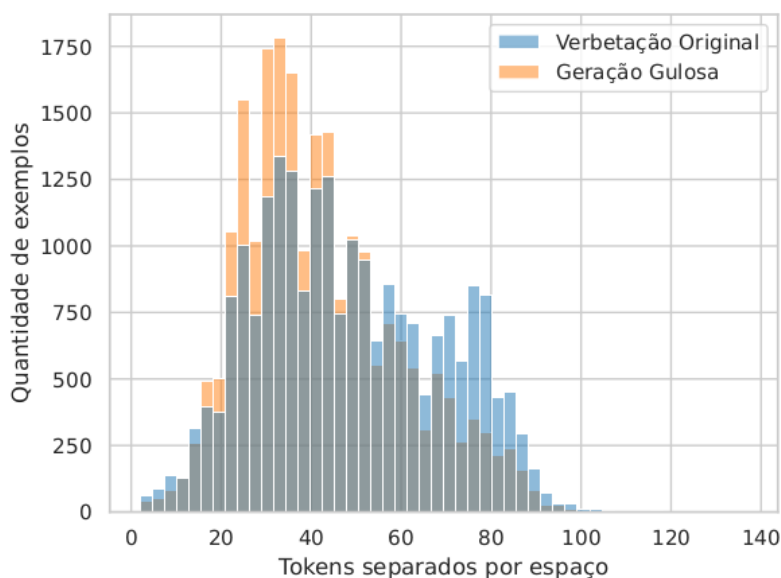
Como última discussão de exemplos, o Quadro 14.6 ilustra potenciais alucinações do modelo avaliado. Como o modelo PTT5 está na família de Transformers generativos, o mesmo não está isento de alucinações, como é o caso dos mais conhecidos (GPT’s, por

exemplo). Dentro do contexto desta aplicação e focando no Quadro 14.6, é possível notar a presença de termos e citações de artigos que não constam no texto original.

Embora esses termos possam surgir em contextos parecidos, não se deve ignorar a possibilidade de que tenham sido gerados por alucinações do Transformer. Sem o conhecimento específico do domínio, não é possível assegurar que tais termos “novos” estejam realmente associados ao texto de entrada. Pode-se apenas supor que eles ocorrem em contextos semelhantes. Assim, existe sim o risco de adicionar textos factualmente incorretos às verbetações geradas, devido ao processo generativo.

A investigação da métrica BLEU permite estimar a quantidade de termos em comum entre as verbetações de referências e as geradas, ao comparar a interseção de n-gramas em ambas. Como a métrica é computada considerando o *corpus* de teste inteiro, ela apresenta uma boa estimativa do casamento de n-gramas entre todas as verbetações originais e geradas. Para aprofundar a comparação entre as verbetações, na Figura 14.2 é apresentada uma comparação entre o número de *tokens* das verbetações originais e das verbetações geradas com decodificação gulosa.

Figura 14.2: Histogramas de comparação entre a quantidade de *tokens* das verbetações originais e das verbetações geradas.



Embora as distribuições apresentadas pelos dois histogramas sejam semelhantes, fica evidente que as verbetações geradas tendem a ter uma concentração maior de exemplos abaixo de 60 *tokens*. Além disto, ao analisar todas as verbetações mostradas anteriormente (Quadros 14.2, 14.4, 14.5 e 14.6) fica evidente que as mesmas tendem a ter menos palavras do que as referências. Assim, podemos concluir que as verbetações geradas tendem a ser mais curtas que as originais.

14.5.2 Experimentos com RI

A Seção anterior apresentou comparações qualitativas entre as verbetações geradas pelo método PTT5 e as originais. Resta verificar como as verbetações originais e geradas se comportam na avaliação, usando um caso real de RI descrito na Seção 14.4. Vale observar

que as métricas mostradas em todas as tabelas desta Seção consistem em médias obtidas para cada métrica considerando as 482 consultas.

14.5.2.1 Comparação de documentos com e sem verbetações

A Tabela 14.2 mostra as métricas obtidas ao realizar o primeiro experimento com documentos com e sem as verbetações originais. Observa-se que o uso de verbetações melhora todas as métricas avaliadas, com maiores ganhos para o método BM25. Tais resultados eram esperados, pois a adição de termos discriminatórios nos documentos beneficia os métodos esparsos. Assim, os resultados indicam que o uso de palavras-chave tem um impacto positivo na tarefa de busca. O maior ganho foi observado na métrica R@10 (4,2 pontos percentuais) para TF-IDF e na MAP@10 (6,1 pontos percentuais) para BM25.

Tabela 14.2: Métricas de RI obtidas para cada modelo avaliado, avaliando a influência das verbetações originais. ‘*’ indicam diferenças estatisticamente significativas, de acordo com um teste-T pareado (p-valor < 0,05).

TF-IDF					
	MRR@10	NDCG@10	R@10	P@10	MAP@10
Sem verbetações	0,806	0,745	0,790	0,191	0,691
Verbetações originais	0,825*	0,780*	0,832*	0,201*	0,729*
BM25					
Sem verbetações	0,819	0,805	0,878	0,212	0,754
Verbetações originais	0,879*	0,858*	0,916*	0,222*	0,815*

14.5.2.2 Avaliação das verbetações geradas

Iniciando os experimentos com as verbetações geradas, a Tabela 14.3 apresenta a avaliação das verbetações geradas com decodificação gulosa. Observa-se que, ao adicionar as verbetações artificiais aos documentos, houve melhorias em quase todas as métricas para ambos os modelos avaliados. Para o método TF-IDF, a diferença foi estatisticamente significativa em todas as métricas avaliadas.

Tabela 14.3: Métricas obtidas para cada modelo de recuperação, ao utilizar documentos sem verbetação e utilizando as verbetações geradas com decodificação gulosa. ‘*’ indicam diferenças estatisticamente significativas, de acordo com um teste-T pareado (p-valor < 0,05).

TF-IDF					
	MRR@10	NDCG@10	R@10	P@10	MAP@10
Sem verbetações	0,806	0,745	0,790	0,191	0,691
Geradas	0,822*	0,770*	0,815*	0,196*	0,718*
BM25					
Sem verbetações	0,819	0,805	0,878	0,212	0,754
Geradas	0,854*	0,819*	0,877	0,211	0,768

Quanto ao BM25, houve ganhos em quase todas as métricas. No entanto, a diferença foi estatisticamente significativa apenas para duas métricas (MRR@10 e NDCG@10). Embora



tenham ocorrido melhorias para o TF-IDF, os resultados sugerem que o modelo BM25 foi mais sensível a falsos positivos e falsos negativos, introduzidos no *corpus* de busca por verbetações ruidosas geradas pelo PTT5. Assim, houve quase nenhuma alteração nas métricas baseadas em revocação e precisão.

As métricas analisadas mostram que ambos os modelos tiveram desempenho inferior em relação ao experimento anterior utilizando as verbetações originais (Tabela 14.2). Este resultado já era esperado, pois os resultados com verbetações originais atuam como um limite superior para este experimento. Mesmo com este cenário, o uso de verbetações artificiais melhorou ambos os métodos de RI em comparação à ausência de verbetação, evidenciando o potencial da geração automática de verbetações.

14.6 Conclusão

Nesta aplicação, utilizou-se modelos Transformers para a automação da escrita de verbetações presentes em ementas de processos judiciais brasileiros. O Transformer usado, PTT5, alcançou uma pontuação BLEU superior a 37% na geração textual. Em outras palavras, ao comparar as verbetações geradas com as originais, foi possível comprovar que as verbetações geradas foram capazes de reproduzir com fidelidade as características principais das verbetações escritas por especialistas.

Além disso, notou-se que as verbetações geradas capturaram com sucesso o estilo linguístico das verbetações originais. As frases curtas e objetivas, presentes nas verbetações, também são características das geradas. O modelo também foi capaz de reconhecer e expandir siglas comuns (como “CF” e “EC”).

Como a geração não é perfeita, também foi identificada a ocorrência de alucinações do modelo generativo usado e seus potenciais malefícios. Verificou-se também que as verbetações geradas tenderam a conter menos termos que as originais. Mesmo com estas ressalvas, os resultados qualitativos e quantitativos são evidências de que a geração automática de textos jurídicos é de fato possível.

Note que, em virtude das ressalvas (principalmente das alucinações), a automação completa da escrita de verbetações não é recomendada. Contudo, a metodologia apresentada aqui pode ser útil para auxiliar na escrita de ementas. Como exemplo, os Transformers podem ser incorporados a sistemas existentes, gerando automaticamente sugestões de verbetações (ou *templates*), que podem, então, serem corrigidas por especialistas, conforme a necessidade.

Considerando que o setor judiciário brasileiro convive com um crescente volume de processos para análise e que todo processo demanda tempo e recursos financeiros, qualquer esforço rumo à automatização é sem dúvida muito bem-vindo. Sendo assim, acredita-se que os resultados aqui apresentados possam estimular outros trabalhos no domínio jurídico, mostrando o potencial de modelos generativos em domínios com grandes demandas sobre recursos textuais.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001. Os autores agradecem também ao CEMEAI, pelo acesso ao *cluster* Euler, para a realização de experimentos.



Capítulo 15

PLN e Humanidades Digitais

Renata Vieira
Helena Freire Cameron
Fernanda Olival
Fatima Farrica
Maria José Bocorny Finatto
Ana Paula Banza
Ana Sofia Ribeiro
Cassia Trojahn

Publicado em: 13/03/2024
Atualizado em: 20/11/2024

 <https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/>

15.1 Introdução

A área de humanidades digitais (HD) tem ganhado força e adeptos nas últimas décadas, em paralelo com o desenvolvimento de ferramentas digitais que ampliam as possibilidades de armazenamento, acesso e processamento de dados. Essas capacidades estendem os horizontes de atuação de pesquisadores, permitindo a captura, organização e análise de um volume muito grande de dados. Com esta interação, que pode ser classificada de tecnológica, as humanidades ganham visibilidade, atravessam fronteiras disciplinares e enfrentam desafios sem precedentes.

Este capítulo apresenta as bases comuns para projetos na área de HD, relacionados ao Processamento de Linguagem Natural (PLN), em particular sobre fontes textuais com grafia pré-contemporâneas, o que imprime uma maior complexidade no processamento. Além disso, nosso enfoque é a relação entre as HD e o PLN no contexto da língua portuguesa, entendendo que o alcance da área de HD vai muito além da língua, e do escopo das questões e projetos aqui mencionados.

Na área de HD, relativamente aos trabalhos baseados em fontes textuais, encontramos uma grande variação, tanto nos períodos históricos das fontes, no seu suporte (manuscritos em papel, impressos, fotografados etc.), como no seu estágio de digitalização, que pode variar entre imagens digitais, textos em PDF e textos digitalizados em outros formatos. Todas essas variações adicionam esforços extras de processamento. Desta forma, apresentaremos um panorama geral e discutiremos os requisitos de preparação e organização das fontes, objetos de análise que depois de transcritas e digitalizadas, podem ser submetidas a processamentos mais avançados. O objetivo é mostrar não apenas como o PLN é útil e relevante nesse domínio, mas também como a área de HD é rica em despertar novas questões para o desenvolvimento do PLN.

Este capítulo está organizado da seguinte maneira. Na Seção 15.2 apresentamos os tópicos e os desafios relacionados à preparação das fontes, desde a digitalização até a



anotação. Na Seção 15.3, discutimos os processos de transformação de fontes em dados, nos quais várias tarefas conhecidas de PLN (discutidas em outros capítulos deste livro) podem ser empregadas. Esses processos de transformação de textos em dados organizados possibilitam ao pesquisador realizar análises diferenciadas sobre textos, por exemplo, ao agrupar tipos específicos de informação, ou rapidamente quantificar fenômenos observáveis. Na Seção 15.4, apresentamos alguns projetos relacionados às HD, em especial aqueles desenvolvidos pelas autoras. São projetos que apresentam diferentes fontes e objetivos de pesquisa, mas nos quais o PLN se faz presente em vários níveis, desde a preparação e a anotação até a extração, organização e partilha da informação. Por fim, apresentamos nossas considerações finais na Seção 15.5.

15.2 Preparação das fontes

Trabalhos em HD que se baseiam em textos podem se beneficiar das atuais técnicas de Inteligência Artificial (IA) e PLN para um melhor acesso às fontes. Os textos em HD podem ter relevância por suas características históricas e literárias, mas também podem estar relacionados a estudos sociais ou de outra natureza. Na verdade, é difícil distinguir os limites das HD, mas se pode dizer que a área está ligada ao emprego e à compreensão de novas maneiras de se desenvolver pesquisa, com base em recursos digitais e computacionais. Em relação ao processamento da língua, essa influência tecnológica se observa em uma variedade de processos referentes ao tratamento das fontes e sua informação, tais como:

- **Digitalização de documentos:** preparação no nível mais básico, para o tratamento computacional, e assim possibilitar a sua leitura por meio de programação.
- **Adição de metadados:** inserção de níveis de informação extra-textuais, que podem ser importantes para a estruturação e o armazenamento, e também para o acréscimo de interpretações sobre o conteúdo de uma coleção de interesse. Os metadados podem trazer informações de contexto, como autores, datas, locais de produção, origem, ou ainda informações como volume, páginas, etc. Ou podem ser de análise linguística: morfológica, sintática ou semântica (por exemplo, identificando entidades ou eventos).
- **Normalização:** a normalização para uma língua padrão e contemporânea pode ser necessária em situações como fontes históricas, textos de redes sociais que usam muitas abreviações, símbolos associados à emoções, ou comentários como *hashtags*, etc.
- **Anotação de corpora:** muitas vezes o estudo de uma fonte ou *corpus* necessita a adição de informações extras sobre o registro, que são representadas como metadados (mencionado acima). Os processos podem ser automáticos ou manuais. No entanto, geralmente o desenvolvimento de um processo automático requer um processo de anotação manual anterior, quer para o seu desenvolvimento (treino de algoritmos de aprendizado), quer para a avaliação da correção da anotação realizada por máquina.

Os métodos usuais de tratamento textual, desenvolvidos em pesquisas de PLN e IA, podem requerer adaptação a diferentes necessidades de investigação, estilo textual, objetivos da pesquisa e também ao uso pretendido e seus usuários. Idealmente, são necessárias novas interfaces para que os métodos desenvolvidos sejam usados de forma facilitada e



fora do contexto das estruturas de programação. Apresentaremos, a seguir, esses passos iniciais para projetos de HD que lidam com fontes textuais e, posteriormente, discutiremos sobre as técnicas de PLN.

15.2.1 Digitalização

Muitos projetos na área de HD necessitam lidar com a digitalização de manuscritos originais. A área de paleografia, em particular, lida com a leitura e transcrição de versões manuscritas para um suporte atual. Mesmo registros em suportes muito antigos, como a escrita cuneiforme, podem ser transcodificados para o digital (Liu et al., 2016), e a partir daí, podem passar por processos computacionais.

A paleografia, deste modo, também se desenvolveu de habilidade tradicional do intelecto, e manual, para uma prática digital (Ackel, 2021). Algumas ferramentas de transcrição, como por exemplo o Transkribus¹ ajudam na digitalização de manuscritos (Kahle et al., 2017), através de um sistema de algoritmos HTR (*Handwritten Text Recognition*).

Alguns documentos impressos podem necessitar a aplicação de tecnologia OCR (*Optical Character Recognition*). A qualidade da saída de sistemas OCR varia muito dependendo do formato e da qualidade da entrada. É um problema básico, mas não totalmente resolvido (Strien et al., 2020). Por exemplo, nas primeiras levadas de digitalização de acervos arquivísticos e bibliotecários, muitas digitalizações foram feitas em formatos de imagem, sem a preocupação de possibilitar a pesquisa dos seus conteúdos de texto. Por meio de diferentes ferramentas de transcrição automática com OCR, como o eScriptorium², podemos fazer essa transcrição de imagem para texto de forma automática.

Muitas vezes, as saídas de OCR precisam passar por processamento extra de correção, por vezes manual, para tornar a fonte adequada para as próximas etapas. Esse processamento se faz necessário para que outras ferramentas, como as de anotação de texto, tenham maior eficácia na aprendizagem dos *tokens* anotados. É preciso preparar os textos, por exemplo, retirando números de páginas, textos de cabeçalhos ou rodapés, e evitar a translineação/hifenização de palavras. Se o texto a ser anotado for composto por um conjunto de textos, pode ser necessário que estejam devidamente separados. Outro problema que pode emergir são os textos organizados em mais de uma coluna nas páginas, pois será preciso ordená-los numa segmentação textual contínua. Essas etapas devem ser consideradas em projetos de HD envolvendo textos e PLN e é, portanto, necessário antever-se o tempo e recursos necessários para o seu desenvolvimento.

15.2.2 Metadados

Uma vez digitalizada a fonte ou *corpus* de estudo, é necessário pensar na organização dos seus metadados. O material digital deve ser bem descrito, ou seja, deve conter a informação sobre a qual acervo pertence, identificar unicamente cada arquivo e, quando pertinente, associar autoria, data e outros elementos pertinentes. Os metadados também podem descrever a estrutura do documento, quando necessário, identificar volumes, capítulos, páginas, por exemplo. É importante, ainda, separar os metadados dos dados originais. Para materiais transcritos, por exemplo, devem ser identificados os cabeçalhos, notas de rodapé, numeração de páginas ou comentários adicionados aos originais. Esta organização e distinção dos elementos extra-textuais são essenciais para garantir o bom

¹<https://www.transkribus.org>

²<https://www.resilience-ri.eu/blog/resilience-tool-escriptorium/>



processamento posteriormente. Os metadados são essenciais para conectar uma fonte a outras e as conectar aos dados abertos ligados (Nair; Jeeven, 2004).

Há vários padrões de metadados bastante difundidos. Por exemplo, a Text Encoding Initiative (TEI) é um consórcio que desenvolve e mantém coletivamente um padrão para a representação de textos em formato digital. O seu principal resultado é um conjunto de diretrizes que especifica métodos de codificação para textos legíveis por máquina, principalmente em ciências humanas, ciências sociais e linguística³. Outro exemplo é o grupo de trabalho sobre “Publicação Sustentável de Metadados”, ativo na rede europeia da *Digital Research Infrastructure for the Arts and Humanities* (DARIAH) desde 2016. Esse grupo tem como objetivo apresentar informações sobre a reprodução e o compartilhamento de metadados entre institutos e pesquisadores no campo das ciências humanas e sociais.

Vários trabalhos trataram da representação de metadados por meio de vocabulários semânticos, para descrever quer recursos digitais, como vídeos, imagens, páginas web, quer recursos físicos, como livros ou obras de arte. Esse vocabulários podem ser usados para fornecer interoperabilidade para o uso de metadados na nuvem de dados ligados. Exemplos são o Dublin Core⁴, VoID (Vocabulary of Interlinked Datasets⁵), schema.org, DCAT (Data Catalog Vocabulary⁶) e PROV-O (Provenance Ontology⁷). Em particular, o DCAT é um vocabulário projetado para facilitar a interoperabilidade entre catálogos de dados publicados na Web. Permite a descrição de conjuntos de dados e serviços, usando um modelo e um vocabulário padrão que facilitam o consumo e a agregação de metadados de vários catálogos. Os metadados agregados do DCAT podem servir como um arquivo de manifesto⁸ como parte do processo de preservação digital.

15.2.3 Normalização

Em HD, as fontes textuais, quer manuscritas, quer impressas, sendo de uma época anterior ao estágio atual da língua, apresentam variações ortográficas ou morfossintáticas, não só em relação ao padrão atual como também dentro da mesma época.

A variação gráfica no século XVIII, por exemplo, é muito expressiva, quer em textos manuscritos, quer em impressos, especialmente ao nível da representação das vogais e de ditongos nasais, por exemplo “am-ão”, em “fizerão”, ou “oens-ões”, em “embarçaçoens”, etc. Também as consoantes duplas como “ll-l”, em “delles”, “tt-t”, em “Cratto”, entre outras, têm uma elevada frequência nesses textos. Note-se, também, a existência do fenômeno das chamadas grafias cultas ou pseudo-etimológicas, que já não integram o cânone atual da língua portuguesa. Estas podem trazer uma maior complexidade ao processamento lexical, como “th-t”, em “athé”, “ch-c” em “Christo” e “y-i”, por exemplo na palavra “Rey”. A palavra “coro”, que frequentemente é registada como “choro” em textos do século XVIII, é um exemplo da complexidade da variação gráfica deste período.

Do ponto de vista do leitor, a abordagem destes documentos requer um forte conhecimento do contexto para a sua desambiguação e exige, frequentemente, conhecimentos linguísticos e históricos para que possa desvendar o conteúdo de forma adequada. A variação gráfica tem uma elevada incidência nos documentos desta época. Grande parte

³<https://tei-c.org>

⁴<https://www.iso.org/standard/71339.html>

⁵<http://vocab.deri.ie/>

⁶<https://www.w3.org/TR/vocab-dcat-3/>

⁷<https://www.w3.org/TR/prov-o/>

⁸*Manifest file*, arquivo que contém metadados para um grupo de arquivos que são parte de um conjunto ou de uma unidade coerente.



das palavras tem pelo menos uma variante gráfica e algumas palavras têm múltiplas num mesmo texto. A título ilustrativo, veja-se a palavra “circunstância”, com nove variantes gráficas recolhidas num mesmo *corpus* textual (Cameron et al., 2023).

Do ponto de vista do PLN, a variação, quer pela sua grande incidência, quer pela sua imprevisibilidade, vem trazer uma perturbação acrescida. Variantes de uma mesma palavra são tratadas como se fossem unidades lexicais diferentes, diminuindo a eficácia do processamento. Para que se possa processar o texto de forma mais célere e eficaz, frequentemente têm de ser desenvolvidas tarefas intermediárias de normalização, quase sempre de forma manual. Unificam-se as variantes gráficas em torno de um lema, normalmente a forma existente na ortografia padrão atual. Essa tarefa, morosa e minuciosa, muito ganharia com processos automáticos ou semi-automáticos de normalização sistemática de textos desses estágios da língua portuguesa. Essa tarefa não só os tornaria aptos para futuros processamentos, mas também permitiria que leitores não historiadores ou não linguistas pudessem ter acesso mais facilmente a textos valiosos culturalmente e de valor patrimonial, com interesse para várias áreas do saber. Compreender essas diferenças e conseguir traduzir ou associar escritos antigos aos padrões atuais é um passo importante para outros níveis de processamento (Cameron et al., 2020).

Outra forma de mitigar as variantes de escrita é criar modelos de linguagem que incluam *corpora* de outros períodos de tempo, com suas variantes de ocorrência natural, numa fase adicional de treinamento (ajuste por *fine tuning*), para adaptar o modelo às variantes (Arevalo; Fonteyn, 2021). Em *corpora* históricos transcritos, a presença de vários critérios de transcrição pode dificultar a operação. Além disso, há a dificuldade em reunir um volume substancial de textos de uma mesma época que permita aprendizagens eficientes e com elevada acuidade.

15.2.4 Anotação de textos

A anotação de textos é um processo que por vezes pode ser complexo, necessitando de equipes interdisciplinares, de modo a prever todas as vertentes. A anotação requer a identificação de categorias de interesse para o estudo realizado, isso implica em impor alguma objetividade sobre a língua, sendo que a interpretação é fundamentalmente um processo subjetivo.

A anotação textual ocorre em diferentes níveis, sintático, semântico lexical, e discursivo. Como para qualquer outra área, também nas HD, é útil identificar elementos de interesse para embasar diferentes estudos. Geralmente a anotação sintática permite um tipo de análise diferenciada da anotação semântica. Estudos linguísticos de filologia podem se beneficiar das anotações de nível morfológico (lexical estrutural) e sintático (sentencial), como poderemos conferir em exemplos de projetos apresentados nas Seções 15.4.1 e 15.4.4.

A anotação de entidades é frequentemente adotada, uma vez que se trata de uma tarefa com desenvolvimento bastante avançado, com utilidade para estudos de literatura, história, geografia, entre muitos outros campos do conhecimento. Conforme a complexidade do fenômeno e a respectiva disponibilidade de recursos, a anotação pode ser manual ou automática.

Identificar personagens, instituições ou os locais mencionados em documentos históricos ou obras literárias possibilita uma série de análises. Quase sempre permite uma resposta mais consistente a perguntas básicas, mas essenciais em qualquer estudo rigoroso, como são as associadas a questões como “quem?”, “o quê?”, “onde?”. Isso poderá ser verificado nas Seções 15.4.2, 15.4.3 e 15.4.5. No processo de anotação, é fundamental respeitar o



contexto nas decisões que o anotador deve tomar, pois há muitos casos que se prestam a interpretação ambígua.

15.3 Transformação de textos em dados

Naturalmente, com a evolução dos estudos em HD, são produzidos dados diferenciados, mais elaborados, mais numerosos, e úteis aos investigadores em humanidades. Amplia-se, assim, a capacidade de análise, pela possibilidade de trabalhar com um volume maior de informação e pela capacidade de organizar essa informação de forma mais ágil e rápida em estruturas bem definidas. O PLN atua como área responsável por possibilitar essa transformação de textos em dados. São comuns nesses processos a presença dos seguintes elementos:

- extração de informação para a criação de conjunto de dados;
- uso de técnicas de representação e organização de conhecimento, para organizar os dados;
- criação de bases de dados interligadas, de forma a estimular o re-uso e a colaboração.

Contudo, tais processos, quando aplicados em pesquisas na área das humanidades, requerem não apenas ferramentas atuais de PLN, mas também uma interação mais próxima com os pesquisadores das respectivas áreas. Isso é necessário para o desenvolvimento das adaptações de ferramentas a diferentes objetivos e para a construção de interfaces adequadas ao seu uso. É, de fato, uma elaboração interdisciplinar. É importante aliar os interesses dos pesquisadores em humanidades às possibilidades mais atuais e eficientes de obtenção de informação por meio da aplicação de tecnologias da linguagem, especialidade dos pesquisadores em PLN. Discutiremos a seguir os elementos, mencionados acima, concernentes à transformação de textos em dados.

15.3.1 Extração de informação

Entre as subáreas de PLN, a extração de informação é o processo mais fortemente relacionado ao objetivo de transformação de dados não estruturados (textuais) em dados estruturados (tabelas ou banco de dados). O capítulo 17 deste livro apresenta em detalhes essa área do PLN. Enquadram-se nela as tarefas de reconhecimento de entidades nomeadas e de identificação e classificação de eventos, aplicadas, por exemplo, nos projetos descritos nas Seções 15.4.2, 15.4.3 e 15.4.5 para o estudo de fontes históricas.

O reconhecimento de entidades nomeadas possibilita a identificação de personagens importantes de um dado período; possibilita também mapear áreas de atuação, fazer relações com questões de cartografia e sistemas de informação geográfica. As instituições de diferentes perfis também podem ser identificadas. Se usarmos técnicas mais avançadas, os relacionamentos entre essas entidades também poderão trazer informações relevantes para os pesquisadores. A identificação de eventos e sua respectiva classificação podem ajudar a abordar zonas e contextos significativos, como, por exemplo, no projeto *Monsoon* (Seção 15.4.3) que realiza uma análise dos eventos inerentes a conflitos na coleção “Livro das Monções”.

Em textos das áreas das humanidades que estejam já digitalizados e normalizados, podem ser aplicadas ferramentas de PLN já desenvolvidas para diferentes tarefas, como



por exemplo, reconhecimento de entidades nomeadas, extração de eventos, resolução de referências, e sistemas de respostas a perguntas.

Uma outra maneira de fazer uso de ferramentas prontas, em textos não normalizados, com características diferenciadas do padrão contemporâneo, seria adaptar os modelos de linguagem já treinados. Essas adaptações podem ser feitas para datas históricas ou para a diversidade de mídias. Uma adaptação pode melhorar a performance dos sistemas, uma vez que os grandes modelos são treinados em versões contemporâneas da língua. Não conhecemos alguma adaptação de modelos para o português histórico, mas em breve deverão estar disponíveis, pois esta é uma área em rápida evolução.

15.3.2 Representação de conhecimento e ontologias

Ontologias são artefatos conceituais interpretados com diferentes nuances em filosofia, lógica e sistemas de informação. A ideia central é representar conceitualizações de forma estruturada. Em HD, a nuance mais comumente relacionada é aquela empregada em sistemas de informação, ontologias como estruturas conceituais que podem ajudar na organização dos dados e na comunicação entre sistemas. O nível de detalhe de uma ontologia pode variar entre estruturas taxonômicas, definição de instâncias, relações e axiomas lógicos. Elas auxiliam a esclarecer (e combinar) diversas conceitualizações que compõem diferentes disciplinas.

Ontologias têm sido usadas em diversos domínios do conhecimento em diferentes tarefas:

- organizar e anotar grandes quantidades de dados (anotação semântica);
- integrar dados de diversas fontes (integração semântica);
- representar conhecimento de domínios complexos;
- dar ancoragem para o raciocínio automático;
- suportar buscas em grandes quantidades de dados (busca semântica).

Em HD, além das tarefas citadas acima, ontologias têm sido utilizadas para representação de metadados, pois, geralmente, é necessário estender ontologias de domínio existentes ou desenvolver novas para atender às especificidades de cada *corpus* ou fonte. Essas descrições formais permitem a integração de dados de múltiplas fontes, de maneira independente de software e de esquema. Um passo essencial para melhorar a qualidade dos dados é usar vocabulários padronizados e ontologias para representação de dados e metadados (Guizardi, 2020).

Uma ontologia bastante referenciada em HD, em especial na área do patrimônio cultural, é o CIDOC-CRM. É entendida como uma ferramenta teórica e prática para a integração de informação. O objetivo é ajudar os pesquisadores, os administradores e o público a explorar questões complexas, relacionadas ao passado, disponíveis em conjuntos de dados diversos e dispersos. O CIDOC-CRM faz isto fornecendo definições e uma estrutura formal para descrever os conceitos e relações implícitos e explícitos, utilizados em documentações do patrimônio cultural.

15.3.3 Dados ligados e dados FAIR

Os esforços de partilha de dados produzidos nas pesquisas ainda têm um longo caminho a percorrer em termos de padronização. É muito importante tornar os dados compatíveis com os princípios FAIR (*Findability*/Encontrabilidade, *Accessibility*/Acessibilidade,



Interoperability/Interoperabilidade e *Reuse*/Reutilização) (Wilkinson et al., 2016). Esses princípios correspondem a um conjunto de 15 recomendações que visam facilitar a reutilização de dados por humanos e máquinas. Eles são independentes de domínio e podem ser implementados principalmente através de:

- (F) atribuição de identificadores exclusivos e persistentes a conjuntos de dados, descrevendo-os com metadados ricos que permitem sua indexação e descoberta;
- (A) uso de protocolos abertos e de padrões para acesso a conjuntos de dados;
- (I) uso de linguagens formais e de vocabulários padronizados *FAIR* para representar (meta)dados;
- (R) uso de metadados ricos sobre licença de uso, proveniência e qualidade de dados.

Portanto, o primeiro passo para o cumprimento dos princípios *FAIR* é atribuir metadados aos conjuntos de dados e definir esquemas precisos de metadados. De fato, 12 dos 15 princípios *FAIR* (Wilkinson et al., 2016) se referem a metadados. Para dar um passo adiante no aprimoramento *FAIR* dos dados, os esquemas de metadados devem se basear em modelos semânticos (ou seja, ontologias) para uma representação de metadados mais rica (Guizzardi, 2020). Graças à sua capacidade de tornar os tipos de dados explícitos, em um formato que pode ser processado por máquinas, as ontologias são essenciais para tornar os dados *FAIR*, mesmo para dados já publicados na Web (Jacobsen et al., 2020).

A ideia dos dados ligados é que os esforços possam ser reaproveitados e as bases enriquecidas com dados já identificados e organizados por outras iniciativas. A *open linked data cloud* (LOD)⁹ fornece uma visão geral dos conjuntos de dados vinculados que estão disponíveis na Web. Representa um grafo de conhecimento, composto por diversos padrões abertos, como URI, URL, HTTP, HTML, RDF, a linguagem de consulta SPARQL, entre outros. Esse padrão são mantidos por curadores de dados amadores e profissionais na indústria e na academia¹⁰.

Por exemplo, a *Europeana* é uma biblioteca virtual desenvolvida por países da União Europeia (Borin; Donato, 2023; Coneglian; Santarem Segundo, 2017). O protótipo agrega milhões de itens digitais, todos eles em domínio público, e está ligado ao LOD Cloud¹¹. Em (Koho et al., 2021) é proposta uma infraestrutura semântica compartilhada, baseada na ideia de representar conceitos de guerra como uma sequência espaço-temporal de eventos, nos quais participam soldados, unidades militares e outros atores. O esquema de metadados usado é uma extensão do CIDOC-CRM, complementado por várias ontologias de domínio histórico militar, e é ligado ao LOD Cloud.

Em termos de avaliação, várias estruturas foram propostas para avaliar o grau de *FAIRness* de um determinado objeto digital (Sun et al., 2022). Em vários deles, a avaliação é realizada respondendo a um conjunto de perguntas, também chamadas de métricas ou indicadores, ou preenchendo uma lista de verificação¹², como o *FAIRshake* (Clarke et al., 2019) ou o *FAIR Data Maturity Model* (FAIR Data Maturity Model Working Group RDA, 2020).

Outros autores (Devaraju et al., 2020; Wilkinson et al., 2019) propuseram abordagens automatizadas. Recentemente, além de avaliar o grau de *FAIRness* dos dados, as propostas

⁹<https://lod-cloud.net>

¹⁰<https://www.w3.org/wiki/LinkedData>

¹¹<https://www.europeana.eu/en>

¹²<https://fairassist.org/>



abordaram a avaliação de vocabulários e ontologias (Cox, 2021; Garijo; Poveda-Villalón, 2020).

Nesta seção discutimos alguns elementos do caminho de construção de bases de pesquisa a partir de textos. Muitos projetos na área de HD são desenvolvidos com o objetivo de mapear informações textuais volumosas, dispersas, ou ainda não totalmente exploradas em novas descobertas. Uma vez mapeadas e organizadas, elas podem estar disponíveis para outros investigadores. As áreas de extração de informação, construção de ontologias e dados ligados caminham em paralelo, embora esforços interdisciplinares ainda sejam necessários para sua integração. Na próxima seção serão apresentados exemplos de projetos que caminham nesse sentido.

15.4 Exemplos de projetos de HD envolvendo o processamento da língua portuguesa

Apresentamos aqui exemplos de trabalhos envolvendo HD e PLN com foco em língua portuguesa. Esta não é uma apresentação completa; há diversos outros projetos e trabalhos importantes e reconhecidos nesse domínio. Em especial, damos atenção aqui aos projetos relacionados com as autoras deste capítulo, e que colaboram no contexto do Laboratório Chronos¹³, o Laboratório de Humanidades Digitais do Centro Interdisciplinar de História Culturas e Sociedades, CIDEHUS, da Universidade de Évora, Portugal. São, em geral, projetos atuais em desenvolvimento, com exceção do primeiro, que tem uma relevância histórica no desenvolvimento dessa área no Brasil.

15.4.1 O *corpus* Tycho Brahe

O *corpus* histórico português Tycho Brahe (Sousa, 2014) é um *corpus* eletrônico de textos escritos em português por autores nascidos entre 1380 e 1978. O objetivo principal do projeto foi possibilitar, de forma ampla, a recuperação de informações filológicas e linguísticas dos textos. Esse pode ser considerado um dos trabalhos pioneiros na área de *corpus* histórico, com adição de anotação linguística e que tenha envolvido recursos computacionais na sua criação, sendo um importante contributo para essa área (Britto et al., 2002; Finger, 2000).

No site do projeto¹⁴, 95 textos se encontram disponíveis para pesquisa. Os textos contam com anotação linguística em dois níveis: anotação morfológica (*part of speech*) (60 textos, total de 2.204.889 palavras); e anotação sintática (31 textos, total de 1.311.834 palavras). Estão disponíveis ainda versões sem anotação. A coleção engloba o português brasileiro e o europeu, contendo, por exemplo, a Gazeta de Lisboa e o Jornal da Bahia.

15.4.2 As Memórias Paroquiais

As Memórias Paroquiais constituem uma das principais fontes sobre o Portugal metropolitano dos meados do século XVIII. São, aliás, das mais citadas pelos historiadores. A coleção é composta pelas respostas dos párocos a um inquérito lançado em 1758, três anos depois do grande sismo de 1755. Incluía 60 perguntas sobre o território da freguesia, pelo que reúne uma plêiade diversificada de informações, desde elementos administrativos e

¹³<https://sites.google.com/view/hdlabcidehus>

¹⁴<https://www.tycho.iel.unicamp.br/corpus/>



demográficos aos recursos existentes: piscícolas, minerais, aquíferos com propriedades específicas, produções agro-pecuárias, etc. Interessam tanto ao historiador como ao botânico ou ao sismógrafo, para referir apenas alguns dos campos abrangidos pela observação dos párocos informantes.

Estas respostas, encadernadas no século XIX, fez com que os volumes passassem a ser conhecidos como “Dicionário Geográfico de Portugal”. Aliás, a iniciativa de 1758 também tinha em vista dar continuidade a um projeto de dicionário com esse perfil, do Padre oratoriano Luís Cardoso (1697-1769), e que o terremoto de 1755 interrompera. A atual designação só se impôs no limiar do século XX (Olival et al., 2022).

Os originais manuscritos, num total de 43 volumes e um de índices, encontram-se na Torre do Tombo, desde o século XIX. Entre 1993 e 2003, os textos foram microfilmados e em 2005 esses rolos de 35 mm¹⁵, em preto e branco, foram digitalizados e disponibilizados online.

Por volta de 2007-2008, o CIDEHUS iniciou a tarefa de transcrever as Memórias das localidades a Sul do Tejo. O trabalho foi concluído em 2023 em um projeto colaborativo. Por vezes, até aproveitou transcrições já publicadas, sempre que foi possível entrar em contato com o Autor(a) da edição. Parte desse trabalho encontra-se disponível online, em acesso aberto, e com licenças *Creative Commons*, que facilitam a reutilização.

Nos últimos dois anos começaram a ser aplicadas técnicas de PLN ao conjunto transcrito. Principiou-se por anotar categorias simples e básicas: pessoas, locais e organizações (Vieira et al., 2021); já em 2023, usaram-se categorias de anotação mais complexas e com várias subdivisões. Teve-se em vista responder de modo mais adequado às necessidades do historiador e de outros cientistas ou pessoas interessadas. Também foi efetuado um ensaio de reconhecimento de entidades nomeadas, usando sistemas previamente desenvolvidos e que foram calibrados para esse efeito (Santos et al., 2024a). Baseavam-se em técnicas de aprendizagem de máquina e modelos de linguagem.

Está ainda em curso a normalização da grafia do texto, quer em versão com léxico explicativo para o grande público, quer em versões para processamento, trabalhando-se na possibilidade de normalizar de forma automatizada (Cameron et al., 2023).

Nos próximos anos, será dada continuidade à tarefa de transformar estes textos em dados confiáveis e de ligá-los a outros repositórios de conhecimento.

15.4.3 O projeto MONSOON

Extração de eventos é uma tarefa bastante conhecida de PLN, com recursos desenvolvidos para o português (Sacramento; Souza, 2021). O desafio, neste caso, é adaptar essas ferramentas à língua portuguesa do século XVII, esforço que está sendo explorado no projeto *Monsoon*: o Estado da Índia Habsburgo em perspectiva digital (1580-1640).

Este projeto pretende estudar as dinâmicas internas desse espaço macro que constituía o Estado da Índia português, combinando o estudo da presença formal portuguesa com a presença de comunidades portuguesas que se instalaram na Ásia fora do território administrado pela Coroa portuguesa, buscando relacionar estas duas esferas com o Outro – seja o europeu ou o asiático ou africano. Por outro lado, o projeto tem o objetivo de criar uma metodologia de análise semi-automática que possa ser replicada em coleções documentais semelhantes de outras potências europeias (Ribeiro, 2022).

Este projeto tem como base documental a correspondência trocada entre o monarca português e as suas instituições com o vice-rei da Índia. Este conjunto ficou conhecido

¹⁵Agradece-se à Dra. Anabela Ribeiro, do Arquivo Nacional da Torre do Tombo, estas informações.



como Livros das Monções, dado que era na altura da monção no Oceano Índico que os navios portugueses podiam partir de Goa para Portugal ou podiam chegar ao Índico. Só no regime climático de monções existem ventos no Índico que permitem que navios sem motor aí circulem. Este conjunto documental, dividido entre o Arquivo Nacional Torre do Tombo, em Lisboa, Portugal, e o Arquivo Histórico do Estado de Goa, em Pangim, Índia, possui informações sobre os mais variados aspectos da presença portuguesa na Ásia e abarca uma geografia que vai desde Moçambique e toda a costa oriental africana até ao Japão.

Este *corpus* documental é composto por um grande número de volumes, sendo que nenhum deles possui um índice sobre os assuntos tratados. Por meio da extração automática de eventos, os historiadores podem facilmente ter acesso a passagens relevantes a partir de uma palavra relacionada com uma determinada ação. É possível identificar e classificar semanticamente eventos, o que nos permite ter uma percepção clara dos temas abordados nesta correspondência, de forma eficiente e rápida. A partir da aplicação de extração de eventos numa amostra destas cartas, compreendemos que, em determinada cronologia, havia no Oceano Índico uma conjuntura generalizada de conflito, quer devido à concorrência de outras potências ultramarinas europeias, como a Holanda e a Inglaterra, quer entre os portugueses e outras entidades políticas asiáticas (Albuquerque et al., 2024a).

Através do sistema de extração de eventos, temos acesso a estatísticas de classificação que nos permitem uma percepção mais evidente dos assuntos representados em cada carta e assim analisar conjunturas históricas. Isso, no que diz respeito à metodologia da análise histórica, é algo relativamente novo.

Por outro lado, este projeto utiliza também a anotação de entidades nomeadas de quatro tipos: pessoas, locais, grupos de pessoas e instituições. Nos próximos passos deste projeto pretendemos conjugar as entidades extraídas por meio de anotação de texto e ligá-las aos eventos extraídos no texto, podendo assim articular agentes, instituições e locais às ações em que estão envolvidos.

15.4.4 A História do Futuro

Nos projetos em curso sobre a “História do Futuro” (Banza, 2022), de Padre António Vieira, após alguns testes com a ferramenta de transcrição de manuscritos Transkribus (Kahle et al., 2017), concluiu-se que, tratando-se de um texto relativamente curto e dadas as características do manuscrito, de leitura muito difícil, em parte devido às más condições do suporte, o treino necessário para uma utilização rentável da ferramenta não se justificaria neste caso, ao contrário de projetos envolvendo um grande volume de textos, uma vez que seria sempre necessária uma correção manual. Assim, o texto foi lido e transcrito manualmente para suporte digital, aplicando-se apenas posteriormente determinadas ferramentas de análise, de acordo com os objetivos pretendidos.

A primeira transcrição, feita a partir da leitura paleográfica do manuscrito, sendo maximamente conservadora, mantém um número muito significativo de traços do português seiscentista, e do português do Padre António Vieira em particular, que dificultam muito ou impossibilitam mesmo a aplicação da maioria das ferramentas de PLN. Assim, optou-se por, a partir da primeira, realizar uma segunda transcrição, igualmente em formato , mas normalizadora. Dadas as características do texto, a normalização foi, mais uma vez, feita manualmente, depois de definidos critérios rigorosos que, sem desvirtuarem as principais características do texto e da sua época, o tornassem apto a processamento computacional e a aplicação de ferramentas automáticas.



Sobre a versão normalizadora, tem sido possível aplicar com êxito algumas ferramentas e testar outras. Inicialmente, constituiu-se um *corpus* não lematizado, que foi anotado lexicalmente (permitindo, por exemplo a elaboração de um léxico da obra) e morfologicamente (o que permitirá rastrear automaticamente alguns aspetos inovadores na obra de Vieira neste domínio, como é o caso do uso dos pronomes clíticos), com recurso à ferramenta LX-Tagger¹⁶, disponível no repositório PORTULAN-CLARIN¹⁷.

Nesses domínios, a anotação automática, verificada depois manualmente, revelou-se particularmente eficaz. Elaboraram-se, ainda, com auxílio da ferramenta AntConc¹⁸, listas lexicais, alfabéticas, por frequência descendente, ou ordenadas pelo final de palavra, permitindo ver a disponibilidade do sistema sufixal. O *corpus* contém ainda um valioso intertexto latino, que foi processado manualmente de modo a constituir um sub-*corpus* autónomo.

Ainda no domínio da análise linguística, prevê-se a utilização do esquema de Dependências Universais, por exemplo, para a análise morfossintática, cujos testes têm revelado um desempenho bastante razoável, apesar de, como nos demais casos, não se poder excluir a revisão manual. Encontram-se também ainda em fase de avaliação ferramentas que permitam identificar similaridades semânticas (Lopez-Gazpio et al., 2017) entre o texto consensualmente aceito como “História do Futuro” e outros, fisicamente distintos, mas conceitualmente idênticos. Neste aspecto, as ferramentas presentemente disponíveis têm-se revelado pouco adequadas, quer ao tipo de texto, quer ao tipo de similaridades que se procura detectar. O objetivo será, neste caso, adaptar os modelos atuais à “História do Futuro”, conseguindo, essencialmente, mais eficiência, em termos de precisão e cobertura. Ainda assim, algumas ferramentas já disponíveis, como por exemplo, para o reconhecimento de entidades nomeadas ou para a análise de sentimentos, úteis para a análise histórico-cultural e literária da obra, poderão necessitar de uma menor adaptação.

15.4.5 Manuais médicos do século XVIII

Com a digitalização de acervos bibliográficos de bibliotecas físicas, impulsionada pelas HD (Finatto, 2023), tornou-se mais fácil acessar obras médicas antigas impressas em português. Todavia, tais obras precisam ter seus conteúdos sistematizados para que possam ser compreendidas e situadas em relação ao conhecimento médico e científico atual.

Mesmo com a digitalização e com as transcrições dos textos, conforme sua apresentação original, a interpretação e sistematização dessa informação requer consulta a toda uma série de materiais de apoio: compêndios de História da Ciência e da Medicina, edições filológicas, bases de dados com textos do português de diferentes épocas e vários tipos de dicionários. Assim, no âmbito do projeto “Humanidades digitais e conhecimento médico em língua portuguesa no século XVIII”, são analisados e disponibilizados¹⁹ textos médicos desse período, reunidos em um *corpus* de arquivos de transcrições (Lazzari; Finatto, 2023). Esse *corpus* é de acesso público e oferece manuais escritos por médicos, enfermeiros e cirurgiões. Nele, têm destaque as obras do médico Curvo Semedo (1635-1719) e o primeiro manual de Enfermagem publicado em português, de 1741.

O objetivo desse projeto é descrever e sistematizar terminologias, conceitos e modos de dizer relacionados a doenças e seus tratamentos, criando-se bases para um futuro hiper-

¹⁶<http://lxcenter.di.fc.ul.pt/tools/en/conteudo/LXTagger.html>

¹⁷<https://portulanclarin.net>

¹⁸<https://www.laurenceanthony.net/software/antconc/>

¹⁹<https://sites.google.com/view/projeto38597>



dicionário (Wives, 1997) de epidemiologia histórica luso-brasileira. Nesse hiperdicionário, a ideia é associar nomes de doenças, partes do corpo, perfis de pessoas/doentes, dados geográficos e datas, com nomes de remédios, de processos e de tratamentos que chegam aos seus correspondentes terapêuticos atuais. Esses dados podem ser correspondidos à identificação das Entidades Nomeadas ao longo dos textos (Zilio et al., 2022).

Naturalmente, há todo um trabalho de geração e tratamento desse *corpus* histórico, que é feito também com auxílio de recursos e técnicas de PLN. Isso porque, no período dessas obras, verificava-se toda a sorte de variação de escrita e de tipografia, de modo que em uma mesma página, de um mesmo autor, encontra-se, por exemplo a palavra “água” escrita como “agua” ou “agoa”, além de haver itens de vocabulário hoje desconhecido, como o termo anatômico “madre” por “útero”. Por isso, o projeto já tenta aproximar, com apoio computacional, a escrita antiga da sua forma atualizada, investindo-se na normalização automática dos textos (Zilio et al., 2024).

Embora sejam enfrentados problemas com a qualidade das informações automaticamente extraídas em dados brutos, pois lida-se com textos apenas na ortografia antiga, plenos de variações nas formas das palavras conforme escritas na época, ferramentas genéricas de PLN já se mostram capazes de identificar informações relevantes (Quaresma; Finatto, 2020). São ferramentas com bom potencial para ajudar especialistas que lidam com esses textos na criação de bases de conhecimento histórico. Enfim, o processamento desse *corpus* ajuda a identificar e refinar a informação nele contida. E isso pode ser muito relevante para várias aplicações, produzindo-se verdadeiros “mapas de navegação” da sua organização e conteúdos.

A ligação de ontologias é, por fim, outro dos objetivos desse projeto de pesquisa. Assim, as antigas terminologias de Saúde poderão ser encontradas e mapeadas em seus equivalentes em ontologias da atualidade (Schröml et al., 2012). Este mapeamento será mais uma forma de enriquecer os recursos derivados do *corpus* e estabelecer quais as áreas da Medicina, da Anatomia e que tipo de doenças e medicamentos eram conhecidos daquela época. Atualmente, as terminologias e expressões relacionadas, conforme usadas nos manuais de Curvo Semedo, já estão contrastadas com o vocabulário de Enfermagem empregado em 1741 (Finatto et al., 2023). Esse contraste, entre outros, pode ajudar a identificar diferentes perspectivas, gêneros textuais, conceitos e saberes dos vários práticos e profissionais de Saúde em ação e formação no século XVIII.

15.5 Considerações finais

Há um grande potencial de aplicações de PLN na área de humanidades digitais, muitas ainda a serem exploradas. Como visto neste capítulo, o PLN se faz presente no tratamento de problemas em diferentes fases, desde as etapas iniciais relacionadas à transcrição de manuscritos, ao tratamento de textos como imagem com técnicas como o OCR, a adição de metadados, e a normalização textual. Para as fases de processamento, a partir de um determinado texto preparado, as tecnologias de linguagem podem prestar auxílio a tarefas diversas como tradução, recuperação e extração de informações, criação de bases de conhecimento, e sua associação a ontologias ou outros dados. O ideal é que exista uma atenção especial aos padrões de disponibilização de dados, como os propostos pelos princípios *FAIR* de Dados Abertos.

Nessa área de intersecção entre humanidades e tecnologias é essencial um trabalho interdisciplinar. São várias questões relevantes a serem combinadas: os objetivos de um pesquisador da área de humanidades, como linguística e literatura, ciências sociais, his-



tória, e, na perspectiva do processamento de língua, os objetivos de encontrar, aplicar ou mesmo desenvolver soluções apropriadas para cada tipo de problema, maximizando a correção e a performance das soluções encontradas de acordo com a disponibilidade de recursos.

É crucial que uma perspectiva de IA, centrada no ser humano, seja levada em consideração para fornecer interfaces de usuário adequadas para preparar e acessar as fontes e os dados extraídos. Os projetos, muitas vezes considerando coleções distintas, ou mesmo possuindo objetivos diferentes, podem se beneficiar com a troca de experiências e com o uso das mesmas ferramentas para lidar com os textos e seus conhecimentos codificados.

No contexto do evento principal da língua portuguesa, o PROPOR, tem sido organizado o *Workshop* de Humanidades Digitais e Processamento de Língua Natural, que em 2024 estará na sua terceira edição. Nos anais desse *workshop* disponíveis *on-line*²⁰, podem ser encontrados mais detalhes sobre os projetos mencionados neste capítulo, e muitos outros.

Agradecimentos

Os trabalhos aqui mencionados contam com o suporte da Fundação de Ciência e Tecnologia (FCT) - projetos CEECIND/01997/2017, 2022.07730.PTDC, UIDB/ 00057/2020 (<https://doi.org/10.54499/UIDB/00057/2020>), e e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) - processos 308926/ 2019-6 (PQ) e 401770/2022-2 (PDE).

²⁰<https://sites.google.com/view/dhandnlp-propor>



Capítulo 16

PLN em Redes Sociais

Brenda S. Santana

Larissa A. de Freitas

Publicado em: 26/09/2023

Atualizado em: 13/03/2024

 <https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/>

16.1 Introdução

O Processamento de Linguagem Natural (PLN) desempenha um papel cada vez mais significativo no cenário das redes sociais. O volume de dados advindos de redes sociais a todo instante é imenso. Dentre os dados gerados, podemos citar os dados textuais, os quais variam desde conversas informais até discussões complexas. Nas redes sociais, as pessoas expressam ideias e opiniões de maneiras diversas. Isso inclui o uso de gírias, abreviações, emojis e outros elementos da linguagem cotidiana. Tratar esse tipo de dado não é uma tarefa trivial e é desafiador para os sistemas de PLN.

Os sistemas de PLN são ferramentas indispensáveis para compreender, analisar e extrair informações. O estudo dos estilos de linguagem utilizados nas redes sociais ajuda a melhorar a compreensão de textos informais. As redes sociais são fontes valiosas de informação e seus conteúdos podem ser utilizados como *corpora* para treinar e testar algoritmos de PLN, permitindo que pesquisadores e desenvolvedores trabalhem com exemplos reais e relevantes. Visto que o Brasil é um dos países com maior presença nas redes sociais, e o português é o idioma predominante nessas interações, tem-se aqui uma área muito fértil para o desenvolvimento de estudos de aplicações de abordagens e PLN.

Neste capítulo buscamos abordar algumas das principais áreas de aplicação de PLN em redes sociais, discutindo os desafios encontrados. Ainda, buscamos apresentar alguns dos recursos disponíveis para suporte no desenvolvimento de estudos voltados para as tarefas apresentadas, focando em dados em língua portuguesa. Para tanto, este capítulo se organiza da seguinte maneira: na Seção 16.2, apresentamos a definição de redes sociais e descrevemos sobre os conteúdos nela postados; na Seção 16.3, apresentamos as principais áreas de aplicação de PLN que utilizam essas redes sociais. E, na Seção 16.4, apresentamos as considerações finais.

16.2 Redes Sociais

Uma rede social é definida como um conjunto de dois elementos: atores e suas conexões (Wasserman; Faust, 1994). Nos últimos anos, as redes sociais (como: Facebook, Reddit, Youtube, Twitter/X, Whatsapp e Instagram) têm revolucionado a forma como indivíduos, grupos e comunidades interagem. Nelas, são compartilhados textos, fotos, vídeos e outros



tipos de conteúdo. Assim, as redes sociais estabelecem um ambiente rico e dinâmico que oferece inúmeras oportunidades para o estudo e o aprimoramento de abordagens em PLN. Segundo Recuero (2009), o estudo das redes sociais na Internet objetiva analisar como as estruturas sociais surgem, de que tipo elas são e como são compostas.

De acordo com Farzindar; Inkpen (2018), usar PLN em textos provindos de mídias tradicionais (como jornal, rádio e televisão) tem sido um tópico de pesquisa popular nos últimos 25 anos. Hoje, usar PLN em textos provindos de redes sociais é uma área de pesquisa que requer adaptações dos métodos tradicionais, já que os textos provindos de redes sociais têm várias peculiaridades, principalmente devido a sua natureza. Ainda, eles podem estar escritos em diferentes idiomas e pertencerem a diferentes fontes.

As redes sociais se popularizaram no Brasil em 2004, com a criação do Orkut¹. Desde lá, novas redes surgiram e com elas a percepção da necessidade e viabilidade de aplicação de abordagens em PLN para o estudo de conteúdos e comportamentos gerados nesse meio. Dentre as áreas de aplicação dessas abordagens, destacam-se a detecção de discurso de ódio e linguagem ofensiva, a detecção de ironia/sarcasmo/humor, a detecção de notícias falsas, a análise de sentimento, entre outras (Ferreira et al., 2017).

Na literatura, existe uma predominância do Twitter/X como fonte de dados, isso se deve, provavelmente, ao fato de ele oferecer uma API² que, de forma muito simples, consegue acessar mensagens publicadas e os dados associados a seus usuários (por exemplo, o número de seguidores deste). No caso do Facebook, é necessário criar um aplicativo e obter a autorização dos usuários para que seus dados possam ser acessados/capturados (Coello; Junqueira, 2019), o que pode ser visto como um limitante na extração de informações desta rede.

No ano de 2023, algumas mudanças ocorreram nas APIs do Twitter/X e do Reddit. No Twitter/X, os pesquisadores terão que se adaptar às restrições da versão gratuita ou assinar alguns dos planos pagos para manter suas atividades. Já, no Reddit, o uso da API³ passou a ser cobrado. Portanto, é de se esperar que mudanças aconteçam nas pesquisas que utilizam *corpora* advindos dessas redes sociais.

Como mencionado anteriormente, o conteúdo postado nas redes sociais pode variar muito de acordo com a plataforma, o público-alvo e a intenção por trás da postagem. Abaixo descrevemos brevemente as redes sociais mais utilizadas em trabalhos de PLN sobre redes sociais em língua portuguesa.

16.2.1 Facebook

O Facebook⁴ é atualmente a maior rede social do mundo, com 2.9 bilhões de usuários ativos em 2023. Ela permite que os usuários criem perfis pessoais, adicionem amigos, compartilhem textos, fotos, vídeos e atualizações de status. Os usuários podem interagir com as postagens de outros usuários através de curtidas, comentários e compartilhamentos. Além disso, essa rede social também permite a criação de páginas para empresas, tornando-se uma ferramenta importante para marketing e divulgação.

¹<http://orkut.com/>

²<https://developer.twitter.com/en/docs/twitter-api>

³<https://www.reddit.com/dev/api/>

⁴<https://www.facebook.com/>



16.2.2 Reddit

O Reddit⁵ é uma plataforma online de compartilhamento de conteúdo e discussões, organizada em comunidades chamadas *subreddits*. Os usuários podem enviar postagens, comentar, votar em conteúdos e interagir uns com os outros. Essa plataforma abrange uma ampla variedade de tópicos e interesses, permitindo que os usuários encontrem comunidades específicas que correspondam aos seus interesses. É um espaço onde os usuários podem trocar informações, debater, compartilhar histórias, memes e muito mais.

16.2.3 Youtube

O Youtube⁶ é uma plataforma de compartilhamento de vídeos que permite que os usuários compartilhem e assistam vídeos de uma variedade de gêneros, incluindo filmes, programas de TV, vídeos musicais, documentários, entre outros. Também é uma ferramenta de marketing importante para muitas empresas e indivíduos, os quais usam a plataforma para compartilhar conteúdo promocional e aumentar a conscientização sobre seus produtos ou serviços.

16.2.4 Twitter/X

O Twitter/X⁷ é um serviço de microblogging que pode ser utilizado para transmitir pequenas atualizações de status (Russell, 2011). Nele podem ser analisados os vínculos entre amigos e seguidores, grafos sociais e descobertas de mais informações sobre os usuários, inspecionando as entidades presentes em seus tweets. Os tweets são mensagens curtas (contendo até 280 caracteres, incluindo texto, imagens, GIFs, vídeos e links para outros sites) e públicas postadas no Twitter/X. Eles têm um alcance imediato e podem se tornar viral rapidamente, dependendo do conteúdo e da quantidade de interação que recebem de outros usuários. Isso faz do Twitter/X uma plataforma poderosa para disseminar informações, ideias e tendências em tempo real.

16.2.5 Whatsapp

O Whatsapp⁸ permite que os usuários troquem mensagens privadas. Apesar de ser usado principalmente para conversas individuais, o WhatsApp possui recursos de grupos de conversação, onde podem participar até 256 usuários, e encaminhamento de mensagens (Cabral et al., 2021). Concebido como um aplicativo de mensagens instantâneas, o WhatsApp evoluiu para uma plataforma multifacetada, permitindo não apenas conversas privadas, mas também a formação de grupos e comunidades, compartilhamento de mídia, chamadas de voz e vídeo e até mesmo recursos empresariais.

16.2.6 Instagram

Instagram⁹ é uma rede social para compartilhamento de fotos e vídeos. Nela também é possível acompanhar (seguir) outras contas, curtir, comentar e compartilhar publicações.

⁵<https://www.reddit.com/>

⁶<https://www.youtube.com/>

⁷<https://twitter.com/>. O Twitter mudou de nome em 2023 e agora se chama X. Por esse motivo, ao longo do texto, iremos usar o termo Twitter/X quando mencionarmos essa rede social.

⁸<https://www.whatsapp.com/>

⁹<https://www.instagram.com/>



Todas as publicações realizadas no aplicativo são mostradas por meio do *feed* e o usuário pode visualizar as postagens das contas que ele segue. Ainda, esta rede social oferece diversas outras funcionalidades, como: *boomerang*, *live* e *stories*.

16.3 Áreas de Aplicação

Abaixo são descritas quatro áreas de aplicações que surgiram com a finalidade de compreender, analisar e extrair informações de textos que são publicados diariamente nas redes sociais, são elas: detecção de discurso de ódio e linguagem ofensiva, análise de sentimento, detecção de notícias falsas e detecção de ironia/sarcasmo/humor.

16.3.1 Detecção de Discurso de Ódio e Linguagem Ofensiva

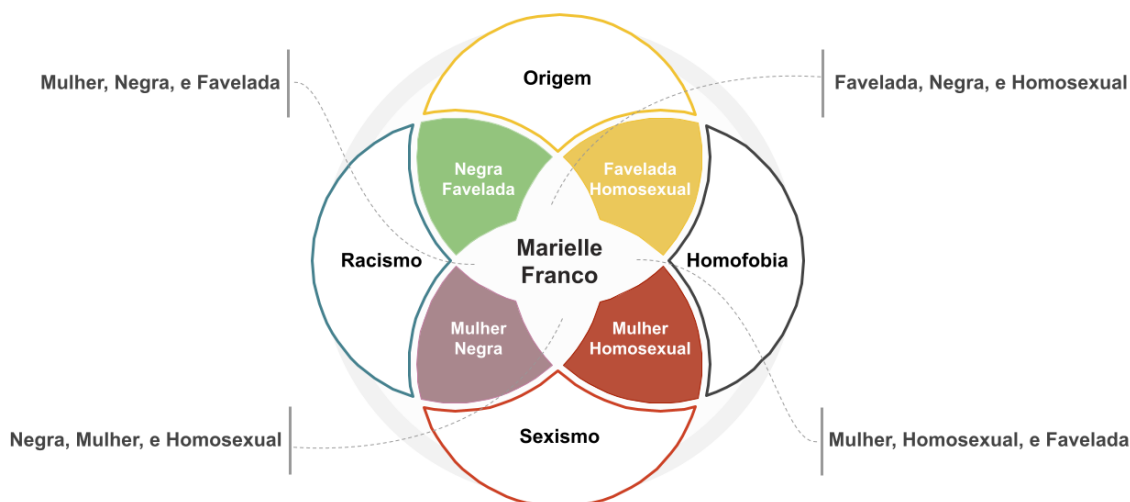
A partir de definições encontradas na literatura, termos diferentes, porém semelhantes, podem ser enquadrados como discursos simbolicamente prejudiciais (por exemplo, discurso perigoso, discurso tóxico, discurso de ódio, discurso intolerante e outros). Certos discursos possuem o potencial de causar danos significativos, inclusive críticos, e podem ser considerados tóxicos (Tirrell, 2018). Discursos tóxicos podem assumir diversas formas, podendo ser um discurso persistente ou momentâneo, afetar indivíduos ou a sociedade como um todo, causando danos temporários ou permanentes. O impacto de toxinas discursivas é de natureza social, afetando comunidades e prejudicando indivíduos pertencentes aos grupos-alvos. Essas toxinas podem incluir palavras ofensivas, insultos, discriminação, discurso de ódio, difamação, ameaças ou qualquer forma de linguagem que busque macular, menosprezar ou ferir a dignidade e a integridade de indivíduos pertencentes ao grupo-alvo. De acordo com Kumar et al. (2023), comentários tóxicos são a principal forma de ódio e assédio online.

Os grupos-alvos de discursos tóxicos podem variar dependendo do contexto e da natureza do discurso. Grupos frequentemente alvos de discursos tóxicos incluem minorias étnicas e raciais, comunidade LGBTQIA, mulheres, religiões minoritárias, portadores de deficiência, refugiados e imigrantes, e grupos políticos ou ideológicos. Contudo, qualquer grupo ou indivíduo pode ser alvo de discursos tóxicos, e a disseminação desse tipo de linguagem é prejudicial para a sociedade como um todo. Indivíduos pertencentes a diferentes grupos discriminados podem ainda ser alvos de discursos interseccionais que os atacam por múltiplas frentes. A Figura 16.1, adaptada de Santana (2023), ilustra um caso em que ataques interseccionais dirigidos a uma entidade foram disseminados na Internet. A imagem busca ilustrar o que aconteceu após o assassinato em 2018 da socióloga e política brasileira Marielle Franco, quando uma rede de ódio e desinformação gerou diversos comentários online atacando sua imagem por diversas características que ela possuía, e até outros traços que foram indevidamente atribuídos a ela. Teixeira; Zamora (2019) destacam que Marielle - mulher negra, assumidamente bissexual, favelada, defensora política dos direitos humanos - foi, sem dúvida, atravessada por todo tipo de opressão desencadeada pelo sistema machista, racista e classista. Os ataques registrados neste caso foram motivados pelo ódio.

A análise e a detecção de diferentes tipos de discursos tóxicos são um tópico de crescente interesse tanto para a área de PLN quanto demais áreas de interesse social como um todo. De acordo com Guimarães et al. (2020), quando focamos em comentários tóxicos, especialmente em notícias, o Facebook é a rede social que mais se destaca. O Reddit também introduz uma inclinação em relação à linguagem tóxica e ofensiva (Mohan et al.,



Figura 16.1: Exemplo de interseccionalidade entre diferentes formas de discurso de ódio observado no caso de Marielle Franco.



Fonte: Adaptada de (Santana, 2023)

2017). Por esse motivo, o conteúdo do Reddit tem sido usado para estudar microagressões (Breitfeller et al., 2019; Mollas et al., 2022) e depressão (Pirina; Çöltekin, 2018). De acordo com estudos realizados por Kumar et al. (2023), perfis que postam comentários tóxicos representam 3,1% de todas as contas que postam comentários no Reddit. Entretanto, ainda de acordo com os autores, apesar de seu percentual relativamente pequeno, tais contas desempenham um papel ativo e de alto impacto na plataforma.

Apesar dos diversos avanços pelos quais a área de PLN vem passando, a detecção de discursos tóxicos ainda é um desafio latente. O desenvolvimento de algoritmos de PLN e Aprendizado de Máquina (AM) para detectar esses tipos de conteúdo depende da disponibilidade de *corpora* anotados para treinamento. Conforme identificado por Trajano et al. (2023) quase todos os sistemas de detecção de toxicidade usam modelos de aprendizado supervisionado que requerem uma grande quantidade de dados rotulados¹⁰. Entre estes *corpora*, podemos ressaltar recursos para a língua portuguesa como o ToLD-Br¹¹ desenvolvido por Leite et al. (2020).

O ToLD-Br (Leite et al., 2020) é um conjunto de dados capturado do Twitter/X entre julho e agosto de 2019 com a ferramenta GATE Cloud's Twitter Collector¹². Elaborado para estudos sobre classificação automática de comentários tóxicos, este conjunto de dados tem como objetivo equilibrar o viés de anotação. Para tanto, 42 anotadores foram selecionados, com base em suas informações demográficas. Este *corpus* apresenta um conjunto de 21000 tweets em português manualmente anotados por três diferentes anotadores em sete categorias: LGBTQ+fobia, obsceno, insulto, racismo, misoginia e/ou xenofobia. Exemplo 16.1 apresenta uma instância de tweet obsceno e Exemplo 16.2 uma instância de tweet insulto contidos no *corpus* ToLD-Br.

¹⁰Em <https://hatespeechdata.com> alguns *corpora* anotados sobre discurso de ódio, abuso online e linguagem ofensiva são catalogados e podem ser utilizados como base para o desenvolvimento de estudos que buscam detectá-los.

¹¹<https://github.com/JAugusto97/ToLD-Br>

¹²<https://cloud.gate.ac.uk>

Exemplo 16.1:

“Aonde tem um monte que fala mal, mas ninguém vai embora do morro.” acha que alguém mora aqui por que quer, c*****o!? Que idéia. [...]

Exemplo 16.2:

[...] VAI SE F***R IRMÃO VC NÃO É FELIZ PQ NAO QUER

O estudo de discursos tóxicos é de suma importância por várias razões, abordando questões sociais, éticas e técnicas. O volume de informações gerado a partir das redes sociais e plataformas online aumenta a exposição a discursos tóxicos, o que pode causar impactos negativos na saúde mental e emocional dos usuários. Compreender e identificar esses discursos é fundamental para criar um ambiente digital mais saudável e seguro para todos. Discursos tóxicos frequentemente incluem discursos de ódio e manifestação de linguagem imprópria, ou seja, atos que podem promover a violência, intolerância e discriminação contra grupos específicos. A análise desses discursos permite identificar padrões prejudiciais e trabalhar para mitigar seus efeitos negativos.

Adicionalmente, muitos discursos tóxicos envolvem a disseminação intencional de informações incorretas, desinformação e notícias falsas (Seção 16.3.3). Ao estudar esses discursos, podemos desenvolver técnicas para detecção precoce de conteúdo enganoso, ajudando a manter a qualidade da informação nas redes. Além disso, técnicas de PLN auxiliam na detecção de linguagem irônica, sarcástica e outros formatos frequentemente usados nas redes sociais para mascarar discursos de ódio (Seção 16.3.4). Essas abordagens avançadas permitem que plataformas de redes sociais aprimorem suas ferramentas de moderação, identificando automaticamente discursos de ódio e adotando medidas para removê-los ou sinalizá-los.

Muitas iniciativas têm sido empreendidas com o intuito de possibilitar a detecção automatizada de discursos de ódio nas diferentes plataformas. Conforme mencionado por Fortuna; Nunes (2018), esse crescente interesse não se restringe apenas à ampla cobertura midiática, mas também à crescente relevância política do tema. No entanto, os autores também destacam desafios latentes, como a falta de técnicas automáticas adequadas e a escassez de dados confiáveis sobre o discurso de ódio, que continuam motivando pesquisas nessa área. Analisando estatísticas brasileiras, Dadico (2020) explana que os dados indicam que o ódio sobrevitimizava pessoas de grupos identificados por critérios de raça, cor, etnia, sexo, orientação sexual, identidade de gênero, origem nacional e regional, sem-teto ou deficiência, entre outros atributos que os expõem a maior vulnerabilidade social. Apesar da normalização do ódio, esse discurso é parte de uma narrativa socio-histórica que traz em si os modos de pensar de uma cultura. É pela língua que nos mostramos como somos, e enquanto ela pode ser um instrumento de empoderamento, também pode gerar exclusão, opressão. O avanço de estudos de aplicação de abordagens de PLN para a detecção de tais conteúdos é essencial. Entretanto, avanços nesta área de estudos dependem fundamentalmente de conjuntos de dados anotados, ferramentas de análise de texto e modelos específicos disponibilizados para tal.

Para o português, Fortuna et al. (2019) criou um conjunto de dados para a classificação do discurso de ódio, o HLPDSD¹³. As instâncias deste conjunto foram coletadas através do uso da API do Twitter/X. Para isso, foram usadas palavras-chave e hashtags como #dyke

¹³<https://b2share.eudat.eu/records/9005efe2d6be4293b63c3c44d4cf193e>



ou #womensPlaceIsInTheKitchen coletadas entre janeiro e março de 2017 (majoritariamente). Este conjunto de dados contém conteúdo de 1156 usuários diferentes e abrange diferentes tipos de discriminação, com base sobre religião, gênero, orientação sexual, etnia, e migração. Nele foram feitas duas anotações: binária (“é discurso de ódio” ou “não é discurso de ódio”) e hierárquica (“racismo”, “sexismo”, ou “homofobia”). Exemplo 16.3 apresenta uma instância de tweet que é discurso de ódio e Exemplo 16.4 uma instância de tweet que não é discurso de ódio contidos no *corpus* HLPHSD.

Exemplo 16.3:

Os negros deveriam voltar para suas terras!!

Exemplo 16.4:

Carne e feijão preto são deliciosos!

Na anotação binária, cada tweet foi anotado por três diferentes anotadores. Por fim, uma votação majoritária para determinar classificação final foi realizada nos 3059 tweets. Os autores realizaram experimentos utilizando uma LSTM combinada com *embeddings* pré-treinados para realizar uma classificação base a partir deste conjunto de dados e assim demonstrar seu potencial de uso. O resultado obtido foi a medida-F de 78%.

Outro conjunto de dados disponível na literatura que foi elaborado para estudos sobre a classificação do discurso de ódio é o HateBR¹⁴, elaborado por Vargas et al. (2022b). O HateBR é composto por 7000 textos sobre o domínio político coletados através da API do Instagram. Neste conjunto de dados, constam postagens de seis contas pré-definidas (gênero - 4 mulheres e 2 homens, posição política - 3 liberais e 3 conservadores). Sua anotação foi feita de três maneiras: binária (“é ofensivo” ou “não é ofensivo”), granularidade (“levemente ofensivo”, “moderadamente ofensivo” e “altamente ofensivo”) e grupos de discursos de ódio (“partidarismo”, “sexismo”, “intolerância religiosa”, “apologia pela ditadura”, “gordofobia”, “homofobia”, “racismo”, “anti-semitismo” e “xenofobia”). Exemplo 16.5 apresenta uma instância de texto que é discurso de ódio e Exemplo 16.6 uma instância de texto que não é discurso de ódio contidos no *corpus* HateBR.

Exemplo 16.5:

Vagabunda. Comunista. Mentirosa. O povo chileno nao merece uma desgraça desta

Exemplo 16.6:

Pois é, deveria devolver o dinheiro aos cofres públicos do Brasil. Canalha.

Tal como diversas outras tarefas de aplicação de abordagens de PLN, apesar dos esforços recentes, a detecção de discurso de ódio em português fica muito atrás do inglês (Jahan; Oussalah, 2023). A detecção de discursos de ódio em língua portuguesa é, sem dúvida, uma área promissora de pesquisa no campo do PLN. Redes sociais são um terreno fértil para a disseminação de discursos de ódio em qualquer idioma. Dada a popularidade da língua portuguesa nas redes sociais, tem-se aqui uma área muito fértil para o desenvolvimento de estudos de aplicações de abordagens e PLN. A detecção de discursos de ódio

¹⁴<https://github.com/franciellelvargas/HateBR>



em português envolve desafios únicos, como a diversidade linguística, o uso de gírias e expressões regionais, além das particularidades culturais. Isso torna a pesquisa nessa área empolgante e relevante, não apenas do ponto de vista técnico, mas também do ponto de vista social e ético.

Embora muito do que é visto em discursos tóxicos seja também discurso de ódio, cabe ressaltar que outras formas de toxicidade também são manifestas através de discursos. Há também o que chamamos de linguagem ofensiva. Diferentemente de discursos de ódio, os quais são voltados para indivíduos ou grupos específicos de pessoas com base em características identitárias, a linguagem ofensiva tem a intenção de magoar, insultar ou provocar os sentimentos das pessoas, sem necessariamente ter um objetivo discriminatório. É importante notar que a linha entre discurso de ódio e linguagem ofensiva nem sempre é clara, e o contexto em que o conteúdo é apresentado pode influenciar a percepção do quão prejudicial ele é. Isto é, todo discurso de ódio é uma linguagem ofensiva, mas nem toda linguagem que é ofensiva é também um discurso de ódio. Ambos podem ser prejudiciais e problemáticos em diferentes aspectos, e muitas vezes é necessário avaliar cuidadosamente o conteúdo para entender suas implicações e tomar medidas apropriadas para mitigar seus efeitos negativos. Tal qual os demais discursos considerados tóxicos, é importante também o desenvolvimento de meios de detecção de linguagem ofensiva.

Conjuntos de dados voltados para a detecção deste tipo de linguagem podem ser usados em um contexto que não é necessariamente de ódio. Para o português brasileiro, Trajano et al. (2023) construíram um conjunto de dados voltados a detecção de linguagem ofensiva, nomeado OLID-Br¹⁵. Exemplo 16.7 apresenta uma instância de texto que é insulto e sexismo e Exemplo 16.8 uma instância de texto que é insulto e ideologia contidos no *corpus* OLID-Br.

Exemplo 16.7:

Pior do que adolescentezinhas de merda...são pessoas que levam filmes tão a sério! O livro/filme é dela, ela faz o que quiser! E por mais ruim que seja, ta rendendo milhões (:

Exemplo 16.8:

“Heterofobismo?” Pelo que eu saiba héteros nunca foram perseguidos, mortos e torturados por centenas de anos pela sua orientação sexual. Em alguns países do Oriente, gays são presos e , as vezes, mortos só por causa de orientação sexual. Isso NUNCA aconteceu e nem vai acontecer com héteros. A implantação nas escolas da ideia de que existem casais diferentes é ótima! O conhecimento é a arma contra a ignorância e tenho certeza que vai ajudar muitas crianças confusas por aí! “Heterofobismo?” Estude mais um pouco antes de abrir a boca e passar vergonha, por favor.

Inspirado em outros *corpora* similares (do inglês, *Offensive Language Identification Datasets* ou OLID), construídos para outros idiomas, o OLID-Br reúne dados de diferentes fontes: Twitter/X, YouTube, e ainda de outros conjuntos de dados em português anotados com um esquema de anotação distinto do proposto. Os conjuntos de dados utilizados foram o OffComBR de Pelle; Moreira (2017), NCCVG¹⁶ de Nascimento et al.

¹⁵<https://github.com/DougTrajano/olid-br>

¹⁶<https://github.com/LaCAfe/Dataset-Hatespeech>



(2019), HLPHSD de Fortuna et al. (2019), e ToLD-Br de Leite et al. (2020). O conjunto de dados OLID-BR contém anotações para cinco tarefas, são elas: (1) classificação de comentário tóxico: classificação binária utilizada para identificar se um comentário é ou não tóxico; (2) detecção do tipo de toxicidade: classificação multi-rótulo que identifica os rótulos de toxicidade presentes em um comentário tóxico; (3) classificação de alvo de toxicidade: classificação binária que prevê se um comentário tóxico é direcionado ou não; (4) identificação do tipo de alvo de toxicidade: classificação multiclasse que identifica o tipo de alvo de um comentário direcionado; e (5) categorização de *spans*: tarefa voltada a detecção de *spans* (parte de um texto) em um comentário tóxico. O conjunto de dados contém 6.354 (extensível para 13.538) comentários rotulados usando um esquema de anotação de três camadas com granulação fina compatível com conjuntos de dados em outros idiomas, o que permite o treinamento de modelos multilíngues.

16.3.2 Análise de Sentimento

Com a proliferação das redes sociais e das plataformas de avaliação online (tais como: TripAdvisor¹⁷, Booking¹⁸ e Airbnb¹⁹), assim como em diversos sites de e-commerce, uma infinidade de textos opinativos são publicados diariamente. Estes textos têm grande potencial para apoiar os processos de tomada de decisão (Zhang et al., 2023). A Análise de Sentimento (AS) estuda as opiniões, sentimentos, avaliações, apreciações, atitudes e emoções em relação a entidades como produtos, serviços, organizações, indivíduos, problemas, eventos, tópicos e seus diferentes aspectos expressos em textos (Liu, 2012). Nesta área desenvolvem-se aplicações em diversos campos do conhecimento como: política, finanças e marketing²⁰.

Existem muitos nomes e tarefas ligeiramente diferentes, por exemplo, análise de sentimento no nível de aspecto, reconhecimento/classificação de emoções etc. A AS visa encontrar soluções computacionais para extrair e analisar as opiniões das pessoas sobre uma entidade e seus diferentes aspectos (Pereira, 2021). Como as opiniões podem ser categorizadas com polaridades (por exemplo, positivo e negativo), a AS pode ser considerada uma tarefa de classificação de texto (Zhang et al., 2023). E como se trata de uma tarefa de classificação de texto, de acordo com Tan et al. (2023), três tipos de abordagens podem ser utilizadas, são elas: AM, Aprendizado Profundo (AP) e Aprendizado Conjunto (AC), usualmente referido como *ensemble learning*. Abordagens baseadas em AM, como classificador Ingênuo de Bayes (em inglês, *Naïve Bayes* ou NB) de Zhang (2004) e Máquina de Vetor de Suporte (em inglês, *Support Vector Machine* ou SVM) de Cortes; Vapnik (1995), usam modelos matemáticos para prever sentimentos. Já, as abordagens baseadas em AP, como Redes de Memória Longa de Curto Prazo (em inglês, *Long Short-Term Memory* ou LSTM) de Hochreiter; Schmidhuber (1997), utilizam Redes Neurais Artificiais para prever sentimentos. O AC combina vários classificadores para obter um melhor desempenho de AS.

No trabalho de Pereira (2021) é apresentada uma pesquisa de AS em língua portuguesa. Nele são apresentados os principais tipos de abordagens de AS, as quais podem ser baseadas em AM (classificação também proposta por Tan et al. (2023)), em léxico de sentimento, em conceitos, e híbrida. Abordagens baseadas em AM utilizam algoritmos de AM tradicionais.

¹⁷<https://www.tripadvisor.com.br/>

¹⁸<https://www.booking.com/>

¹⁹<https://www.airbnb.com.br/>

²⁰Veja exemplo na área do Direito no Capítulo 12.



Já, abordagens baseadas em léxico de sentimento obtêm o grau de polaridade de opinião ou emoção de um léxico de sentimento. As abordagens baseadas em conceito usam redes de conceito (por exemplo: ontologias) para realizar a análise semântica do texto. Por fim, as abordagens híbridas, combinam as abordagens mencionadas anteriormente.

Em geral, a AS tem sido investigada principalmente em três níveis de granularidade: documento, sentença ou aspecto (Liu, 2012). No nível de documento, um sentimento é atribuído ao documento como um todo (Exemplo 16.9 com polaridade positiva). No nível de sentença, um sentimento é atribuído a cada sentença do documento (Exemplo 16.10 com polaridade positiva). No nível de aspecto, um sentimento é atribuído a cada aspecto de determinada entidade. É uma análise mais refinada, onde os aspectos podem ser atributos ou componentes de uma entidade (Exemplo 16.11 com polaridade positiva para o aspecto **café da manhã**).

Exemplo 16.9:

O café da manhã é **incrível**. O hotel é um **ótimo** lugar para relaxar e curtir cada momento.

Exemplo 16.10:

O café da manhã é **incrível**.

Exemplo 16.11:

O **café da manhã** é **incrível**.

Abordagens de AS são altamente dependentes do uso de ferramentas de PLN, pois precisam interpretar textos em linguagem natural. Logo, desenvolver soluções específicas para a língua portuguesa está diretamente condicionado ao desenvolvimento de recursos linguísticos para a língua. Segundo Lo et al. (2017), o português é uma das línguas com poucos recursos linguísticos disponíveis, apesar de estar entre as línguas mais utilizadas na Web.

Dentre os recursos para o português brasileiro, podemos citar os léxicos de sentimentos: OpLexicon²¹ (Souza et al., 2011b), OpenWordNet-PT²² (De Paiva et al., 2012), SentiLex²³ (Silva et al., 2012), Reli-Lex²⁴ (Freitas, 2013), Brazilian Portuguese LIWC Dictionary²⁵ (Balage Filho et al., 2013), Word NetAffect-BR²⁶ (Pasqualotti, 2015), Personalitatem Lexicon (Machado et al., 2015), AffectPT-BR²⁷ (Carvalho et al., 2018) e LexReli (Machado et al., 2018).

- O OpLexicon (Souza et al., 2011b) possui 30.322 palavras (23.433 adjetivos e 6.889 verbos) e foi construído com base em um *corpus* do português brasileiro (composto por 346 resenhas de filmes e 970 textos jornalísticos), no *thesaurus* denominado TEP²⁸ (do português, Thesaurus Eletrônico Básico para o Português do Brasil) de Dias-da-Silva; Morales (2003) e no léxico de sentimento de Hu; Liu (2004) traduzido para o português.

²¹<https://www.inf.pucri.br/linatural/wordpress/recursos-e-ferramentas/oplexicon/>

²²<https://github.com/own-pt/openWordnet-PT>

²³<https://b2find.eudat.eu/dataset/b6bd16c2-a8ab-598f-be41-1e7aeecd60d3>

²⁴<https://www.linguatca.pt/Repositorio/ReLi/>

²⁵<http://143.107.183.175:21380/portlex/index.php/pt/projetos/liwc>

²⁶<https://www.inf.pucri.br/linatural/wordpress/recursos-e-ferramentas/wordnet affect br/>

²⁷<https://github.com/LaCAfe/AffectPT-br>

²⁸<http://www.nilc.icmc.usp.br/tep2/>



- A base de dados da OpenWordNet-PT (De Paiva et al., 2012) é o resultado da tradução da base de dados da WordNet de Princeton²⁹, portanto, contém uma base de dados com grande abrangência, possui 62034 sentidos de pares de palavras e 45421 palavras únicas.
- A versão 2 do SentiLex (Silva et al., 2012) é composta por 82347 formas flexionadas, organizadas em adjetivos (16863), substantivos (1280), verbos (29504) e expressões idiomáticas (34700).
- O ReLi-Lex (Freitas, 2013) é derivado do *corpus* ReLi de Freitas et al. (2012), que é composto por resenhas de livros publicadas na internet e possui 1600 resenhas de treze livros (sete autores), este léxico contém 609 entradas.
- O Brazilian Portuguese LIWC Dictionary (Balage Filho et al., 2013) é um léxico disponível para a língua portuguesa, construído a partir do LIWC de Pennebaker et al. (2001), ou seja, foi resultado de tradução automática, utilizando diversos dicionários bilíngues português-inglês e possui 127.149 instâncias.
- O WordNetAffect-BR (Pasqualotti, 2015) é um vocabulário de emoções que possui 289 palavras (adjetivos e substantivos).
- O Personalitatem Lexicon (Machado et al., 2015) contém lexemas de conotação afetiva baseada nos traços de personalidade e foi construído com base no *Linguistic Inquiry e Word Count* (LIWC) 2.015.
- O AffectPT-BR (Carvalho et al., 2018) tem um total de 1.139 palavras atribuídas na categoria “afeto”, 479 em “posemo” e 661 em “negemo”.
- O LexReli (Machado et al., 2018) é uma combinação de três léxicos, OpLexicon (Souza et al., 2011b), SentiLex (Silva et al., 2012) e Brazilian Portuguese LIWC Dictionary (Balage Filho et al., 2013), especializado em identificar a polaridade de aspectos em textos opinativos sobre livros e contém 1.543 entradas.

Além dos léxicos há também os *corpora* anotados para a tarefa de AS: ReLi (Freitas et al., 2012), comentários sobre hotéis publicados no TripAdvisor (Freitas, 2015), comentários sobre produtos publicados no Buscapé³⁰ (Avanço; Nunes, 2014), comentários sobre restaurantes (Farias et al., 2016), TweetSentBR³¹ (Brum; Nunes, 2018), UTLCorpus³² (Sousa et al., 2019) e tweets sobre a pandemia de COVID-19 (Vargas et al., 2020). Exemplo 16.12 apresenta uma instância de texto que é positivo e Exemplo 16.13 uma instância de texto que é negativo contidos no *corpus* ReLi. Exemplo 16.14 apresenta uma instância de texto que é positivo, Exemplo 16.15 uma instância de texto que é negativo e Exemplo 16.16 uma instância de texto que é neutro contidos no *corpus* TweetSentBR. Exemplo 16.17 é um comentário sobre restaurante onde o aspecto **sobremesa** é positivo e o **atendimento** é negativo.

Exemplo 16.12:

impossível abandonar o livro pela metade

Exemplo 16.13:

a tradução deveria ser CRAPúsluco

²⁹<https://wordnet.princeton.edu/>

³⁰<https://github.com/avanco/LBC>

³¹<https://metatext.io/datasets/tweetsentbr>

³²<https://github.com/RogerFig/UTLCorpus>



Exemplo 16.14:

Essa mulher que faz voz de siri e do google tradutor é mó linda #TheNoite

Exemplo 16.15:

Nunca fiquei tão bravo numa eliminação quanto hoje, Mas fazer o que, né?
#HellsKitchenBR

Exemplo 16.16:

Vamos acorda esse prédio!!!!!! #AltasHoras @luansantana

Exemplo 16.17:

As **sobremesas** deste restaurante são **deliciosas**, mas o **atendimento** é muito **ruim**.

No ano de 2022, foi proposto um desafio sobre AS no nível de aspecto (em inglês, *Aspect-based Sentiment Analysis* ou ABSA) para língua portuguesa no IberLEF³³ denominado ABSAPT³⁴. A proposta do ABSAPT foi inspirada em competições propostas em outros idiomas, como SemEval (Pontiki et al., 2014, 2015, 2016) para o inglês e EVALITA (Mattei et al., 2020) para o italiano. Além disso, tinha como público-alvo acadêmicos, pesquisadores e profissionais de empresas privadas. Na competição participaram cinco equipes de diferentes universidades e institutos do Brasil. O *corpus* disponibilizado na competição foi desenvolvido por Freitas (2015) e Corrêa (2021). Os participantes usaram diferentes tipos de abordagens para resolver a tarefa de ABSA, a qual foi dividida em duas, identificação de aspectos e extração de polaridade destes aspectos. O time da UFS-CAR (Assi et al., 2022) propôs uma solução baseada em regras e léxico de sentimento, os times do NILC (Machado; Pardo, 2022) e da UFPR (Heinrich; Marchi, 2022) propuseram soluções baseadas em AM, utilizando algoritmos de AM tradicionais como *Conditional Random Field* (CRF) e os times Deep Learning Brasil (Gomes et al., 2022), PiLN (Ricarte Neto et al., 2022) e UFPR (Heinrich; Marchi, 2022) propuseram soluções baseada em AP, utilizando Transformers (Silva et al., 2022a). Enfim, estratégias como estas, especialmente para línguas com poucos recursos como o português, são extremamente importantes.

16.3.3 Detecção de Notícias Falsas

Uma notícia falsa é uma mensagem transmitida conscientemente por um remetente para promover uma falsa crença ou conclusão por parte do destinatário (Fuller et al., 2006). Segundo Oliveira et al. (2020), a classificação de notícias falsas pode ser vista como uma execução de uma classificação binária entre falso ou verdadeiro. A principal diferença entre a definição dos problemas de classificação de notícias falsas é em função dos diferentes esquemas de anotação ou contextos de aplicação em diferentes conjuntos de dados. Em geral, os dados são coletados de declarações anotadas em sites de verificação de fatos, com

³³<https://sites.google.com/view/iberlef2022/>

³⁴<https://sites.google.com/inf.ufpel.edu.br/absapt2022/>



os rótulos “verdadeiro” ou “falso”. No Brasil, algumas agências de checagem são: Agência Lupa³⁵, Aos Fatos³⁶, Fato ou Fake³⁷ e Comprova³⁸.

A detecção de notícias falsas, também conhecidas como *fake news*, em redes sociais é uma área de pesquisa crítica e desafiadora. A aplicação eficaz de técnicas de PLN nesse contexto é crucial para preservar a integridade da informação online e combater a desinformação que pode trazer consequências sérias tanto na política, quanto na economia, e ainda na sociedade como um todo. O PLN desempenha um papel fundamental no desenvolvimento de abordagens eficazes para lidar com esse problema (veja Capítulo 8). Apesar do problema de disseminação de notícias falsas estar presente em todas as redes sociais, algumas tendem a ter o compartilhamento deste tipo de conteúdo mais dissipado. De acordo com Cabral et al. (2021), o Whatsapp facilita a disseminação rápida de desinformação. No Brasil, cerca de 35% das notícias falsas são compartilhadas através do WhatsApp (Newman et al., 2020), e 40,7% destas mensagens são compartilhadas após serem desmentidas (Resende et al., 2019).

Na literatura, encontramos alguns *corpora*, descritos na língua portuguesa, anotados para a tarefa de detecção de notícias falsas em língua portuguesa, são eles: COVID-19³⁹, FakeTweetBr⁴⁰ de Cordeiro; Pinheiro (2019), Fake.br-Corpus⁴¹ de (Monteiro et al., 2018) e FakeWhatsApp⁴² de Cunha (2021). O COVID-19 contém notícias sobre a cura da COVID-19 postadas no Twitter/X. O FakeTweetBr é um *corpus* de notícias falsas também advindo do Twitter/X. O Fake.br-Corpus contém notícias classificadas em seis grandes categorias (política, TV e celebridades, sociedade e notícias diárias, ciência e tecnologia, economia, religião) extraídas do G1⁴³, Folha de São Paulo⁴⁴ e Estadão⁴⁵. O FakeWhatsApp possui mensagens anônimas do WhatsApp de grupos públicos do português brasileiro para detecção automática de desinformação textual e de usuários maliciosos. Exemplo 16.18 apresenta uma mensagem falsa e Exemplo 16.19 apresenta uma mensagem verdadeira contidas no *corpus* FakeWhatsApp.

Exemplo 16.18:

Olhem reportagem da TV italiana RAÍ Essa matéria é de 2015 e apresenta a preocupação da comunidade científica a respeito das pesquisas feitas pelo Instituto de virologia de Wuhan com o coronavírus de morcegos, combinando - o geneticamente com o SARS para fazê-lo mais contagioso, a fim de estudar seus efeitos.

Exemplo 16.19:

Em novembro de 2015 a televisão italiana apresentou uma matéria que denunciava que cientistas chineses haviam desenvolvido um super virus capaz de desencadear pneumonia aguda em seres humanos.

³⁵<https://piaui.folha.uol.com.br/lupa/>

³⁶<https://www.aosfatos.org/>

³⁷<https://g1.globo.com/fato-ou-fake/>

³⁸<https://projeto comprova.com.br/>

³⁹<https://wp.ufpel.edu.br/midiars/datasets/>

⁴⁰<https://github.com/prc992/FakeTweet.Br>

⁴¹<https://github.com/roneysco/Fake.br-Corpus>

⁴²<https://github.com/cabrau/FakeWhatsApp.Br>

⁴³<https://g1.globo.com/>

⁴⁴<http://m.folha.uol.com.br/>

⁴⁵<https://www.estadao.com.br/>



Um dos trabalhos recentes que utiliza Aprendizado Profundo (AP) na detecção de notícias falsas é o trabalho de Narde (2021). Nele foram utilizados diferentes modelos (ELECTRA de Clark et al. (2020), RoBERTa de Liu et al. (2021), XLM-R de Conneau et al. (2020), Multilingual BERT de Devlin et al. (2019) e BERTimbau de Souza et al. (2020a)) para detectar notícias falsas em redes sociais. O modelo BERTimbau (Souza et al., 2020a) com 6 épocas foi o que obteve acurácia e medida-F superior a todos outros os modelos utilizados nos experimentos, medida-F de 95%.

16.3.4 Detecção de Ironia/Sarcasmo/Humor

Apesar dos avanços na área de Análise de Sentimentos (Seção 16.3.2), ela ainda se depara com vários desafios. Entre eles, destaca-se o entendimento de figuras de linguagem (ironia/sarcasmo/humor). As figuras de linguagem são difundidas em quase qualquer gênero de texto e são especialmente comuns nos textos da Web e das redes sociais, em plataformas como o Twitter/X (Ghosh et al., 2015).

Existe uma linha tênue entre os conceitos de ironia, sarcasmo e humor. Reyes et al. (2012) define a ironia como uma “ligeira fronteira no significado do sarcasmo e da sátira”. Gibbs; Colston (2001) afirmam que o sarcasmo, combinado com jocosidade, hipérbole, perguntas retóricas e eufemismo, são tipos de ironia. Tradicionalmente, a ironia é conhecida como o oposto do significado literal (Grice, 1975).

Dentre os trabalhos aplicados à tarefa de detecção de ironia para a língua portuguesa, podemos citar: Carvalho et al. (2009), Freitas et al. (2014) e Silva (2018). Nos trabalhos de Carvalho et al. (2009) e Freitas et al. (2014) são propostas pistas linguísticas para detectar ironia. Em Carvalho et al. (2009), os autores mostraram que é possível identificar opiniões irônicas em comentários, com precisão relativamente alta (de 45% a 85%), usando padrões linguísticos relativamente simples, tais como: emoticons, expressões onomatopeicas para risos, sinais de pontuação, aspas e interjeições positivas (“viva”, “parabéns”, “força” etc.). Ainda, em Freitas et al. (2014) padrões linguísticos foram aplicados no *corpus* sobre o assunto “fim do mundo” extraído do Twitter/X (Exemplo 16.20 de tweet irônico com emoticons e Exemplo 16.21 de tweet irônico com sinais de pontuação e expressões onomatopeicas para risos presentes no *corpus* sobre o assunto “fim do mundo”).

Exemplo 16.20:

O fim domundo não chegou ainda porque esta engarrafado naBR116. #Gau-chanoFimdoMundo :)

Exemplo 16.21:

Todo mundo tirando sarro do fim do mundo...oquei!!!...mas quero ver amanhã, qualquer temporalzinhovai todo mundo se mijar de medo...hahaha

Esses padrões estão relacionados ao português brasileiro, mas a metodologia pode ser facilmente transferida para análises em outras línguas. Isso foi feito no trabalho de Freitas et al. (2020), no qual um subconjunto de padrões foi testado em *corpora* de diferentes idiomas (inglês, italiano e espanhol). Em Silva (2018), o autor descreve sobre o processo de geração de um *corpus* de ironia para a língua portuguesa, bem como, a criação de um modelo pré-treinado de uma Rede Neural Convolucional (em inglês, *Convolutional Neural*



Network ou CNN) para detectar ironia. A CNN foi capaz de adaptar-se e detectar automaticamente as figuras de linguagem em questão. Tal abordagem mostrou-se satisfatória para detecção de ironia, obtendo medida-F de 89,78%.

No ano de 2021, foi proposto um desafio sobre detecção de ironia para língua portuguesa no IberLEF⁴⁶ denominado IDPT⁴⁷. A proposta do IDPT foi inspirada em competições proposta em outros idiomas, como SemEval (Hee et al., 2018) para o inglês, IronITA (Cignarella et al., 2018) para o italiano, IroSvA (Bueno et al., 2019) para o espanhol e IDAT (Ghanem et al., 2019) para o árabe. Participaram da tarefa seis equipes de universidades e de empresas de quatro diferentes países: Brasil, China, Portugal e Espanha. Os *corpora* disponibilizados na competição contêm textos (tweets e notícias) sobre diferentes temas. Exemplo 16.22 de notícia irônica, Exemplo 16.23 de notícia não irônica, Exemplo 16.24 de tweet irônico e Exemplo 16.25 de tweet não irônico contidos no *corpus* do desafio IDPT.

Exemplo 16.22:

A criatura chegou ao Rio de Janeiro e deve conceder entrevista na tarde de hoje.

Exemplo 16.23:

A superlotação de hospitais da região metropolitana de Belo Horizonte preocupa a capital, já que a cidade é referência hospitalar da macrorregião Central e, caso não haja vagas na cidade de origem, os pacientes são transferidos para BH. No momento, cerca de 30% dos leitos de terapia intensiva e enfermaria da capital estão ocupados com pacientes de outras cidades.

Exemplo 16.24:

Agora deu vontade de me alistar... #sqn #exercitosuperfaturaatecerveja
https://t.co/Pp6vx36jzN

Exemplo 16.25:

Agora a festa dos europeus comemorando o Mundial é contagiante hein?

O conjunto de dados de treinamento foi desenvolvido por Freitas et al. (2014), Silva (2018) e Schubert; Freitas (2020). Os participantes usaram abordagens tradicionais de AM (como: SVM, NB e outros) e/ou AP (como: Transformers). Os times que atingiram os melhores resultados foram o BERT4EVER (Jiang et al., 2021) e PiLN (Anchiêta et al., 2021). BERT4EVER (Jiang et al., 2021) utilizou Transformers e obteve uma acurácia balanceada de 92% para conjunto de dados de notícias. Para o conjunto de dados composto por tweets, a equipe PiLN (Anchiêta et al., 2021) utilizou *superficial features* e SVM e obteve uma acurácia balanceada de 52%.

⁴⁶<https://sites.google.com/view/iberlef2021/>

⁴⁷<https://sites.google.com/inf.ufpel.edu.br/idpt2021/>



16.4 Considerações finais

Falar sobre aplicações de PLN em redes sociais é de grande importância por diversas razões. As redes sociais desempenham um papel fundamental na comunicação e interação social na sociedade moderna. Compreender como o PLN é aplicado nessas plataformas é essencial para entender as dinâmicas sociais e o impacto da tecnologia na vida das pessoas. Este capítulo forneceu uma visão geral sobre aplicações de abordagens de PLN em conteúdos de redes sociais. Damos ênfase ao desenvolvimento de pesquisas desenvolvidas com foco na língua portuguesa, dado o foco deste livro e de esta ser ainda uma língua com recursos escassos para algumas tarefas. Nessa primeira versão, deixamos de cobrir tópicos relevantes e atuais como reconhecimento/classificação de emoções, rastreamento de transtorno mental, e detecção de postura. Reconhecendo a importância destes tópicos, pretendemos contemplá-los na versão seguinte deste livro.



Capítulo 17

PLN e Responsabilidade Social

Aplicações da FrameNet-Br

Maucha Andrade Gamonal

Livia Vicente Dutra

Mariane de Carvalho Pinto

Tiago Timponi Torrent

Publicado em: 16/04/2026

17.1 Introdução

Este capítulo apresenta um panorama da aplicação da Semântica de *Frames* e da infraestrutura computacional da FrameNet Brasil em tarefas de Processamento de Linguagem Natural (PLN) voltadas à construção de sentido em aplicações práticas, como no enfrentamento à violência de gênero e na acessibilidade cultural. No Capítulo [Semântica com técnicas simbólicas](#), as autoras discutem métodos simbólicos voltados ao processamento e à compreensão de textos em linguagem natural, apresentando três modelos como exemplos de bases de conhecimento semântico: a FrameNet (Ruppenhofer et al., 2016), a WordNet (Fellbaum, 1998) e a ConceptNet (Speer et al., 2017). A FrameNet e a WordNet, bem como suas versões em português, são classificadas como recursos léxico-semânticos, enquanto a ConceptNet é caracterizada como uma base de conhecimento de senso comum. Naquele capítulo, as principais referências teóricas e os conceitos fundamentais dessas abordagens são detalhados e contextualizados, oferecendo um suporte conceitual para as aplicações que apresentamos aqui. Assim, ao longo da Seção [17.2](#), retomaremos e aprofundaremos aspectos do modelo FrameNet Brasil introduzidos anteriormente, além de inserir novos conceitos necessários para sustentar as aplicações descritas.

Partindo da teoria desenvolvida por Fillmore (1982) e da FrameNet como sua aplicação computacional, discutimos como os *frames* e suas relações semânticas estruturam a compreensão linguística e servem de base para a anotação e o processamento de dados em português brasileiro. A tarefa de PLN que orienta este capítulo é a **anotação semântica baseada em *frames***, aplicada a diferentes domínios e modalidades comunicativas. Trata-se de uma tarefa central para o desenvolvimento de modelos computacionais capazes de operar sobre dados reais em contextos sensíveis como a saúde pública e a acessibilidade audiovisual.

O capítulo busca mostrar como demandas sociais do dia a dia podem ser objetos de aplicações computacionais ao envolverem desde a identificação de padrões de violência de gênero em registros de saúde até a avaliação de traduções intersemióticas em produtos audiovisuais. A primeira parte do texto apresenta os fundamentos teóricos da Semântica de *Frames*, detalhando o funcionamento da FrameNet Brasil e seus refinamentos semânti-



cos, como as Relações Qualia Ternárias¹ que se dão por meio da ferramenta de anotação Webtool. Em seguida, são discutidos dois estudos de caso que ilustram o potencial da anotação semântica para enfrentar problemas reais.

O primeiro estudo está relacionado ao projeto **DataToStopGBV**, voltado à identificação de casos de violência baseada em gênero a partir de registros do sistema público de saúde. Por meio da modelagem de domínios específicos (saúde e violência), da modelagem de *frames* e da anotação manual e automática dos dados, o projeto mostra como a estrutura semântica da FrameNet pode ser empregada para apoiar políticas públicas de prevenção à violência.

O segundo estudo apresenta o *dataset Audition*, um recurso multimodal anotado semanticamente com o objetivo de apoiar a geração e a avaliação de traduções audiovisuais acessíveis, como a audiodescrição. A partir da articulação entre modos comunicativos, como áudio original, vídeo e audiodescrição, o *Audition* permite análises de similaridade semântica e contribui para o avanço de tecnologias assistivas baseadas em teoria linguística.

Apesar de compartilharem a mesma base teórica e metodológica – a Semântica de *Frames* e a FrameNet Brasil –, os dois estudos de caso exploram abordagens de anotação semântica distintas: enquanto o projeto **DataToStopGBV** faz uso de **anotações semânticas automáticas**, com apoio de modelos supervisionados, o *dataset Audition* foi construído com **anotação manual** especializada sobre os dados multimodais. Essa diferença reflete tanto os objetivos como as particularidades de cada tarefa.

Ao integrar teoria linguística, recursos computacionais e compromisso social, o capítulo demonstra como a Semântica de *Frames* pode orientar o desenvolvimento de ferramentas e metodologias eficazes para o processamento semântico em português brasileiro, com aplicações relevantes na saúde pública e na acessibilidade cultural.

17.2 Semântica de *Frames*

Como apresentado no Capítulo **E o significado?**, a Semântica ocupa-se do estudo do significado expresso por palavras, expressões e enunciados. Dentro desse campo, a Semântica de *Frames*, também discutida na Subseção **FrameNet**, é uma teoria desenvolvida por Charles J. Fillmore (Fillmore, 1982), que propõe uma maneira particular de se entender o significado lexical. Segundo essa abordagem, compreender o sentido de um item lexical envolve mais do que conhecer sua definição no dicionário: é necessário também ativar um conjunto de conhecimentos socialmente compartilhados sobre o mundo e sobre as situações em que ele costuma ser usado. Esse conjunto é chamado, na teoria, de *frame*.

Na definição de Fillmore (1982, p. 111), um *frame* é “qualquer sistema de conceitos relacionados de tal maneira que, para entender qualquer um deles, é preciso entender toda a estrutura na qual ele se insere”². Em outras palavras, para o autor, o significado de uma palavra está associado a uma rede de conceitos interligados, que funciona como uma espécie de pano de fundo conceitual para a construção de sentido. Essa rede de conceitos, os *frames*, está ancorada na forma como experienciamos e categorizamos o mundo ao nosso redor e possibilita a atribuição de significado no momento da comunicação.

¹Conjunto de relações semânticas inspiradas na Teoria do Léxico Gerativo (Pustejovsky, 1995), utilizadas na FN-Br para modelar aspectos internos dos conceitos, como finalidade, composição e origem. Este tópico será abordado na Seção 17.3.2 deste capítulo.

²“[...] By the term 'frame' I have in mind any system of concepts related in such a way that to understand any one of them you have to understand the whole structure in which it fits.” [Tradução nossa]



Ao ouvir uma palavra como “medicar”, por exemplo, não pensamos apenas na ação de administrar um remédio mas também em tudo que geralmente a acompanha: há um profissional de saúde, um medicamento, uma condição a ser tratada. O interessante é que, mesmo quando esses elementos não aparecem explicitamente na frase, eles ainda são mobilizados para a sua compreensão. Se alguém diz “O plantonista medicou a criança”, por exemplo, já se pressupõe que a criança estava com algum sintoma ou doença, que foi administrada uma substância com fins terapêuticos e que esse ato se deu num contexto de cuidado em saúde.

Isso acontece porque vivemos em uma sociedade em que as práticas de atenção à saúde são parte do cotidiano, ou seja, esse tipo de situação bem como os participantes e as circunstâncias que a compõem integram nosso repertório sociocultural. Assim sendo, quando o verbo “medicar” é usado, ele ativa uma rede de conhecimentos compartilhados, isto é, o *frame* que permite tal entendimento. Na Semântica de *Frames*, diz-se que o verbo “medicar” é uma Unidade Lexical que **evoca** um *frame* do cenário da saúde.

Ao destacar os conhecimentos compartilhados que sustentam o significado das palavras, a Semântica de *Frames* amplia a visão tradicional sobre a compreensão linguística. Em vez de analisar apenas sentenças isoladas, essa abordagem valoriza o papel da experiência humana e do contexto cultural para explicar como usamos e entendemos a linguagem no cotidiano. Por isso, Fillmore (1985) caracteriza essa teoria como uma **semântica da compreensão**, em contraste com a semântica das condições de verdade, que se concentra em avaliar se uma sentença é verdadeira ou falsa com base na realidade. Para ele, a linguagem não é simplesmente um espelho do mundo, mas uma ferramenta por meio da qual construímos e negociamos sentidos, ancorados na cognição e nas práticas culturais. Por isso, de modo geral, a teoria propõe uma pergunta central: como as pessoas organizam sua experiência por meio da linguagem? Para respondê-la, os estudos em Semântica de *Frames* analisam como as Unidades Lexicais são utilizadas em contextos autênticos e quais *frames* são acionados nesses usos.

Desde 1997, essa abordagem tem fundamentado a construção da FrameNet³ (Fillmore et al., 2003), projeto de lexicografia computacional que descreve o léxico com base nos *frames* evocados pelas ULs em ocorrências reais extraídas de *corpora*. Fundada por Charles J. Fillmore em 1997, no International Computer Science Institute (ICSI), localizado em Berkeley, Califórnia, a iniciativa se apoia na Semântica de *Frames* (Fillmore, 1982; Fillmore, 1985) e tem como objetivo investigar como os sentidos são construídos nas línguas naturais. O projeto tem como foco o desenvolvimento de um banco de dados lexical capaz de descrever a língua inglesa sob uma perspectiva semântica e sintática, utilizando exemplos reais extraídos de *corpora*. Como desdobramento desse trabalho, Fillmore também contribuiu para a proposição do **Constructicon**⁴, recurso voltado à descrição de construções gramaticais que evocam *frames*, e que busca representar a integração entre forma e sentido em estruturas mais complexas e idiomáticas.

Com o tempo, os recursos criados pela FrameNet começaram a ser adotados em diferentes áreas, incluindo o avanço de aplicações em linguística computacional, especialmente em tarefas de PLN, e o projeto foi expandido para outras línguas. Um dos principais exemplos dessa expansão é a FrameNet Brasil (FN-Br), um laboratório de linguística computacional

³A ferramenta de consulta da FrameNet original, bem como informações sobre os dados anotados, estão disponíveis em: <https://framenet.icsi.berkeley.edu/>.

⁴Seu desenvolvimento foi adotado por diferentes framenets, como as do japonês, sueco e português brasileiro, embora não tenha avançado da mesma maneira no projeto original em inglês. A versão do *Constructicon* da FrameNet Brasil pode ser consultada em: <https://webtool.frame.net.br/report/cxn>.



vinculado à Universidade Federal de Juiz de Fora (UFJF), em Minas Gerais. O projeto brasileiro desenvolve anotações semânticas para o português brasileiro e também propõe extensões à base de *frames* e à estrutura de dados originalmente concebidas pela Berkeley FrameNet.

17.3 FrameNet Brasil

Para compreender a metodologia de anotação utilizada pela FN-Br, é fundamental conhecer as nomenclaturas essenciais ao projeto. Algumas foram apresentadas na Subseção *FrameNet* e outras serão detalhadas ao longo desta seção. A primeira delas é o ponto de partida de toda anotação, a **Unidade Lexical** (UL), que corresponde à associação de um lema ao *frame* evocado por ele. Na FN-Br, o lema “medicar.v” está associado ao *frame* *Intervenção_em_saúde*⁵, apresentado na Figura 17.1.

Figura 17.1: O *frame* *Intervenção_em_saúde*

Intervenção_em_saúde		@Health	#1155	Medical_intervention [en]	PDF
Definition					
As intervenções baseadas em procedimentos ou em medicina são usadas no Paciente para tentar aliviar uma Condição_médica . Essas intervenções podem ter uma Frequência_de_sucesso e Efeitos_colaterais . Esse frame difere-se de Cura, já que este frame lida apenas com tentativas de aliviar uma Condição_médica , enquanto Cura lida com situações em que a aflição ou a Condição_médica foi curada.					
Frame Elements					
Core					
Condição_em_saúde	Uma descrição holística do estado médico do paciente (ou uma parte do estado do paciente), que pode ou não indicar a causa de um desvio do normal.				
Intervenção	Um medicamento ou procedimento é administrado ou realizado para tratar uma Condição_médica .				
Paciente	Indivíduos que recebem Intervenção por causa de alguma condição em saúde.				
Profissional_em_saúde	O indivíduo ou equipe que tenta melhorar a condição médica do Paciente .				
Resultado	A consequência da Intervenção .				
Peripheral					
Efeitos_colaterais	Efeitos da Intervenção que são não-intencionais ou indesejáveis.				
Extensão	O grau em que um Intervenção afetou a Condição_médica ou os sintomas.				
Frequência_de_sucesso	A regularidade com a qual uma Intervenção é eficaz para tratar uma Condição_médica ou os sintomas.				
Tempo	Momento em que a intervenção acontece.				

Fonte: Webtool 4.1

Na Figura 17.1, é possível visualizar a estrutura básica de um *frame* na interface gráfica

⁵De acordo com a metodologia das FrameNets, os nomes dos *frames* são grafados em fonte Courier.

da FrameNet Brasil, a **Webtool**⁶ (Torrent et al., 2024), plataforma responsável pelo gerenciamento do banco de dados da FN-Br. Na ferramenta, os anotadores têm acesso aos *frames* e às ULs do banco de dados e podem interagir com esses elementos diretamente na interface, o que facilita a análise e a anotação.

No topo da página de cada entrada, é exibido o nome do *frame*, seguido por uma breve definição que o caracteriza. No caso do *frame* **Intervenção_em_saúde**, por exemplo, trata-se de um evento em que um profissional da área da saúde realiza uma intervenção médica, como administrar um medicamento, com o objetivo de tratar ou aliviar uma condição ou doença. Logo abaixo da definição, são apresentados os participantes, os instrumentos ou as circunstâncias que compõem a cena representada pelo *frame*, denominados de **Elementos de Frame** (EFs). Na FrameNet, os EFs são divididos em dois grupos: os **nucleares**, que são essenciais para a composição do *frame*, e os **não nucleares**, que desempenham funções circunstanciais.

No *frame* **Intervenção_em_saúde**, são apresentados como EFs nucleares o PROFISSIONAL_EM_SAÚDE⁷, que realiza os procedimentos, o PACIENTE, que sofre a condição ou doença, a CONDIÇÃO_EM_SAÚDE, que motiva a intervenção, a própria INTERVENÇÃO e o RESULTADO obtido. Os outros elementos que aparecem na Figura 17.1, como EFEITOS_COLATERAIS e TEMPO, são considerados não nucleares.

Para realizar a anotação de texto corrido⁸ – modelo de anotação mais seguido atualmente na FN-Br –, os anotadores do projeto recebem lotes de sentenças para análise. Conforme a metodologia descrita por Ruppenhofer et al. (2016), que fundamenta o modelo adotado pela FrameNet, a tarefa central nesse tipo de anotação é identificar, em cada sentença, as ULs e os *frames* que elas evocam. Na anotação, ao selecionar uma UL na interface, são exibidos os *frames* associados a ela na base de dados. O anotador, então, lê as descrições desses *frames* para escolher o mais adequado ao contexto da sentença e, a partir dessa escolha, o sistema gera uma camada para marcar os Elementos de *Frame* (EFs), que devem ser destacados manualmente no texto.

Durante o processo de anotação, pode acontecer de EFs nucleares não serem instanciados explicitamente como material linguístico nos enunciados em análise. Quando isso acontece, dizemos que houve uma **instanciamento nulo**. Na FrameNet, esse fenômeno é dividido em três tipos: **Instanciamento Nulo Definido** (DNI, do inglês *Definite Null Instantiation*), **Instanciamento Nulo Indefinido** (INI, do inglês *Indefinite Null Instantiation*) e **Instanciamento Nulo Construcional** (CNI, do inglês *Constructional Null Instantiation*).

A primeira delas, a DNI, diz respeito à omissão de um EF que pode ser recuperado com base no contexto. Na sentença do Exemplo 17.1⁹, por exemplo, o verbo “observar” está anotado no *frame* **Percepção_ativa**, que tem dois EFs nucleares: o PERCEPTOR_AGENTIVO, que é aquele que realiza a ação de observar, e o FENÔMENO, que é aquilo que é percebido.

Exemplo 17.1:

Um cachorro **observa**^{PERCEPÇÃO_ATIVA} de longe.

Exemplo 17.2:

⁶A ferramenta de anotação semântica WebTool, bem como os dados anotados, estão disponíveis em: <https://webtool.frame.net.br/>. O repositório oficial da FrameNet Brasil no Github pode ser acessado em: <https://github.com/FrameNetBrasil>.

⁷Seguindo a metodologia das FrameNets, nomes de EFs são apresentados em VERSALETE.

⁸Ao longo do texto, as Figuras 17.9 e 17.14 discutem exemplos de anotações de texto corrido.

⁹As sentenças que aparecem ao longo desta seção para ilustrar os conceitos discutidos advêm de curtas-metragens do Audition e foram anotadas por meio da Webtool.



As pessoas descem.

A sentença do Exemplo 17.1, no entanto, expressa apenas o PERCEPTOR_AGENTIVO (“um cachorro”) enquanto o FENÔMENO está ausente. Como aquilo que é observado (“as pessoas”) pode ser facilmente recuperado pelo contexto discursivo – mais especificamente pela sentença do Exemplo 17.2, que ocorre logo antes no *corpus* –, essa omissão é classificada como DNI.

Já a INI ocorre quando um EF nuclear não está explicitamente expresso na sentença nem pode ser determinado com base no contexto imediato. Isso faz com que o elemento ausente seja interpretado de maneira genérica ou indefinida. Na sentença do Exemplo 17.3, por exemplo, o *frame* **Pessoas_por_idade** é evocado pela UL “crianças.n”¹⁰, e um de seus EFs nucleares é a IDADE, que, nesse caso, não está lexicalmente realizado nem pode ser inferido com precisão a partir do contexto. No entanto, como o *frame* é formado por ULs que já pressupõem uma faixa etária aproximada, a ausência do EF IDADE pode ser considerada INI e DNI. Isso demonstra que, em certos casos, há sobreposição entre os dois tipos de instanciação nula, exigindo decisões interpretativas do anotador.

Exemplo 17.3:

As **crianças**^{PESSOAS_POR_IDADE} se reúnem ao redor do globo.

A Instanciação Nula Construcional (CNI), por sua vez, ocorre quando a própria estrutura da língua licencia a omissão de um EF nuclear. No português brasileiro, isso acontece, por exemplo, nas sentenças com sujeito desinencial ou nas construções imperativas. É o caso da sentença do Exemplo 17.4, em que, por se tratar de uma oração com sujeito desinencial, a expressão do agente não é obrigatória no português. Nesse caso, o verbo “brincar” evoca o *frame* **Atividade**, cujo único EF nuclear é o AGENTE, ou seja, aquele que realiza a ação. No exemplo, no entanto, o AGENTE da ação está lexicalmente omitido pela estrutura. Por isso, essa omissão é classificada como uma ocorrência de CNI.

Exemplo 17.4:

Brincam^{ATIVIDADE} de lançar as bolinhas dentro da lata.

Existem casos, ainda, em que a própria UL que evoca o *frame* já inclui um dos EFs em seu radical, caracterizando a ocorrência de Incorporação (INC). Na sentença do Exemplo 17.5, por exemplo, o lema “elefante” está associado ao *frame* **Animais**, cujo EF nuclear é o próprio animal referido. Nesse caso, não há um termo separado na sentença que realize esse EF, pois ele já está presente na própria UL “elefante.n”, configurando assim uma INC (do inglês, *incorporation*).

Exemplo 17.5:

O **elefante**^{ANIMAIS} azul.

Essa questão da ausência linguística ganha novos contornos quando se consideram dados multimodais, como imagens e vídeos. Em contextos audiovisuais, elementos que parecem ausentes no texto verbal muitas vezes estão presentes na imagem, o que desafia os limites

¹⁰Na nomenclatura das FrameNets, antes do ponto é indicado o lema e, após o ponto, a classe gramatical: .n (substantivo, *noun*), .v (verbo, *verb*), .a (adjetivo, *adjective*), entre outros.



da anotação textual tradicional. Estudos realizados em nosso laboratório, como o de Belcavello (2023), demonstraram que objetos visuais também são capazes de evocar *frames* ou complementar o que foi evocado pela narração. Diante disso, tornou-se necessário desenvolver um tipo específico de anotação que integrasse essas diferentes modalidades semióticas. Essa anotação multimodal será abordada na última seção deste capítulo, em que se destaca sua importância para ampliar a representação semântica na FrameNet Brasil e, inclusive, repensar o escopo das instanciações nulas.

Considerando a estrutura interna dos *frames* e de seus elementos apresentada até aqui, a FrameNet Brasil categoriza os *frames* em cinco grandes grupos, definidos a partir da natureza do conteúdo semântico que representam (Torrent et al., 2022). No Lutma¹¹, recurso de criação de *frames* da FrameNet Brasil para não especialistas, eles são definidos da seguinte forma:

- (i) Entidade: um conceito, uma coisa, um ser vivo.
- (ii) Evento: envolve pelo menos um participante e ocorre em um local e hora.
- (iii) Estado: uma situação estável.
- (iv) Relação: uma relação entre duas ou mais entidades, eventos ou estados.
- (v) Atributo: um atributo com um valor é aplicado a uma entidade.

Na prática, esses grupos podem ser ilustrados com exemplos concretos presentes na própria base da FN-Br. O grupo **Entidade**, por exemplo, inclui o *frame* **Pessoas**, que é evocado por ULs gerais para indivíduos (como “povo.n” e “gente.n”). O grupo **Evento** pode ser representado por **Fazer_exame_em_Saúde**, que diz respeito ao ato de realizar um exame médico e é evocado por ULs como “aferir.v” e “coletar.v”. Já o grupo **Estado** é exemplificado por **Condições_em_Saúde**, formado por ULs como “diabético.a” e “cardiopata.a”. O grupo **Relação**, por sua vez, inclui *frames* como **Parentesco**, que reúne ULs que expressam relações familiares (como “irmão.n” e “mãe.n”), enquanto o grupo **Atributo** pode ser exemplificado com o *frame* **Cor**, que associa características como “azul.a” e “acinzentado.a” a uma entidade.

Além de contribuir para a sistematização da base de dados da FrameNet Brasil, essa divisão também facilita a compreensão do conteúdo e do uso de cada *frame*, antecipando o tipo de estrutura conceitual que será encontrada. Porém, para entender como os *frames* se conectam e formam uma rede, é preciso olhar para as relações entre os *frames*, que serão discutidas a seguir.

17.3.1 As relações *frame-a-frame*

Na FrameNet, cada *frame* faz parte de uma grande rede de significados. Isso quer dizer que, ao modelar um novo *frame*, é possível conectá-lo a outros já existentes, formando um sistema interligado. Essas conexões são chamadas de **relações *frame-a-frame*** e são fundamentais para mostrar como eles se organizam e se relacionam na rede.

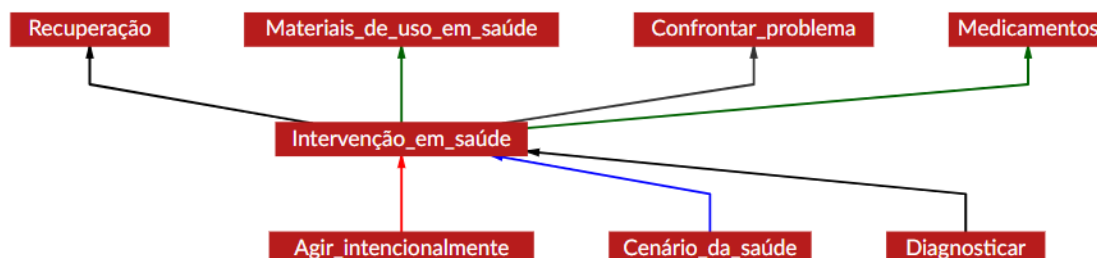
As principais categorias dessas relações são: Herança, Subframe, Uso, Precedência, Causativo_de, Incoativo_de e Perspectiva. Cada uma delas indica um tipo específico de ligação entre os *frames* – como quando um *frame* mais geral se conecta a um mais específico, por exemplo, ou quando dois *frames* representam diferentes pontos de vista de uma

¹¹Disponível em: <https://lutma.frame.net.br/>. Acesso em: 28 jul. 2025.



mesma situação. Na FrameNet Brasil, essas relações podem ser visualizadas por meio do **Grapher**, uma ferramenta gráfica disponível na WebTool. No Grapher, cada tipo de relação aparece com uma cor diferente, facilitando a identificação e o estudo dessas conexões. A Figura 17.2, por exemplo, ajuda a visualizar como o *frame* **Intervenção_em_saúde** se conecta a outros na base.

Figura 17.2: Relações do *frame* **Intervenção_em_saúde**



Fonte: Webtool 4.1

A relação de **Herança** (*Inheritance*), representada pela seta vermelha, indica quando um *frame* filho herda as características de um *frame* mãe, elaborando um ou mais de seus EFs. Um exemplo disso é **Intervenção_em_saúde**, que herda as características de **Agir_intencionalmente** (veja Figura 17.3).

Figura 17.3: O *frame* **Agir_intencionalmente**

Fonte: Webtool 4.1

Acerca disso, em **Agir_intencionalmente**, o EF PESSOA é definido de forma genérica, como qualquer indivíduo que realiza uma ação. Já no *frame* **Intervenção_em_saúde**, esse papel é especificado: o agente da ação passa a ser o **PROFISSIONAL_EM_SAÚDE**, caracterizado como alguém que atua com a intenção de melhorar a condição clínica de um paciente. A mesma situação se dá com outros elementos herdados: o que antes era simplesmente a **AÇÃO**, por exemplo, passa a ser identificada como a **INTERVENÇÃO** realizada. Essa

especificação ajuda a representar de forma mais detalhada o tipo de situação descrita pelo novo *frame*.

A relação de **Subframe** (*Subframe*), indicada pela seta azul, conecta um *frame* que codifica um evento complexo a outros que representam subeventos delimitáveis (ou seja, partes) desse evento. É o caso de **Cenário_da_saúde**, que representa situações em que um paciente precisa de cuidados médicos. Na FrameNet Brasil, esse cenário está relacionado, por meio da relação de Subframe, a *frames* como **Diagnosticar**, **Intervenção_em_saúde** e **Recuperação**, que descrevem, respectivamente, as etapas de diagnóstico, realização da intervenção e recuperação do paciente, momentos específicos desse cenário.

Os *frames* **Diagnosticar**, **Intervenção_em_saúde** e **Recuperação** se articulam também por meio da relação de **Precedência** (*Precedes*), indicada pelas setas pretas na Figura 17.2, que organiza os eventos em uma sequência temporal. No caso, **Diagnosticar** ocorre antes da intervenção, que é seguida pela **Recuperação**, formando assim uma ordem cronológica que reflete o processo dos cuidados médicos ao longo do tempo.

Por sua vez, a relação de **Uso** (*Using*), indicada pelas setas verdes, ocorre quando a compreensão de um *frame* depende de outro que funciona como seu contexto de fundo. É o caso de **Medicamentos** e **Materiais_de_uso_em_saúde**, por exemplo, que pressupõem a cena da **Intervenção_em_saúde** para serem plenamente compreendidos, visto que tanto os medicamentos quanto os materiais são utilizados no contexto de uma ação médica. Assim, **Intervenção_em_saúde** fornece o cenário necessário para interpretar o uso desses elementos, dando sentido às práticas que envolvem seu manuseio ou aplicação.

As relações **Causativo_de** (*Causative_of*), **Incoativo_de** (*Inchoative_of*) e **Perspectiva** (*Perspective_on*), por outro lado, não podem ser demonstradas a partir do *frame* **Intervenção_em_saúde**. Por isso, para exemplificá-las, é necessário recorrer a outros *frames*.

A relação **Causativo_de**, representada por uma seta amarela no Grapher, indica um vínculo de causalidade entre dois *frames*. Por exemplo, na FN-Br, o *frame* **Causar_condição_em_saúde**, que indica a ação de causar (intencionalmente ou não) uma condição em saúde, é causativo do *frame* **Condições_em_saúde**, que reúne ULs relacionadas a doenças ou condições que um paciente apresenta.

Já a relação **Incoativo_de**, representada por uma seta marrom, marca uma mudança de estado. O *frame* **Condições_em_saúde**, por exemplo, tem como incoativo o *frame* **Tornar-se_paciente_em_saúde**, que descreve a transição de um indivíduo, a partir de um diagnóstico, para a condição de estar doente.

A relação de **Perspectiva**, por fim, acontece quando diferentes *frames* representam a mesma situação sob pontos de vista distintos. Ela pode ser exemplificada a partir do *frame* **Emoções**, que descreve um experienciador em determinado estado emocional, geralmente provocado por algum estímulo. A partir dele, surgem duas perspectivas distintas: **Foco_no_estímulo**, que enfatiza o estímulo responsável por causar a emoção, e **Emoção_com_foco_no_experienciador**, que destaca o ponto de vista de quem sente a emoção em relação ao que a provocou.

As relações *frame-a-frame* ajudam a entender como os *frames* se organizam e se ligam dentro da rede semântica da FrameNet Brasil, mostrando diferentes formas de associação entre eles. No entanto, nem todas as conexões importantes acontecem nesse nível. Em muitos casos, é preciso observar ligações mais específicas.

Nesse sentido, para ampliar o refinamento semântico da base, a FrameNet Brasil incluiu novos tipos de relações. Um exemplo é a **relação EF-frame** (*FE-to-frame*), que modela o fato de que um EF, em determinado *frame*, pode estar conceitualmente associado a

outro *frame* na rede. Em outras palavras, essa relação expressa que certos EFs, ao serem instanciados por ULs, tendem a evocar outros *frames*. Assim, se **Intervenção_em_saúde** inclui o EF PACIENTE, como vimos, a relação EF-*frame* pode indicar que esse PACIENTE frequentemente remete ao *frame* Pessoa, já que, em geral, os pacientes são pessoas.

Nesse mesmo esforço de ampliar a capacidade descritiva da rede, também foi proposta a **relação metonímica**. Gamonal (2017) propõe que a metonímia seja representada na FN-Br por meio de relações entre EFs de um mesmo *frame*, permitindo capturar casos em que um EF substitui outro com base em inferências pragmáticas apoiadas em conhecimento compartilhado. Por exemplo, em uma sentença como “O Hospital Central atendeu dezenas de pacientes ontem”, a expressão “Hospital Central” é usada metonimicamente para se referir à equipe de profissionais da saúde que prestam o atendimento naquele local.

Para representar essa relação na base, primeiramente, mapeia-se que o EF PROFISSIONAL_DE_SAÚDE – no *frame* Atendimento_em_saúde – pode ser instanciado em Profissionais_em_saúde; em seguida, estabelece-se uma relação metonímica entre os EFs PROFISSIONAL e LUGAR_DE_TRABALHO nesse *frame*, indicando que o local onde o profissional atua pode ser usado para representá-lo. Temos primeiramente uma relação entre *frames* (EF - *Frame*) seguida de uma relação metonímica considerada *intraframe* por se dar entre EFs de um mesmo *frame*. Na literatura da Linguística Cognitiva, considera-se que haja um mapeamento “intradomínio”.

Além dessas, outras relações foram desenvolvidas com foco nas conexões entre ULs de *frames* distintos, a fim de capturar nuances semânticas importantes, principalmente em domínios mais especializados. Nesse cenário, ganham destaque as relações qualia. Elas foram incorporadas pela FN-Br com base na proposta de Pustejovsky (1995) e servem para capturar sentidos mais detalhados, especialmente quando se trata de áreas mais técnicas ou específicas. Como esse tipo de relação desempenha um papel central na proposta de anotação discutida neste trabalho, ele será abordado de forma mais detalhada na próxima seção.

17.3.2 As relações qualia ternárias

Além das relações *frame-a-frame*, que foram inicialmente desenvolvidas pela Berkeley FrameNet, a FrameNet Brasil adicionou um novo tipo de relação semântica à sua base: as relações qualia. Elas foram implementadas com base nos estudos de Pustejovsky (1995), cuja ideia central é a de que o significado de uma palavra pode depender de diferentes formas pelas quais as pessoas entendem e relacionam os objetos ao seu redor. Para Pustejovsky (1995), essas formas são organizadas em quatro tipos de relações, definidas como apresentado no Quadro 17.1.

Quadro 17.1: Relações qualia ternárias (Torrent et al., 2022, p. 8)¹²

Formal – descreve a categoria básica do item e fornece a informação que distingue uma entidade dentro de um conjunto maior dentro de seu domínio semântico. Constitutivo – expressa uma variedade de relações sobre a constituição interna de uma entidade.

¹²Formal – describes the basic category for the item and provides the information that distinguishes an entity within a larger set inside its semantic domain.

Constitutive – expresses a variety of relations concerning the internal constitution of an entity.

Telic – concerns the typical function or purpose of an entity, i.e., what the entity is for.

Agentive – concerns the origin of an entity, its creator, or its coming into being. [Tradução nossa]

Télico – diz respeito à função ou propósito típico de uma entidade, ou seja, para que a entidade serve.

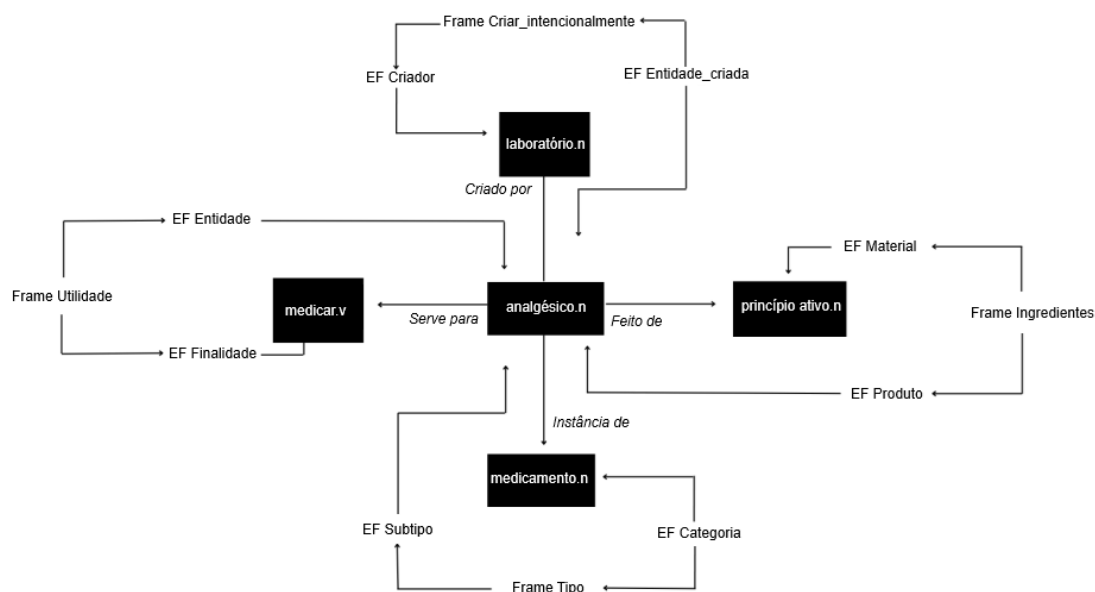
Agentivo – refere-se à origem de uma entidade, seu criador ou seu processo de surgimento.

Essas quatro dimensões – formal, constitutiva, télica e agentiva – ajudam a explicar como os sentidos das ULs se constroem e se conectam no uso da linguagem. Na FN-Br, foi adotada uma forma própria de representar essas relações semânticas em seu banco de dados. Em vez de simplesmente indicar que duas Unidades Lexicais estejam conectadas por uma relação qualia, o projeto propôs uma modelagem mais detalhada, utilizando *frames* para mediar essas conexões. Essa abordagem, chamada de **qualia ternária**, foi desenvolvida por Costa (2020) e permite representar essas relações de forma mais precisa e estruturada.

Nesse modelo, cada relação qualia é representada dentro de um *frame* específico, que funciona como um mediador para a conexão entre duas ULs. Segundo Belcavello et al. (2020), as duas ULs envolvidas são associadas a diferentes EFs num mesmo *frame*. Assim, o *frame* estabelece uma ponte semântica entre os termos, deixando claro que tipo de relação existe entre eles e quais papéis cada um desempenha. Outro aspecto importante dessa modelagem é a direção da relação: ela é sempre unidirecional, ou seja, vai de uma Unidade Lexical para outra, e não o contrário.

Como apresentado por Dutra (2024), a aplicação das relações qualia pode ser ilustrada com o exemplo da UL “analgésico.n”. No banco de dados da FN-Br, essa unidade está conectada a outras cinco, por meio de diferentes tipos de relações qualia, e cada uma dessas conexões é mediada por um *frame* específico, como mostra a Figura 17.4.

Figura 17.4: As relações qualia a partir da UL *analgésico.n*



Fonte: Adaptado de (Dutra, 2024)

Por exemplo, a **relação agentiva** liga *analgésico.n* à UL “laboratório.n” usando o *frame Criar_intencionalmente*. Nesse caso, o analgésico é entendido como aquilo que

é criado (EF ENTIDADE_CRIADA), enquanto o laboratório aparece como o agente da ação (EF CRIADOR). A **relação constitutiva**, por outro lado, conecta “analgésico.n” e “princípio ativo.n” por meio do *frame** *Ingredientes*. Nele, o “analgésico.n” ocupa o papel de PRODUTO, e “princípio ativo.n”, o de MATERIAL, indicando do que o analgésico é feito. Já a **relação formal** é modelada no *frame** *Tipo*, em que “analgésico.n” aparece como um SUBTIPO e “medicamento.n” como uma CATEGORIA, representando a ideia de que analgésico é um tipo de medicamento. Por fim, a **relação télica** envolve o *frame* *Utilidade*. Nesse caso, “analgésico.n” é associado ao EF ENTIDADE (ou seja, um item com uma finalidade), enquanto a UL “medicar.v” representa sua FINALIDADE.

A organização das relações na FN-Br, tanto no nível dos *frames* quanto das Unidades Lexicais, constitui a base do trabalho de anotação desenvolvido no projeto. Para realizar essas anotações de forma padronizada e organizada, a equipe utiliza a WebTool. Na próxima seção, apresentamos um exemplo prático dessa aplicação por meio de um projeto que utiliza a FrameNet Brasil para apoiar pesquisas sobre o enfrentamento da violência contra a mulher.

Exercício prático: Explorando as relações qualia ternárias

Agora é o momento de aplicar os conceitos discutidos! Acesse a lista de *frames* disponíveis na [WebTool](#) e escolha um que lhe pareça interessante. Em seguida, selecione uma das Unidades Lexicais no *frame* (listadas logo após os EFs) e, com base nela, reflita sobre os aspectos que compõem seu significado. Para isso, **associe a UL escolhida às relações qualia que forem possíveis**. Para cada dimensão, pense em uma UL que esteja semanticamente relacionada à original e, sempre que possível, pesquise essa nova UL na WebTool para verificar a que *frame* ela pertence – ou seja, que *frame* poderia ser responsável por **mediar** essa ligação entre as duas ULs.

Exemplo: A UL “pizza.n” pertence ao *frame* *Criação_culinária* e se relaciona com “farinha.n”, que pertence ao *frame* *Ingredientes*. Essas duas ULs estão conectadas por meio desse *frame*, que faz a mediação de uma relação qualia do tipo **constitutiva**, já que *farinha* é parte da composição da *pizza*.

Para ajudar no processo, as perguntas do Quadro 17.2 foram organizadas segundo cada uma das dimensões qualia e podem servir como guia para sua análise.

Quadro 17.2: Perguntas para cada uma das dimensões qualia

Relação qualia	Pergunta-guia
Construtiva	Do que isso é feito? Quais são seus componentes?
Formal	O que o define como sendo daquele tipo?
Télica	Para que serve? Qual é sua finalidade ou função?
Agentiva	Quem o produziu ou como foi criado?

17.4 Projeto DataToStopGBV: Semântica contra a violência de gênero

O **projeto DataToStopGBV** é uma iniciativa que nasceu da colaboração entre Vital Strategies Brasil e a FrameNet Brasil em 2023. A Vital Strategies Brasil é a filial brasileira



da organização global de saúde Vital Strategies, que trabalha em conjunto com o governo e a sociedade buscando implementar medidas que minimizem desafios ligados a questões de saúde pública. A parceria entre a **Vital Strategies Brasil** e a **FrameNet Brasil** se deu com uma intenção clara: desenvolver ferramentas para alerta precoce e intervenção em casos de violência de gênero, que pudessem oferecer maior apoio às equipes de saúde, autoridades locais e criadores de políticas públicas.

Este projeto foi também o foco principal da pesquisa de mestrado “Evaluating The Contribution of Framenet to Gender-Based Violence Identification” (Dutra, 2024). Esta seção apresenta um panorama das discussões iniciadas no âmbito dessa pesquisa.

17.4.1 Contexto e motivação

A **Violência baseada em gênero (VBG)** afeta diversos grupos sociais, no entanto a incidência de casos da violência direcionada às mulheres faz com que este seja o maior foco de pesquisas e normalmente delimita as mulheres como grupo de foco. Seguindo, então, a mesma linha que a Organização Mundial da Saúde (OMS) (WHO, 2024) e o Instituto de Igualdade de Gênero Europeu (EIGE) (EIGE, 2024), o projeto também optou por focar neste grupo em um primeiro momento. Este não é o único ponto de alinhamento entre esta pesquisa e órgãos internacionais que visam aos direitos humanos. A compreensão de que a VBG é uma questão de saúde pública também é defendida pela OMS e corroborada por diversos outros estudos que julgam que a busca por soluções para a sua prevenção é uma prioridade (Garcia-Moreno; Watts, 2011; Öhman et al., 2020; Sweet, 2014).

No Brasil, profissionais da saúde desempenham papel estratégico na identificação e registro de casos de VBG, sendo obrigados por lei a notificar episódios suspeitos ou confirmados ao Sistema de Informação de Agravos de Notificação (SINAN). No entanto, mesmo com a obrigatoriedade legal, o número de subnotificações ainda é alto. Alguns estudos buscam entender as causas que levam a este cenário e evidenciam que o problema parte da falta de preparo para lidar com tais casos, desde o atendimento até a notificação. Entre as principais razões para a subnotificação se destacam o medo de um possível confronto ou represália e também entraves estruturais, tais como a ausência de protocolos, a falta de mecanismos legais de proteção para os profissionais responsáveis pela notificação, dificuldades na identificação de sinais de violência e o conflito com o sigilo profissional (Garbin et al., 2015; Kind et al., 2013).

Considerando este cenário, Vital Strategies Brasil e FrameNet Brasil propuseram o projeto DataToStopGBV, que visa ao desenvolvimento de recursos que possam levar ao alerta precoce de possíveis casos de violência contra mulheres que não foram expressamente reportados aos órgãos competentes. A proposta parte da **análise semântica de dados inseridos nos campos abertos de prontuários eletrônicos do SUS**, com o objetivo de **identificar padrões semânticos**, a partir da **metodologia da FrameNet**, para compreender como tais padrões de violência aparecem nos registros de saúde. Essa proposta vai na contramão da maioria das pesquisas neste campo, que focam em dados parametrizados (Malta et al., 2023; Mascarenhas et al., 2020; Pinto et al., 2021).

No Brasil, a inserção e o registro de informações sobre a população em contextos de saúde pública são realizados por meio de diferentes sistemas. Entre os principais sistemas voltados à coleta de dados sobre atendimentos em saúde e casos de violência estão o **e-SUS AB**, o **Sistema de Informações sobre Mortalidade (SIM)** e o **SINAN**.

O e-SUS AB é responsável por registrar os atendimentos realizados na atenção primária à saúde, a partir de informações do **prontuário eletrônico do paciente**. Já o SIM

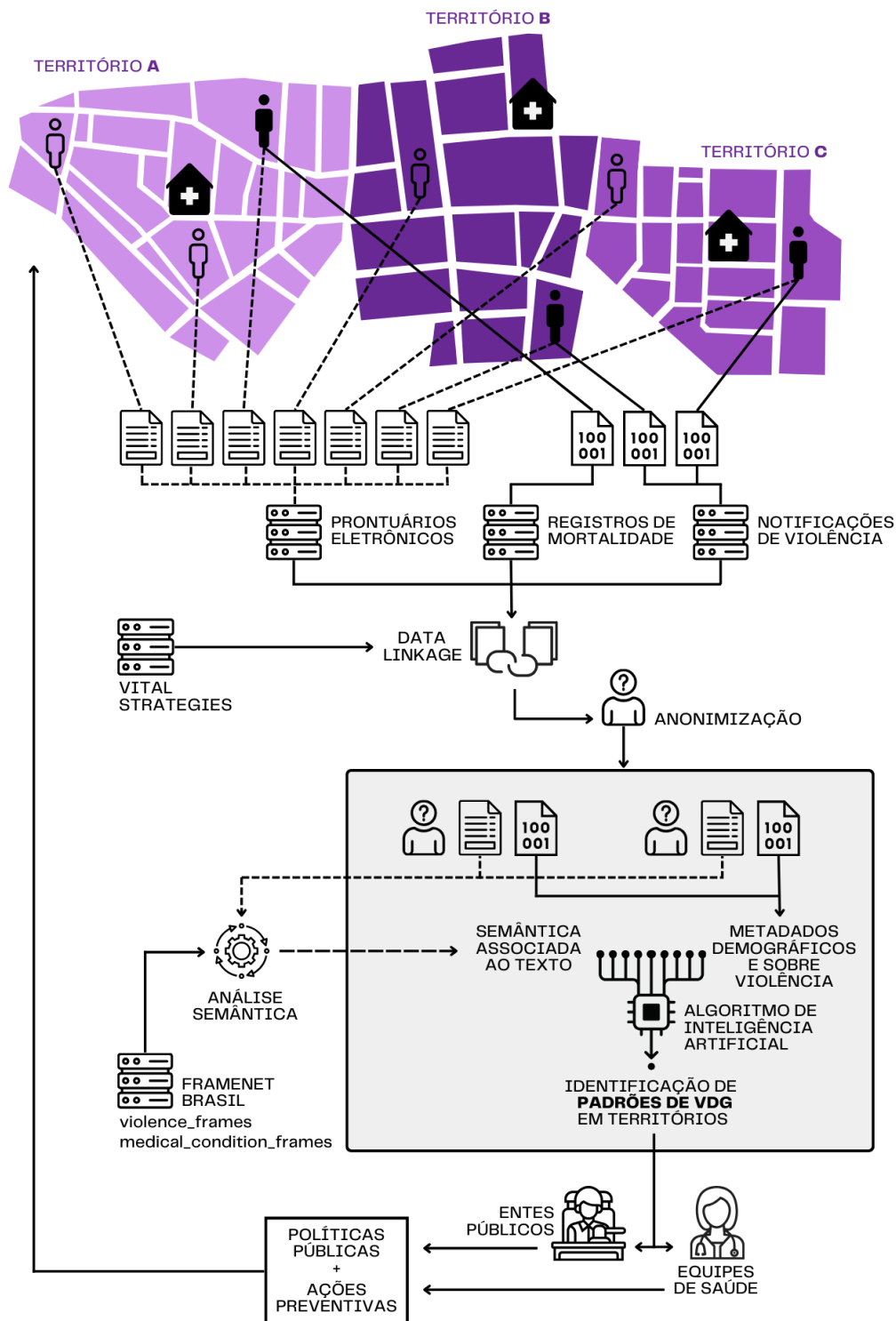


reúne **dados sobre óbitos** tendo como principal fonte a Declaração de Óbito. Grande parte das informações inseridas nesses sistemas segue um formato padronizado: os profissionais devem selecionar códigos a partir de listas pré-estabelecidas, como a **Classificação Internacional de Doenças** (CID-10), o que acaba por limitar o nível de detalhamento dos casos registrados. No entanto, tanto o e-SUS quanto o SINAN oferecem **campos de texto aberto**, que permitem uma descrição mais contextualizada das situações atendidas. Apesar de trabalhar com um sistema diferente, o Capítulo 9, por exemplo, também aborda o trabalho com campos de textos abertos que permitem uma descrição narrativa dos atendimentos aos pacientes. Esses são os campos relevantes para o desenvolvimento deste projeto.

Para atingir esse objetivo, o projeto firmou uma parceria com a **prefeitura de Recife**, que garantiu acesso aos registros do e-SUS, SINAN e SIM para estudo. Com isso, foi necessário o desenvolvimento de uma metodologia capaz de conectar tais sistemas, garantir a segurança dos dados usados e aplicar a metodologia da FrameNet para identificar padrões que pudessem ser usados para criar políticas públicas (veja Figura 17.5).



Figura 17.5: Pipeline do projeto DataToStopGBV



Fonte: (Dutra, 2024, p. 5.)



Desta forma, uma vez que os dados foram disponibilizados, a primeira etapa do processo consistiu no *linkage* dos bancos de dados da saúde pública de Recife – e-SUS, SINAN e SIM – o que permitiu relacionar os registros de uma mesma pessoa em diferentes unidades, além de mapear possíveis padrões de uso do sistema por vítimas de violência. Tais registros, no entanto, contêm dados extremamente sensíveis que são, por lei, protegidos. Portanto, existe uma preocupação ética relacionada ao acesso e armazenamento de tais tipos de dados, temas estes que também são discutidos nos Capítulos 9 e [Questões éticas em IA e PLN](#). Dessa forma, para que o acesso aos dados pudesse ser autorizado, todos os conjuntos utilizados foram previamente anonimizados.

A próxima etapa da metodologia descreve o foco da contribuição da FrameNet Brasil para o projeto que envolve o desenvolvimento do modelo semântico baseado na Semântica de *Frames* e FrameNet. Esta etapa inclui a **modelagem de domínios específicos da Saúde e da Violência** e a **anotação semântica manual e automática** dos dados que permitiu a análise dos registros em campos abertos e identificação de possíveis padrões semânticos ligados a casos de violência baseada em gênero.

A seguir, iremos destacar o processo linguístico-computacional aplicado neste projeto, o recorte de interesse para este capítulo, que descreve a modelagem dos domínios específicos da Saúde e da Violência, o processo de anotação aplicado, assim como os experimentos que avaliaram a contribuição da metodologia para a identificação de casos e de padrões de violência.

17.4.2 A contribuição da FrameNet para identificação da Violência Baseada em Gênero

A construção de uma rede semântica especializada se caracteriza pela **modelagem de um domínio específico**. No contexto da FrameNet, tal processo constitui a criação de uma representação cognitiva estruturada, que organiza conceitos centrais de um determinado tópico. O domínio, então, corresponde a um grupo de *frames* que cobrem eventos, participantes, objetos, entre outros, que são organizados e relacionados entre si via relações *frame-a-frame*. O processo de modelagem de um domínio envolve diferentes etapas que passam pelo estudo de um dado tópico, as ULs e *frames* relacionados a ele, as relações entre *frames* e entre ULs – as relações qualia – e, por fim, a anotação semântica que valida a modelagem.

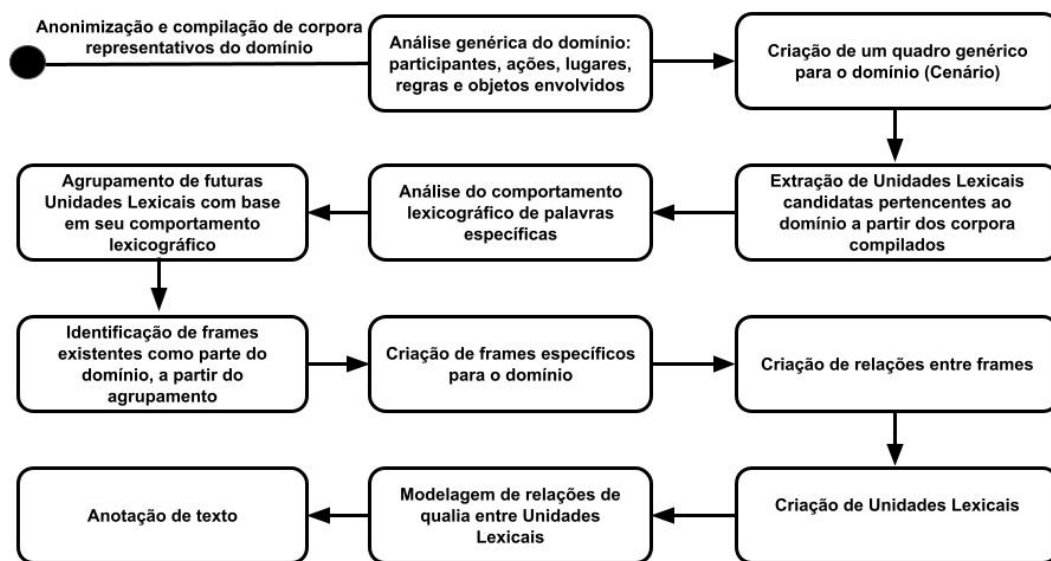
17.4.2.1 Modelagem dos domínios

Este processo foi definido inicialmente por Costa (2020), mas adaptado e ampliado para o escopo deste projeto como é descrito por Dutra (2024). A Figura 17.6 ilustra o passo a passo de 12 etapas da modelagem de domínio específico seguida neste projeto para ambos os domínios.

Como descrito na imagem, o primeiro passo na modelagem dos domínios foi a anonimização e compilação dos *corpora* representativos do domínio: os campos abertos dos registros disponibilizados por Recife do e-SUS e SINAN. Conforme já discutido, tais registros possuem informações extremamente sensíveis que poderiam levar à identificação dos pacientes, o que por si já levanta preocupações. No entanto, quando consideramos que tais pacientes podem ser potenciais vítimas de violência e que sua identificação poderia prejudicar sua segurança, esse fator aumenta o cuidado com que os dados precisam ser trabalhados. Portanto, para garantir a proteção das informações, todos os membros envolvidos assinaram acordos de confidencialidade. Além disso, os dados passaram por



Figura 17.6: Passo a passo da metodologia para modelagem computacional de um domínio específico



Fonte: (Dutra, 2024, p. 25).

um processo de anonimização em três etapas, e seu acesso foi restrito até mesmo entre os participantes, com apenas amostras anonimizadas sendo repassadas para os pesquisadores da FrameNet Brasil. Uma vez que as amostras anonimizadas que compuseram os *corpora* de análise foram disponibilizadas, um estudo dos domínios foi feito para que um panorama de conceitos relacionados a eles pudesse ser desenhado. Isso implicou a compreensão de quais são os possíveis participantes, ações, lugares, objetos que poderiam estar presentes no domínio. A partir deste estudo, foi modelado o primeiro *frame* de um domínio: o de cenário. No caso dos domínios da Saúde e Violência, *frames* existentes foram ajustados para ocupar essa posição. O *frame* *Cenário_de_interações_médicas* foi reestruturado para ser mais amplo e se tornou *Cenário_da_saúde*. Enquanto o *frame* *Violência* foi ajustado para *Cenário_da_Violência*.

Os próximos passos estão relacionados com lemas que são candidatos a ULs. Primeiramente, a ferramenta Keywords do *software* Sketch Engine (Kilgarriff; Al., 2014) foi usada para extrair termos mais relevantes ao domínio. Isso foi feito a partir de uma comparação dos *corpora* específicos com um *corpus* genérico, o Portuguese Web 2020. De acordo com a frequência de ocorrência de um termo, ele era considerado tendo maior ou menor probabilidade de estar relacionado ao domínio. Dessa forma, palavras como “dor de cabeça.n” ou “abuso.n” se destacavam como tendo maior probabilidade de pertencer aos domínios em comparação a palavras como “dia.n” ou “almoço.n”, por exemplo. A lista inicial de tais termos foi composta pelas 1.000 palavras e 1.000 multiwords¹³ (Capítulo *Expressões multipalavras*) com maior chance de pertencer ao domínio.

Uma vez que esta etapa foi concluída, os termos da lista gerada foram analisados quanto

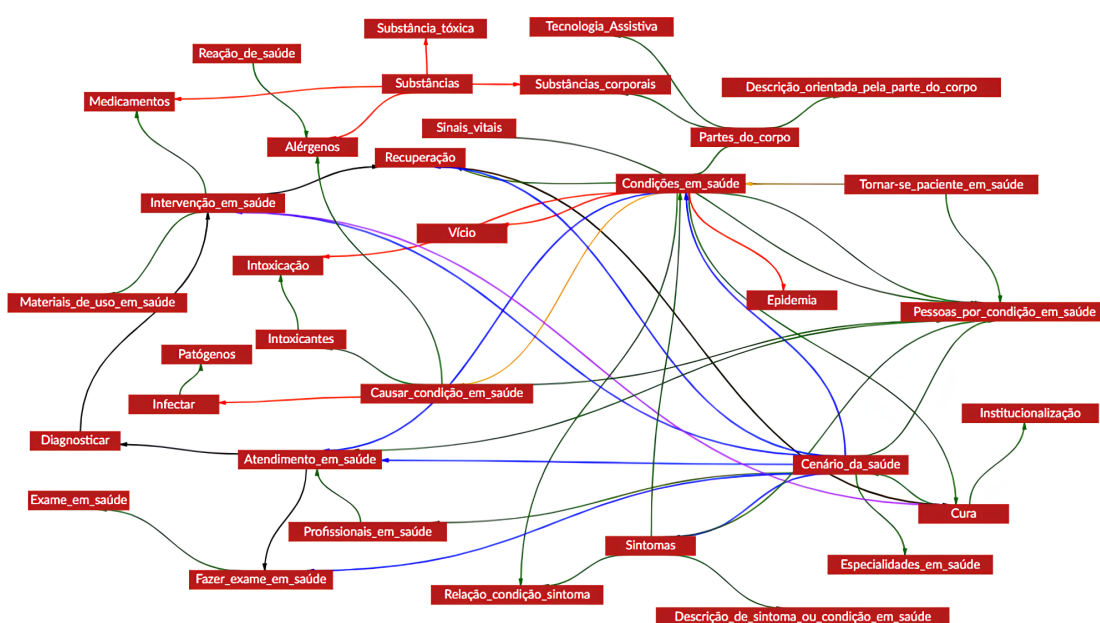
¹³Expressão composta por duas ou mais palavras que juntas possuem sentido próprio, como por exemplo, “pé de moleque”. Expressão que denomina um doce e não o pé de um menino.

ao seu comportamento lexicográfico e o agrupamento das futuras ULs foi feito com o objetivo de identificar ULs pertencentes a *frames* existentes ou novos *frames*. A modo de ilustração, houve uma grande ocorrência de ULs relacionadas ao *frame* **Parentesco** – “pai.n”, “mãe.n”, “filho.n” – no entanto, tal *frame* não faz parte diretamente de nenhum dos domínios em foco. Já ULs que nomeiam doenças – “gripe.n”, “enxaqueca.n”, “infecção.n” puderam ser relacionadas diretamente ao **Condições em saúde**.

Em um segundo contraponto, havia também um grande número de ULs que poderiam estar relacionadas a este *frame*, por se referirem a condições de saúde – tais como “dor de cabeça.n”, “febre.n” e “coriza.n”. No entanto, sua análise apontou a necessidade da modelagem de um novo *frame*, o de **Sintomas**. Esta etapa do processo permitiu uma melhor compreensão de quais *frames* existentes deveriam ser incluídos nos domínios e quais precisariam ser modelados para formar o domínio.

Com isso, os próximos passos foram a modelagem dos *frames* sugeridos e das relações entre *frames* que precisavam ser estabelecidas dentro da rede semântica da FrameNet Brasil. O domínio da Saúde possui um total de 35 *frames*, desses, 17 foram modelados a partir das demandas deste projeto. Dos que já estavam no banco, seis precisaram ser modificados. Já o domínio da Violência possui 48 *frames*, sendo um modificado e 5 novos. Com os *frames* de cada domínio definidos, foi possível conectá-los via relações *frame-a-frame*. As Figuras 17.7 e 17.8 ilustram as relações, sendo que cada cor, como explicado anteriormente, representa um tipo de relação – vermelho: herança, azul: subframe, verde: uso, amarelo: causativo de, marrom: incoativo de, preto: precedência, rosa: perspectiva de e roxo: ver também. As descrições da implicatura de cada uma das relações pode ser revista na Seção 17.3.1.

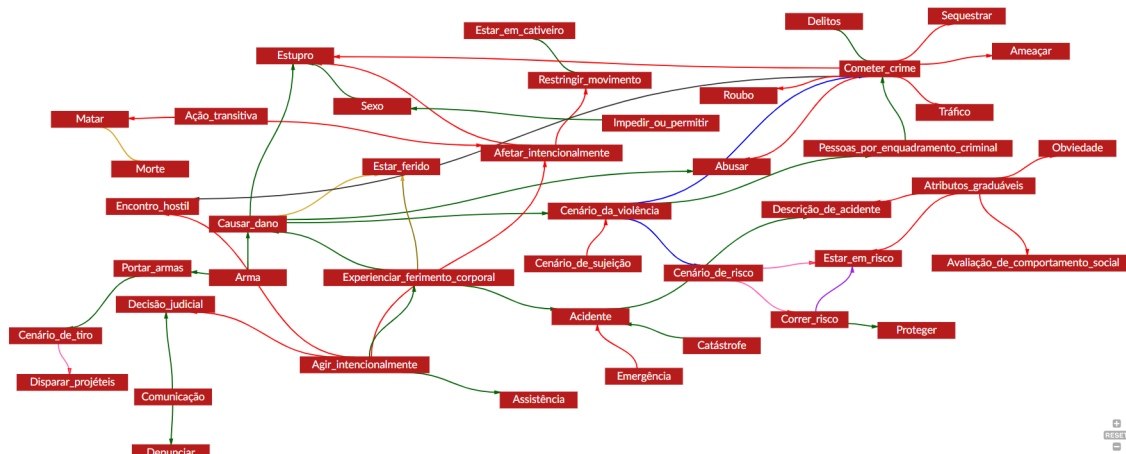
Figura 17.7: Representação do Domínio da Saúde



Fonte: Webtool 4.1

Os próximos passos estão ligados às ULs. Após a definição das relações entre os *frames*, os lemas agrupados como possíveis ULs foram associados aos *frames* na base de dados e às relações qualia foram definidas. Para que isso pudesse ser feito, no entanto, os lemas e

Figura 17.8: Representação do Domínio da Violência



Fonte: Webtool 4.1

wordforms – possíveis variações de cada lema tais como flexão e conjugação - precisavam existir na Webtool para que as estruturas fossem reconhecidas na próxima fase: a anotação.

Para este projeto, no entanto, além das *wordforms* tradicionais foram acrescentadas grafias alternativas e abreviações. Isso foi feito devido ao caráter dos *corpora* que possuem jargões específicos do âmbito da saúde e foram “construídos” durante consultas em unidades de saúde brasileiras que geralmente estão sob alta demanda, o que pode levar a erros ortográficos e de digitação. Ignorar tais ocorrências levaria à perda significativa de dados que não seriam reconhecidos.

Com todos os lemas e *wordforms* adicionados, as ULs foram criadas e as relações qualia definidas. Como introduzido na Seção 17.3.2, as relações qualia relacionam ULs o que permite a expansão de suas informações semânticas. Existem quatro tipos de relações qualia: agentiva, formal, constitutiva e télica que na FrameNet Brasil são intermediadas por *frames* como ilustrado no Quadro 17.1. Até este momento, foram criadas 4.698 relações entre ULs pertencentes aos domínios em foco.

DESAFIO: *Frame* de cenário para domínios específicos

Com base na descrição acerca do processo de modelagem de um domínio, propomos o seguinte desafio: **a proposta de um domínio específico e a modelagem de um frame de cenário para o domínio proposto.** No momento, a base de dados da FrameNet Brasil possui por volta de 1.200 frames e cinco domínios específicos: turismo, esporte, saúde, violência e educação.

Para a modelagem do frame de cenário, o domínio precisa ser estudado para que seja possível definir o frame e os seus elementos de frame – nucleares e não-nucleares. Na Webtool 4.1 é possível acessar outros frames de cenário, para melhor entendimento se necessário. A modelagem do frame deve ser feita pela plataforma Lutma. A plataforma é intuitiva e nela também estão disponíveis vídeos tutoriais.

Dúvidas podem ser enviadas para o email dos autores deste capítulo.



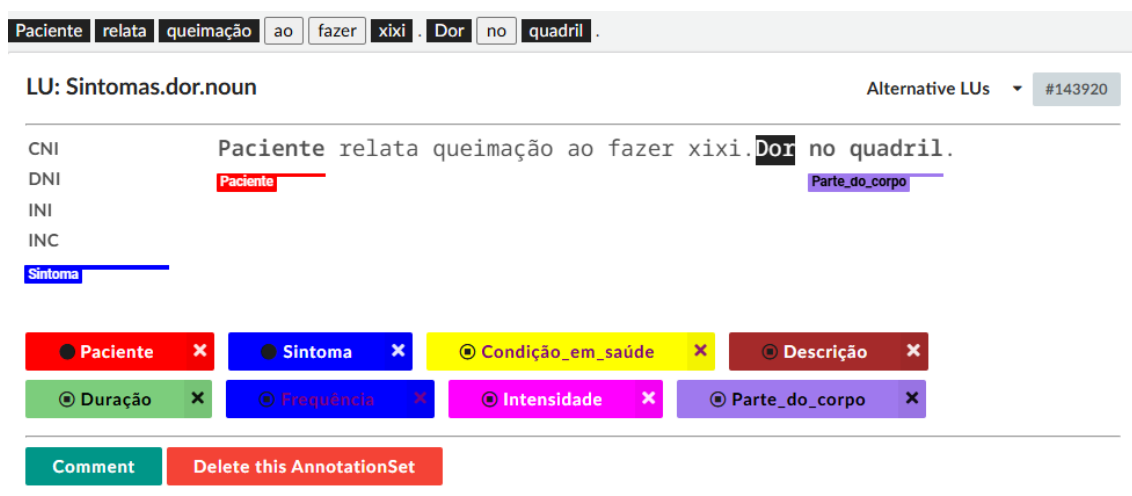
17.4.2.2 Anotação Computacional

A **anotação computacional** neste projeto foi dividida em duas etapas. A primeira delas foi a anotação manual feita na plataforma Webtool. Dada a natureza dos dados, uma versão da plataforma foi criada com acesso restrito aos integrantes do projeto. O primeiro objetivo da anotação manual foi validar a modelagem do domínio. A anotação computacional é o último passo do processo e permite uma análise geral da efetividade do domínio, tais como a suficiência dos *frames* – em cobertura e em estrutura.

A metodologia de anotação utilizada foi baseada na definida por Ruppenhofer et al. (2016) descrita na Seção 17.3. O objetivo do projeto e as características dos *corpora* levaram a pequenos ajustes, tais como (1) o foco apenas na camada semântica; (2) a ausência da anotação de Instanciações Nulas (NI); (3) liberdades quanto à localidade sintática; e (4) a anotação por campo e não por sentença.

Na Figura 17.9, observa-se a anotação da UL “*dor.n*” no registro “*Paciente relata queimação ao fazer xixi. Dor no quadril.*”¹⁴ que exemplifica as alterações nas diretrizes de anotação. Neste exemplo, nota-se que a anotação contém apenas informações semânticas. Outro ajuste feito diz respeito à presença de dois períodos no campo de anotação, quando normalmente “*Paciente relata queimação ao fazer xixi*” e “*Dor no quadril*” estariam em campos de anotação distintos. Observa-se também que a marcação do elemento de *frame* PACIENTE que poderia ter sido anotado como uma Instanciação Nula Definida - pelo contexto do prontuário - ou uma Instanciação Nula Indefinida - por não se saber quem é este paciente, foi marcada na sentença no período anterior. No entanto, como foi optado por não anotar Instanciações Nulas, a anotação do elemento de frame foi feita fora da localidade sintática - neste caso, em outro período.

Figura 17.9: Exemplo de anotação



Fonte: Adaptada de (Dutra, 2024)

Ao todo foram anotadas 2.352 sentenças, sendo que foram criados 6.313 *Annotation Sets* no domínio da Saúde e 8.309 no domínio da Violência. Este processo foi realizado por sete anotadores com diferentes níveis de experiência. Reuniões semanais também foram

¹⁴Devido ao conteúdo sensível dos dados utilizados nesta pesquisa, sentenças reais não puderam ser utilizadas. Desta forma, o exemplo usado foi criado para ilustrar o tipo de ocorrência presente no *corpus*.

realizadas para discutir o processo de anotação e esclarecer possíveis dúvidas.

A segunda etapa de anotação consistiu na anotação automática dos dados. A partir da amostra obtida pela anotação manual, uma versão do **rotulador automático LOME** (Xia et al., 2021) foi treinada do zero – com cerca de 200 *frames* a mais do que os usados no artigo citado – e usada para a anotação em larga escala dos *corpora*. O LOME é um modelo desenvolvido para extração de informações multilíngues com um *pipeline* que inclui um **parser semântico da FrameNet**. O modelo foi usado para anotar os registros do e-SUS AB, SINAN e CID. Ele é capaz de identificar as ULs alvos, rotular o *frame* e seus elementos. A avaliação da anotação utilizou as mesmas métricas propostas por Xia et al. (2021): *exact match* (em) e *span match* (sm), todas considerando *precision*, *recall* e *F1*. A métrica **em** exige correspondência exata de todo o trecho da anotação. Já a **sm** avalia apenas a detecção correta dos limites do trecho, sem considerar rótulos. Os resultados foram similares aos alcançados por Xia et al. (2021), o que aponta uma boa performance do modelo, uma vez que foram usadas informações de domínios mais específicos, *frames* novos e mais de 5.000 ULs que haviam sido anotadas pela primeira vez.

17.4.3 Avaliação da identificação da Violência Baseada em Gênero

Para avaliar os benefícios do uso da Framenet para a identificação de casos de VBG, três experimentos foram feitos. Os experimentos contrastavam o uso da metodologia da FrameNet classificando casos como sendo ou não casos de VBG em uma configuração que se apoiava nos dados dos campos de texto abertos e fechados.

Os registros que foram utilizados para a classificação se enquadram em quatro grupos: (1) violência – casos com código CID para violência ou notificação SINAN; (2) não violência – sem evidências de violência, muitas vezes associados a doenças; (3) provável violência – casos com notificação de violência sem código CID para agressão 30 dias antes da mesma; e (4) desconhecidos – não se relaciona a nenhuma das categorias. Foram considerados 801 casos, sendo 167 deles classificados como violência.

A partir dos experimentos, foram feitas as avaliações quantitativas e qualitativas. A avaliação quantitativa focou na comparação das performances dos experimentos. Já a avaliação qualitativa observou os atributos considerados mais relevantes para o modelo e foi complementada com a análise dos *frames* e unidades lexicais com maior ocorrência na anotação para ambos os domínios.

A avaliação do processo de modelagem e anotação manual dos domínios também foi avaliada no âmbito de Dutra (2024). No entanto, para este recorte, o foco será nos resultados diretamente ligados à avaliação da contribuição da FrameNet para a identificação de casos de VBG. Os experimentos e os processos avaliativos aplicados serão discutidos a seguir.

17.4.3.1 Os experimentos

Os três experimentos usaram o modelo **Support Vector Machine** (SVM) (Cortes; Vapnik, 1995), um modelo frequentemente usado em tarefas de classificação. A escolha foi motivada devido a sua capacidade de lidar com dados com muitas variáveis e flexível para separar padrões tanto em casos lineares quanto não lineares, ou seja, em que não há uma divisão simples entre as classes. Uma das vantagens do SVM é a possibilidade de ranquear a importância dos atributos na escolha da classificação, o que foi feito com o objetivo de apoiar a análise qualitativa e reforçar a relevância da informação semântica da FrameNet.

O primeiro e o segundo experimentos utilizaram o mesmo pipeline de seis etapas: (1) a anotação com LOME; (2) a identificação das LUs, (3) a criação dos atributos; (4) a



filtragem dos vetores com base em sua ocorrência; (5) a geração da representação TF-IDF¹⁵ dos vetores; e (6) a aplicação da análise do componente principal (PCA)¹⁶ às representações TF-IDF. A diferença entre os dois experimentos está nos dados que foram usados. O primeiro experimento considerou apenas os campos abertos de texto anotados, enquanto o segundo, além dos campos abertos anotados, considerou também um grupo de parâmetros do campo fechado anotado para *frames*: orientação sexual e gênero, necessidade de prótese, raça e idade.

Por fim, o terceiro experimento consistiu apenas na criação de atributos a partir das informações dos campos parametrizados, ou seja, os dados considerados não passaram pelo processo de anotação. Consideramos vinte dos campos disponíveis, dentre eles características do paciente, bloco e número CID, localidade e variáveis temporais acerca dos atendimentos.

17.4.3.2 Avaliação Quantitativa

A avaliação quantitativa dos modelos foi feita por meio de validação cruzada. Esta é uma técnica de avaliação que permite o treinamento e teste do modelo com diferentes combinações dos conjuntos de dados utilizados (Kohavi, 1995). As métricas consideradas foram *F1 score*, *recall* e *precision*. O resultado final é uma média calculada a partir das diferentes combinações feitas. Para este projeto, o conjunto de dados foi dividido em cinco conjuntos, portanto foram feitas cinco interações e a média foi calculada a partir do resultado delas. O resultado pode ser visto na Tabela 17.2, que apresenta o desempenho final da validação cruzada, seguido da variação, que indica a dispersão dos resultados entre as diferentes rodadas do processo.

Tabela 17.2: Resultados das performances dos experimentos

Experimento	<i>F1</i>	<i>Recall</i>	<i>Precision</i>
1	0.772 (0.113)	0.838 (0.071)	0.756 (0.190)
2	0.771 (0.114)	0.832 (0.071)	0.759 (0.189)
3	0.461 (0.089)	0.701 (0.173)	0.345 (0.057)

Fonte: (Dutra, 2024)

Os resultados obtidos apontam que, quando a anotação semântica seguindo as diretrizes da FrameNet Brasil foi considerada, a performance do modelo obteve um desempenho melhor na identificação de casos de VBG. O primeiro experimento apresentou o melhor desempenho, com alta *recall* e equilíbrio entre *precision* e *F1*, o que é ideal no contexto da pesquisa, em que falsos positivos são preferíveis a falsos negativos. O segundo modelo, que inclui dados parametrizados, teve desempenho semelhante, que confirma a vantagem do uso da metodologia. Já o terceiro experimento teve resultados inferiores, com alta taxa de falsos positivos e grande variação nos testes. Esses achados reforçam a hipótese de que informações semânticas extraídas de texto livre acrescida de informações do modelo semântico da FrameNet são mais eficazes na identificação de casos de VBG.

¹⁵Técnica que calcula a importância de uma palavra em um texto, levando em conta sua frequência no documento e quão rara ela é em todo o conjunto de textos.

¹⁶Método que reduz o número de variáveis nos dados, transformando-as em componentes principais e mantendo a maior parte da informação original.



17.4.3.3 Avaliação Qualitativa

Para a avaliação qualitativa, foram analisados o primeiro e o terceiro experimento. O foco foi nos atributos que mais influenciaram a classificação dos casos como violência ou não violência. No entanto, é importante ressaltar que neste momento não foi possível determinar para qual classe cada atributo foi mais relevante, apenas que ela teve peso na decisão de classificação do modelo. De modo complementar à análise semântica do primeiro experimento, uma lista dos *frames* e ULs mais evocados dos domínios da Saúde e da Violência de acordo com a anotação automática do LOME também foi analisada em busca de mais padrões. Essa lista considerou apenas a anotação de casos confirmados de VBG.

A análise dos atributos mais relevantes no terceiro experimento revelou uma concentração em dados de localização das unidades de saúde, indicando um viés, já que esse fator não tem correlação direta com a ocorrência de violência; pelo contrário, alimenta estereótipos. O atributo *raça* também teve destaque, o que pode refletir um viés decorrente da inconsistência no preenchimento do campo, muitas vezes deixado em branco. Esse resultado sugere que o modelo tende a considerar mais atributos pessoais, e não o contexto do atendimento.

Por outro lado, alguns códigos CID relacionados ao motivo da busca por atendimento também foram identificados, com destaque para códigos que podem estar associados a casos de violência sexual ou situações que exigem acompanhamento, como o pré-natal. Apesar disso, tais variáveis foram menos frequentes. Isso reforça novamente que o uso dos campos parametrizados não traz informações que podem ser cruciais sobre o atendimento, e também tendem a reforçar estereótipos ao focar no indivíduo e não na causa para a busca pelo serviço de saúde. Esses dados podem ser vistos na Figura 17.10.

Já a análise do primeiro experimento, apesar de preliminar, apresentou melhor resultado quanto à identificação de possíveis padrões de violência. Os atributos de maior relevância estão listados na Figura 17.11. Dentre eles, destacam-se *frames* genéricos como *Relações_pessoais* e *Medo*, que, embora fora do domínio de violência, podem indicar relações com o agressor ou o cenário emocional envolvido. Entre os *frames* do domínio da violência, apenas *Experienciar_ferimento_corporal* foi identificado, o que pode estar relacionado à baixa frequência desses *frames* nos registros do e-SUS. Ainda assim, sua presença sugere que sinais de violência podem surgir indiretamente por meio de dados de saúde. No domínio da saúde, destacaram-se *Condições_em_saúde*, *Medicamentos* (com múltiplos elementos de *frame* relevantes) e *Intervenção_em_saúde*, com ULs como *prescrever.v* e *administrar.v*. Tais resultados reiteram que os registros médicos podem fornecer pistas importantes para a identificação precoce da violência baseada em gênero.

A análise da lista extraída da anotação do LOME dos casos de violência trouxe mais possíveis padrões que merecem atenção. No domínio da saúde, dois grupos de ULs de *Condições_em_saúde* se destacaram. O primeiro relacionado à gravidez. Este fato, no momento, leva a duas hipóteses: é possível que este resultado se dê devido a um enviesamento do modelo, uma vez que o grupo de estudo é composto por mulheres. Uma alternativa, no entanto, é de que a gravidez aumenta a vulnerabilidade da mulher, fazendo com que ela seja mais propensa a permanecer em um relacionamento abusivo. O segundo grupo que chama atenção é o de ULs relacionadas a transtornos mentais. Novamente, existe a possibilidade de enviesamento do modelo, já que mulheres podem buscar ajuda profissional com maior frequência nestes casos do que homens. No entanto, tal estado pode também ser uma consequência da violência experienciada.



Figura 17.10: Resultado dos atributos mais relevantes no terceiro experimento de usando apenas dados parametrizados



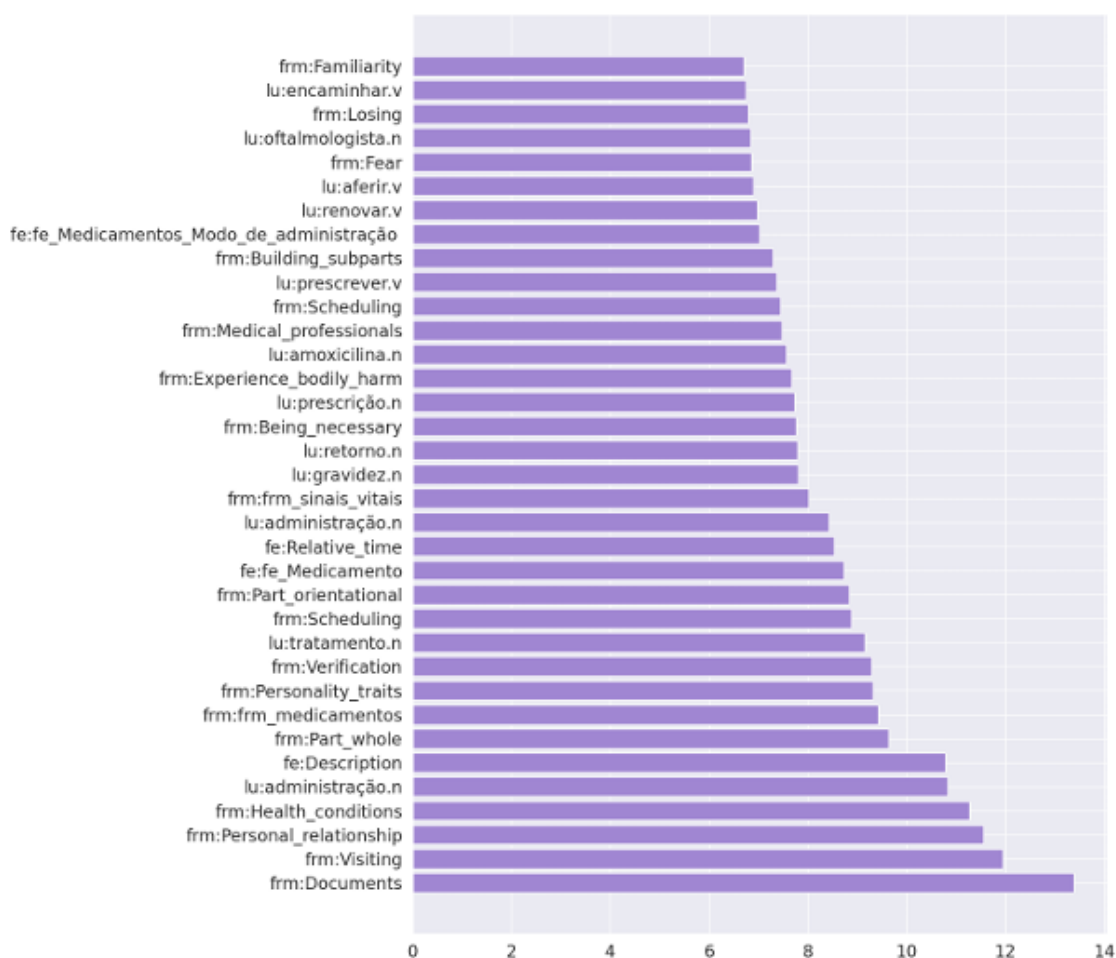
Fonte: (Dutra, 2024)

No domínio da Violência, destacam-se também dois grupos de ULs. O primeiro relacionado à automutilação e ao suicídio. Tais dados foram encontrados com alta frequência e demandam um estudo isolado, mas já exibem forte indício de serem uma consequência da violência. O segundo grupo está relacionado a abuso sexual, resultado esse que foi reforçado nos dados do domínio da Saúde devido à alta incidência de exames e doenças ligadas a doenças sexualmente transmissíveis. Esses resultados, tanto da avaliação semântica do modelo, quanto dos resultados obtidos a partir da anotação LOME, reforçam mais uma vez o valor que a análise de tais dados via FrameNet pode trazer para a compreensão de padrões de violência nos registros de saúde e identificação de casos de VBG.

17.4.4 Impacto Social e técnico

A partir da descrição da metodologia aplicada a este projeto e dos resultados discutidos, são visíveis os benefícios do uso da metodologia da FrameNet para a interpretação semântica de registros nos sistemas de saúde e a identificação de padrões de violência. Considerando

Figura 17.11: Resultado dos atributos mais relevantes no primeiro experimento usando a anotação semântica do campo aberto



Fonte: (Dutra, 2024)

que a maior parte das pesquisas que buscam a prevenção de casos de Violência Baseada em Gênero focam em dados estatísticos, o projeto DataToStopGBV é o pioneiro em avançar este campo de pesquisa para a análise dos campos abertos com resultados que comprovam a eficácia do uso da FrameNet no estudo de casos notificados e, também, na identificação de casos subnotificados.

No entanto, como pode ser observado nos resultados, para além de padrões relacionados à violência física e psicológica, foram identificados também padrões relacionados à saúde da mulher vítima de violência. Tais dados levam para dois possíveis desdobramentos futuros: a busca da mulher por atendimentos relacionados a sua saúde que mascaram a violência sofrida e as condições de saúde que se sobressaem em mulheres que são vítimas de violência. O estudo e a análise destes dados em conjunto com profissionais da saúde podem levar a desdobramentos com impactos significativos para a prevenção da VBG.

Para além deste âmbito, dada a eficácia da metodologia aplicada, existe também a possibilidade da ampliação deste projeto para outras esferas que extrapolem a VBG. Esse

movimento pode partir da expansão do grupo foco, fazendo com que não apenas os casos de VBG contra mulheres sejam analisados, e pode alcançar outros tipos de violência tais como a violência infantil.

Na perspectiva técnica, este projeto também mostra a capacidade de adaptação da metodologia da FrameNet, não apenas para a criação de domínios semânticos específicos – trabalhos já desenvolvidos anteriormente (Costa, 2020; Gamonal, 2013) –, mas também para objetivos específicos. Observou-se, neste projeto, a adaptação da base de dados em duas esferas: a criação de uma versão espelhada com acesso restrito e a adesão de lemas e *wordforms* que incluíam abreviações e grafias alternativas. Para além, houve a adaptação das diretrizes de anotação manual, que foram ajustadas e ainda assim alcançaram performance significativa quando reproduzidas automaticamente.

A metodologia deste projeto é adaptável não apenas com relação aos métodos da FrameNet, mas ao modo de expansão territorial. Portanto, após possíveis ajustes tanto no âmbito do léxico regional ou de uma possível diferença no sistema público de saúde usado, o trabalho desenvolvido com os dados da cidade de Recife pode ser ampliado para outras cidades e/ ou estados brasileiros.

Dessa forma, o trabalho conjunto entre essa iniciativa, profissionais da saúde e órgãos governamentais para a elaboração de políticas públicas que possam ser implementadas é imperativo. Com a expansão desta pesquisa e possível aplicação em maior escala, esperamos o fortalecimento das políticas públicas para intervenção preventiva de casos de Violência Baseada em Gênero e o decréscimo de casos de subnotificação.

17.5 Multimodalidade, acessibilidade cultural e sentido compartilhado no Audition: *dataset* multimodal de curtas-metragens acessíveis

A **comunicação humana** em contextos reais de uso é fundamentalmente **multimodal**. A noção de multimodalidade, tal como entendida neste capítulo, refere-se à participação simultânea e coordenada de diferentes **modos comunicativos** (verbal, visual, sonoro, corporal entre outros) na construção de sentido. Assumimos uma concepção fenomenológica do termo, como discutido por Torrent; Coneglian (2025), por estarmos interessados nos meios (multi)semióticos pelos quais os indivíduos constroem significado na experiência situada da linguagem.

Essa concepção ultrapassa a ideia de uma simples junção de semioses, pois envolve uma estrutura compartilhada de significados, na qual os elementos provenientes dos diferentes modos se articulam de forma integrada. Em outras palavras, a multimodalidade, tal como concebida em nossas pesquisas, diz respeito à articulação entre modos em torno de um sentido coerente, ativado e inferido no uso situado da linguagem.

É a partir dessa perspectiva que passamos a relacionar a Semântica de *Frames* e a FrameNet Brasil a um contexto multimodal. Partimos do pressuposto de que, assim como o texto verbal evoca *frames*, imagens, sons e gestos também o fazem. Essa premissa abriu caminho para a criação da ReINVenTA (*Research and Innovation Network for Vision and Text Analysis*), uma rede de pesquisa dedicada ao processamento semântico computacional de objetos multimodais.

A atuação da ReINVenTA concentra-se em quatro eixos principais: (i) a ampliação da cobertura da FrameNet para o português brasileiro, incorporando dados multimodais; (ii) a constituição de um *gold standard dataset* de objetos multimodais anotados semanticamente



e validados em perspectiva psicolinguística; (iii) o desenvolvimento de algoritmos e modelos para rotulação automática e descoberta de conhecimento em objetos multimodais; e (iv) a proposição de melhores práticas para tradução audiovisual acessível, com ênfase na audiodescrição fílmica.

Esses objetivos se materializam em três *datasets* distintos que compõem, até o momento, o conjunto de dados da rede.

O primeiro é o Frame² (Belcavello et al., 2024), composto por episódios do programa *Pedro pelo Mundo*. Esse *dataset* foi anotado para *frames*, elementos de *frames* e categorias de objetos reconhecíveis por algoritmos de visão computacional para as modalidades de vídeo, áudio original e legendas.

O segundo é o Framed Multi30k (Viridiano et al., 2024), que expande o *dataset* Multi30k com descrições em português e anotações semânticas visuais. Neste *dataset*, foram produzidas cinco descrições em português e uma descrição traduzida do inglês para cada uma das cerca de 30 mil imagens do *dataset* Flickr 30k. Além disso, o *dataset* inclui anotações automáticas de *frames* para todas as legendas em inglês e português, bem como anotação manual de *frames*, elementos de *frame* e *bounding boxes* aplicadas às imagens do *dataset* Flickr 30k Entities.

O terceiro é o Audition, *dataset* de curtas-metragens brasileiros acessíveis, que será apresentado a seguir.

17.5.1 O *dataset* Audition

O *Audition* é um *dataset* multimodal em desenvolvimento, composto por curtas-metragens brasileiros anotados semanticamente com base na teoria da Semântica de *Frames* e no modelo da FrameNet Brasil. Sua criação foi motivada pela necessidade de recursos que integrem os modos sonoro (verbal e não verbal) e visual em português brasileiro com ênfase na promoção do acesso a conteúdos audiovisuais, especialmente para pessoas com deficiência visual.

Embora os modelos atuais de PLN venham avançando significativamente no tratamento de textos escritos, ainda há pouca atenção às necessidades de usuários que dependem de traduções intersemióticas – como a audiodescrição – para ter acesso a obras audiovisuais, por exemplo.

O Capítulo **Conjunto de dados, *dataset* e *corpus*** oferece um excelente panorama sobre a preparação e a avaliação de dados linguísticos para PLN, discutindo, entre outros pontos, a distinção entre *corpus* e *dataset*. Quem já leu o capítulo deve ter alguma opinião formada sobre o conjunto de dados multimodais que configura o *Audition*. A partir dessa discussão, propomos uma reflexão: o *Audition*, enquanto conjunto de dados multimodais, deve ser classificado como *dataset* ou também pode ser considerado um *corpus*?

De acordo com o Capítulo **Conjunto de dados, *dataset* e *corpus***, um *dataset* é um conjunto de dados linguísticos anotados por pessoas ou feito por máquinas e revisto por pessoas. Já um *corpus* linguístico pode ter sido anotado automaticamente. A depender dos objetivos da pesquisa, receber um material já categorizado previamente pode facilitar a tarefa de estudo desejada. Essa questão diz respeito à utilidade do material em discussão.

Outro critério importante é a origem dos dados. Para configurar-se como *corpus*, o conjunto de textos deve ser composto por material linguístico naturalmente produzido, ou seja, não elaborado especificamente para fins experimentais ou científicos. Em contrapartida, *datasets* podem ser criados com a finalidade explícita de alimentar modelos, validar hipóteses ou compor bancos de dados anotados, como é o caso do *dataset* Multi30k



(Viridiano et al., 2024).

Em se tratando de dados multimodais, como os que compõem o *Audition*, as distinções conceituais entre *corpus* e *dataset* devem ser compreendidas de maneira ampliada. Isso porque os dados anotados não se restringem à linguagem verbal, mas envolvem também o modo visual e sonoro, organizado em alinhamento temporal com base nas unidades significativas do vídeo.

Nesse cenário, a noção de dado linguístico é estendida para abranger modos semióticos diversos que participam conjuntamente da construção de sentido. O alinhamento semântico entre esses modos – textual, visual e sonoro – permite explorar relações intermodais e avaliar traduções intersemióticas na construção de sentido da obra fílmica.

Tendo em vista esse enquadramento conceitual, o *Audition* pode ser caracterizado como um *dataset* multimodal padrão-ouro, cuja cobertura inclui anotação manual de transcrições derivadas de textos verbais (orais e escritos) e de conteúdo visual dinâmico (vídeo). Além do áudio original, ele contempla diversos modos de tradução intersemiótica, tais como audiodescrição, legendas, *closed captions* e texto sobreposto à imagem (*on-screen text*), todos anotados semanticamente por meio de *frames*.

O Quadro 17.3 apresenta a relação entre as modalidades contempladas, os tipos de dados e as anotações semânticas realizadas em cada caso.

Quadro 17.3: Cobertura Multimodal do *Dataset* Audition

Modalidade	Tipo de dado	Anotação semântica
Áudio original (OA)	Texto verbal (Fala, narração)	<i>Frames</i> , LUs, FEs
Conteúdo de Vídeo (VC)	Imagens em movimento (vídeo)	<i>Frames</i> , FEs, <i>Visual Entity</i> (via <i>bounding boxes</i>)
Audiodescrição (AD)	Texto verbal (roteiro escrito)	<i>Frames</i> , LUs, FEs
<i>Closed Captions</i> (CC)	Texto verbal escrito + som não verbal	<i>Frames</i> , LUs, FEs
Legendas	Texto verbal (tradução escrita das falas)	<i>Frames</i> , LUs, FEs
<i>On-screen Text</i>	Texto verbal (de origem gráfica sobreposto à imagem)	<i>Frames</i> , LUs, FEs

Todos esses modos, que atuam como recursos importantíssimos de acessibilidade cultural, estão alinhados semântica e temporalmente, permitindo a observação da sincronidade e da evocação de *frames* na construção de sentido. Esse alinhamento, detalhado na Seção 17.3, oferece subsídios para experimentos de similaridade semântica entre modos, contribuindo para avaliação da qualidade de traduções intersemióticas. Assim, o *Audition* é tanto um recurso para o desenvolvimento de sistemas de tradução audiovisual acessível, como também uma ferramenta para o estudo da construção e compartilhamento de sentido entre diferentes modalidades comunicativas.

Para garantir um bom *dataset* linguístico, o Capítulo Conjunto de dados, *dataset* e *corpus* destaca cinco características centrais. A seguir, cada uma dessas dimensões será discutida a partir das especificidades do *Audition*.

(i) consistência.

“Por consistência, entende-se que fenômenos semelhantes devem ser anotados,



isto é, analisados da mesma maneira. A consistência permite que algoritmos de aprendizado de máquina generalizem a partir dos dados e que as avaliações sejam confiáveis”. (Seção [Características de um bom dataset linguístico](#))

A anotação semântica no *Audition* foi realizada com base na metodologia de anotação *full text* da FrameNet e nas *guidelines* desenvolvidas pela FrameNet Brasil para anotação semântica multimodal. Essas diretrizes asseguram a rotulação de *frames*, unidades lexicais (LUs), entidades visuais (*visual entities*) e elementos de *frame* (FEs) a todas as modalidades comunicativas do *dataset*.

A equipe responsável pela anotação é composta por estudantes de graduação com bolsas de Iniciação Científica, previamente treinados em anotação semântica baseada em *frames*. Todo o trabalho é supervisionado por anotadores experientes vinculados à FrameNet Brasil.

Adotamos uma abordagem de anotação como definida pelo manifesto de dados perspectivista¹⁷ e por Basile et al. (2021). Desse modo, reconhecemos que a seleção de frames pode variar legitimamente entre anotadores em função de fatores como conhecimento de mundo, contexto situacional e a própria perspectiva interpretativa assumida.

Para diminuir inconsistências e assegurar qualidade da anotação, o processo inclui: **tutoriais iniciais de formação**, **encontros semanais** com espaço para discussão coletiva de decisões, **resolução de dúvidas** e **refinamento colaborativo das anotações**. Esse modelo de curadoria ativa e engajamento interpretativo é o que proporciona a consistência no *Audition*.

(ii) variedade:

“Um bom dataset deve ser variado (ou balanceado) com relação aos fenômenos para os quais foi construído” (Seção [Características de um bom dataset linguístico](#))

O *Audition* contempla filmes curtos de diferentes gêneros – ficção, animação, documentário e performance – com diversidade estilística e discursiva. Há ainda **variação** entre os modos comunicativos incluídos, o que amplia o escopo de análise de fenômenos intersemióticos. A diversidade de contextos e gêneros assegura que o *dataset* seja útil para tarefas no PLN e na tradução audiovisual acessível. O quantitativo por gênero cinematográfico e modos comunicativos está exibido na Figura 17.12.

(iii) representatividade:

“Esperamos que os textos que compõem o corpus/dataset sejam representativos do tipo de texto que será alvo da aplicação, isto é, ao qual o modelo baseado nesse dataset será aplicado” (Seção [Características de um bom dataset linguístico](#))

Como conjunto de dados voltado ao estudo e treinamento de tarefas computacionais em tradução audiovisual acessível em português brasileiro, o *Audition* foi construído a partir de obras cinematográficas acessíveis. Embora haja vários recursos disponíveis, a opção por audiodescrição foi a prioridade e está presente em todos os curtas que compõem o *dataset*.

(iv) documentação:

“Um bom dataset é um dataset bem documentado, que informa a origem do conteúdo textual, as classes de anotação e as diretrizes usadas para anotar, as

¹⁷<https://pdai.info/>



características e/ou formação dos anotadores”. (Seção *Características de um bom dataset linguístico*)

A documentação do *Audition* inclui quatro etapas principais: (a) seleção e liberação das obras cinematográficas com audiodescrição; (b) formações teóricas e práticas voltadas à anotação estrutural e à anotação semântica multimodal; (c) realização da anotação estrutural, que incluiu a transcrição e o alinhamento temporal entre os modos transcritos e o conteúdo visual dinâmico; e (d) anotação semântica multimodal, envolvendo a rotulação de texto e de vídeo em termos de *frames*, elementos de *frames*, unidades lexicais e entidades visuais.

Complementarmente, a documentação inclui metadados sobre os anotadores, incluindo formação acadêmica e grau de experiência com a anotação. Esses registros são importantes para proporcionar a transparência do processo, facilitar a replicação de experimentos, além de possibilitar comparações entre modalidades comunicativas e avaliação crítica do material produzido.

(v) tamanho:

“o tamanho é um aspecto que precisa ser levado em conta. E por tamanho não me refiro apenas ao tamanho do corpus em tokens ou palavras, mas também à quantidade de fenômenos rotulados”. (Seção *Características de um bom dataset linguístico*)

Como bem pondera (Capítulo *Conjunto de dados, dataset e corpus*), o critério do tamanho de um *dataset* diz respeito tanto à quantidade total de dados quanto ao número efetivo de itens anotados. Nesse sentido, ele deve ser suficiente para permitir treinamento e avaliação confiáveis, conforme o tipo de anotação realizada. Discussões sobre o tamanho não são conclusivas, como destacam (Berber Sardinha, 2004; Biber, 1993; Bowker; Pearson, 2002) e Capítulo *Conjunto de dados, dataset e corpus*. Isso porque o tamanho precisa ser representativo diante do propósito assumido. Desse modo, por mais que a quantidade seja algo relevante, especialmente no contexto dos *Large Language Models* (LLMs), esse tópico não é fechado. Ele não deve ser tomado isoladamente como critério de qualidade ou suficiência, pois precisa ser pensado em articulação com a finalidade da tarefa para a qual está sendo construído.

O *Audition* é um material que está em fase de conclusão. Contém cerca de 240 minutos de material audiovisual, segmentado, até o momento, em mais de 10.093 unidades visuais anotadas semanticamente via *bounding boxes*. Reúne aproximadamente 5 mil sentenças advindas de diferentes modos de acessibilidade e atuantes em diferentes gêneros cinematográficos. Essas sentenças resultaram, até o momento, em mais de 13 mil conjuntos de anotações (*annotations sets*) e cerca de 118 mil elementos de *frames* anotados. A Figura 17.12 apresenta esses números organizados por gêneros cinematográficos¹⁸.

Checklist do Audition: construindo e avaliando um dataset linguístico multimodal

Baseado nos critérios da Capítulo *Conjunto de dados, dataset e corpus**

Na Figura 17.13, propomos um *checklist* para acompanhamento da criação do *Audition*. Pedimos que você, leitora ou leitor, preencha os campos com base nas informações já

¹⁸Por decisão metodológica, optamos por classificar a animação como um gênero cinematográfico distinto da ficção, com base nas reflexões de (Gordeeff, 2023). Além disso, o *Audition* inclui um exemplar de performance.



Figura 17.12: Panorama quantitativo parcial do *dataset* Audition por gênero cinematográfico

		Ficção	Anima.	Docum.	Perfo.	Total <i>Parcial</i>
Áudio original (OA)	Sentenças	664	337	537	1	1.539
	<i>Annotations sets</i>	1.387	686	2.148	-	4.221
	EFs	12.451	6.112	21.269	-	39.832
Áudio descrição (AD)	Sentenças	1.012	251	491	126	1.880
	<i>Annotations sets</i>	3.748	557	1.642	413	6.360
	EFs	31.445	4.848	15.060	3.585	54.938
Legenda	Sentenças	344	286	521	-	1.151
	<i>Annotations sets</i>	678	16	1.502	-	2.196
	EFs	5.869	145	13.361	-	19.375
Closed Caption (CC)	<i>Sentenças</i>	59	-	73	-	132
	<i>Annotations sets</i>	73	-	138	-	211
	EFs	685	-	1.111	-	1.796
On-screen text	<i>Sentenças</i>	98	39	27	-	164
	<i>Annotations sets</i>	115	13	45	-	173
	EFs	1.150	65	405	-	1.620
Conteúdo de vídeo (VC)	<i>Semantic boxes</i>	3.415	1.576	5.067	35	10.093

apresentadas nesta seção. Após a leitura das seções seguintes, retorne a este quadro para revisar e complementar seus apontamentos. Este exercício pode ajudá-la(o) a refletir sobre os passos metodológicos necessários para a criação de um *dataset* linguístico (e multimodal) de qualidade e a identificar, de forma crítica, como eles se materializam em projetos reais, como este.

17.5.2 Anotação semântica multimodal

A anotação semântica do *Audition* fundamenta-se nos princípios da Semântica de *Frames* (Fillmore, 1982), segundo os quais o significado de uma palavra ou expressão está ancorado em um “quadro” ou “moldura” conceitual, do inglês *frame*. Na linguagem verbal, o *frame* é ativado por unidades lexicais, as ULs, que podem ser verbos, substantivos, adjetivos etc. Cada *frame* envolve um conjunto de papéis semânticos denominados elementos de *frame* (EFs), que representam participantes, objetos, ações e relações relevantes para a cena conceitual evocada.

No contexto multimodal do *Audition*, parte-se da premissa de que não apenas a linguagem verbal, mas imagens, gestos e sons não verbais também evocam *frames*. Assim, a tarefa dos anotadores consiste em identificar e registrar essas evocações em cada modo comunicativo, possibilitando a comparação e o alinhamento semântico entre eles. A anotação



Figura 17.13: *Checklist* do Audition

Critério	Você identificou no Audition?	Comentários e observações
Problema ou tarefa	() Sim () Não () Parcialmente	
Etiquetas de anotação (tagset)	() Sim () Não () Parcialmente	
Diretrizes de anotação (guidelines)	() Sim () Não () Parcialmente	
Compilação ou criação de corpus adequado à tarefa	() Sim () Não () Parcialmente	
Maneira de codificar a anotação	() Sim () Não () Parcialmente	
Pessoas responsáveis pela anotação	() Sim () Não () Parcialmente	
Ferramenta de anotação (se for o caso)	() Sim () Não () Parcialmente	
Estratégias para otimizar o processo de anotação (se for o caso)	() Sim () Não () Parcialmente	
Maneiras de avaliar o resultado da anotação	() Sim () Não () Parcialmente	

deste *dataset* é realizada integralmente por humanos e segue uma **abordagem orientada pelo texto** (*text-oriented*), na qual o processo de anotação parte da análise dos áudios transcritos para guiar a rotulação semântica do conteúdo visual.

No caso específico do *Audition*, adotamos como diretriz metodológica que o conteúdo de vídeo fosse anotado para situações, quais sejam **eventos, estados, processos e relações**. Esse recorte está alinhado com o objetivo do *dataset* de oferecer uma base para a análise da construção de sentido em sequências audiovisuais, priorizando a ação, a mudança de estado e os processos perceptivos. Em contraste, em (Belcavello et al., 2024), o protocolo de anotação previu também a marcação de entidades, por meio de *frames* centrados em participantes ou objetos nas cenas de vídeo. Essa diferença metodológica reflete a flexibilidade do modelo da FrameNet Brasil frente a diferentes objetivos analíticos e tarefas de anotação.

A ferramenta utilizada para anotação é a **Webtool 4.1**, adaptada especificamente para a multimodalidade. Essa plataforma permite trabalhar com os modos em sincronia e oferece funcionalidades para:

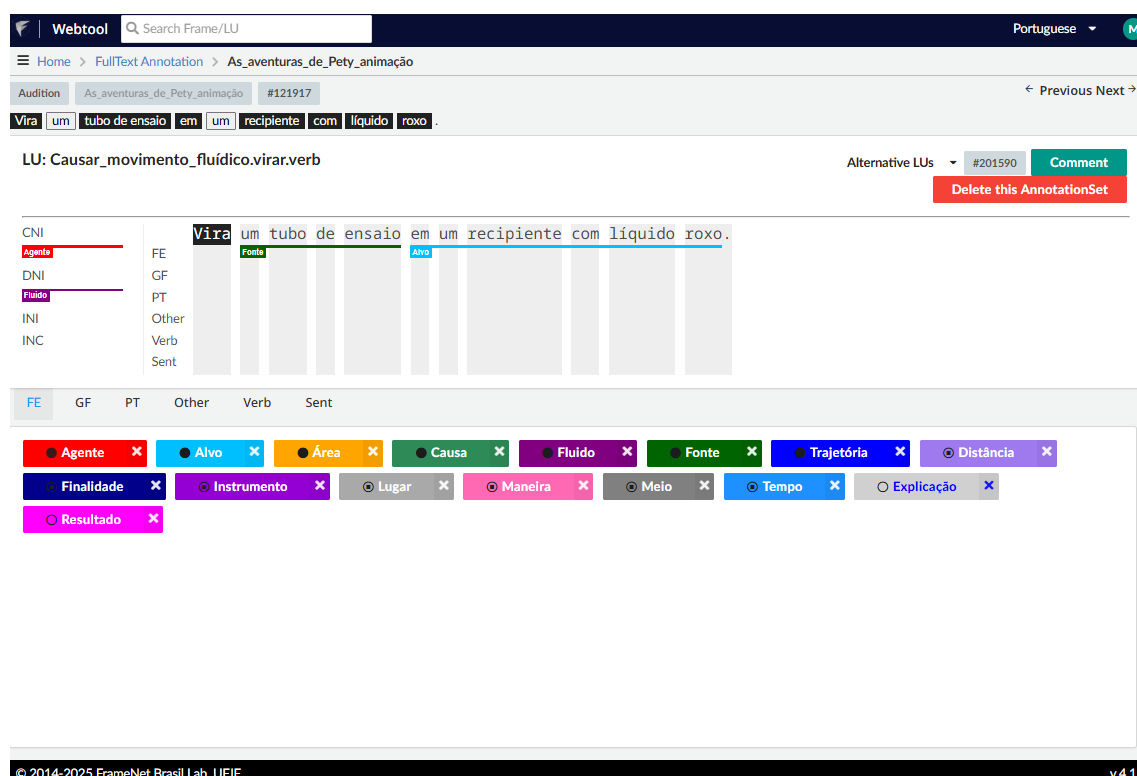
- transcrever o material audiovisual e alinhá-lo ao tempo de exibição do conteúdo de vídeo;
- visualizar simultaneamente os segmentos verbais e visuais;
- criar *bounding boxes* associadas a entidades visuais evocadoras de *frames*;



- selecionar unidades lexicais e entidades visuais;
- associar essas unidades e entidades a *frames* e a seus respectivos elementos, com base na rede semântico-lexical da FrameNet Brasil.

A seguir, serão apresentados exemplos de anotação de texto e de vídeo, com o objetivo de ilustrar os dois processos adotados no *Audition*. O exemplo da Figura 17.14 é um caso de anotação linguística. No caso das modalidades verbais (como Áudio Original e Audiodescrição), a cada UL selecionada, o anotador associa o *frame* correspondente e especifica os elementos de *frame* (EFs) que estiverem lexicalmente instanciados ou inferíveis a partir do contexto verbal ou visual.

Figura 17.14: Anotação linguística no *dataset* Audition



Na sentença “Vira um tubo de ensaio em um recipiente com líquido roxo”, uma das ULs anotadas é “virar.v”, que evoca *Causar_movimento_fluidico*. Esse *frame* prevê EFs como:

- AGENTE: quem faz com que o fluido se mova;
- ALVO: local onde o fluido termina;
- ÁREA: configuração na qual o movimento do fluido ocorre;
- CAUSA: evento ou força que faz com que o fluido se mova;
- FLUIDO: entidade que muda de localização e se movimenta de maneira fluida;
- FONTE: local que o fluido ocupa inicialmente.

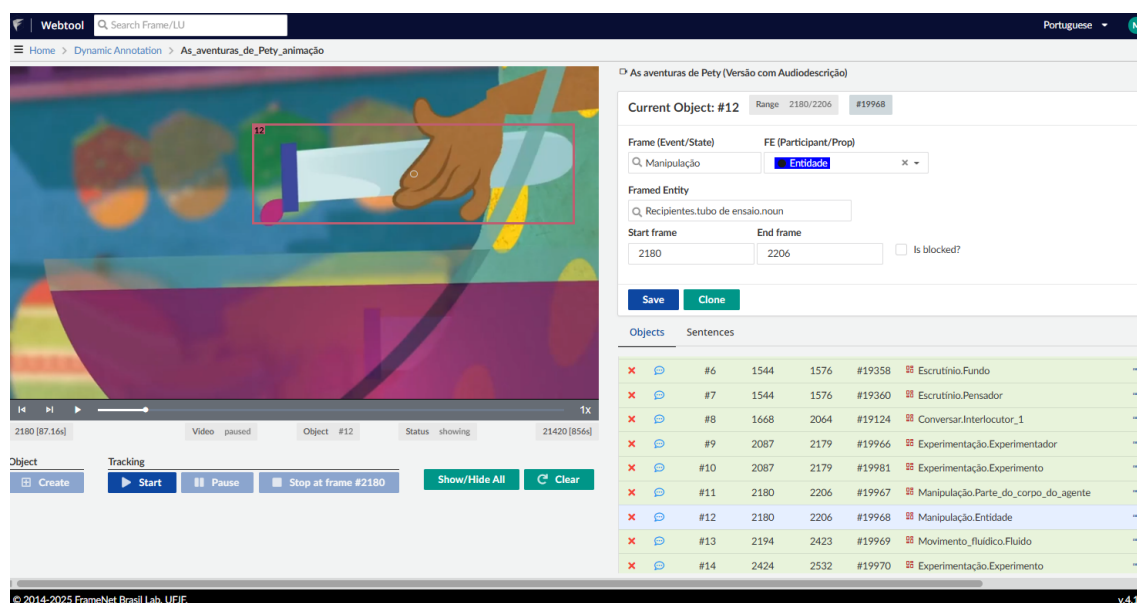


- **TRAJETÓRIA:** A trajetória pela qual o fluido se move.

Nesta sentença, o EF AGENTE é anotado como CNI, pois sua ausência é justificada pela desinência verbal. E o EF FLUIDO foi anotado como DNI, dado que pode ser inferido pelo conteúdo do vídeo. A camada ilustrada na Figura 17.14 corresponde à anotação de texto, de onde se destacam as ULs e seus respectivos *frames* e EFs. A metodologia da FN-Br prevê anotação conjunta de camadas semântica e sintática, ambas organizadas a partir da UL (unidade predicadora). No âmbito da ReINVenTA, no entanto, a anotação ocorre apenas na camada semântica, também chamada de anotação em primeira camada.

No caso do conteúdo de vídeo, a anotação é realizada por meio da marcação de *bounding boxes*, que delimitam visualmente as entidades e situações relevantes no conteúdo audiovisual. Para auxiliar essa tarefa, foi incorporado o YOLOv3 (You Only Look Once), algoritmo de detecção de objetos em tempo real amplamente utilizado em tarefas de visão computacional, como detecção de pessoas, objetos, animais etc. No entanto, no *Audition*, o aproveitamento direto das detecções automáticas foi limitado, de modo que o procedimento predominante consistiu na delimitação manual das entidades visuais, associadas a EFs a partir do *frame* escolhido.

Figura 17.15: Anotação de conteúdo visual dinâmico no *dataset* Audition



A Figura 17.15, referente ao recorte do vídeo correspondente à sentença da AD apresentada na Figura anterior, ilustra um exemplo de anotação de conteúdo de vídeo. Nela, o anotador criou uma *bounding box* para delimitar o artefato em cena. Esse artefato foi anotado como EF ENTIDADE no *frame* Manipulação, que descreve eventos em que um agente manipula uma entidade com um propósito específico. Além da anotação do *frame* semântico que designa o evento ou estado, o objeto visual também é rotulado quanto ao seu próprio *frame*. Assim, a entidade delimitada é um tubo de ensaio, que está associado ao *frame* Recipiente na base da FN.Br. Esse duplo nível de anotação, que especifica tanto o evento quanto a entidade, promove refinamento semântico para o *dataset*. Essa infraestrutura de anotação possibilita o registro das evocações semânticas nos diversos modos presentes no *dataset*, oferecendo base para análises intermodais, comparação entre

recursos de acessibilidade e avaliação de preservação de sentido entre os modos, como será explorado na próxima seção.

17.5.3 Validação, Aplicações em PLN e Impactos na Tradução Audiovisual Acessível

A primeira versão do *dataset Audition*¹⁹ foi empregada em um experimento de avaliação de similaridade semântica entre modos comunicativos, descrito em (Gamonal et al., 2025). O objetivo foi mensurar, por meio de uma **métrica baseada em frames**, o grau de convergência semântica entre os diferentes modos anotados no *corpus* – especialmente o áudio original, a audiodescrição e o conteúdo de vídeo. Trata-se de um exemplo de **validação extrínseca**, tal como definida no Capítulo **Conjunto de dados, dataset e corpus**, pois avalia a capacidade do *dataset* de operar uma tarefa concreta de PLN: a comparação semântica multimodal.

A métrica empregada é uma **medida de similaridade semântica baseada em frames**, inicialmente explorada pela ReINVenTA em (Viridiano et al., 2024). Seu objetivo é quantificar o alinhamento conceitual entre diferentes modos comunicativos, operando em três etapas principais:

- (i) construção de uma **matriz de associação** que representa relações semânticas entre *frames* da FrameNet (como Herança, Subframe, Uso e Perspectiva), segundo sua proximidade conceitual;
- (ii) aplicação de um algoritmo de ativação propagada (***spread activation***) que dissemina a ativação a partir dos *frames* anotados, com decaimento conforme a distância entre os nós no grafo semântico;
- (iii) geração de **vetores de ativação** para cada instância anotada, nos quais cada dimensão representa um *frame* e seu valor de ativação acumulado.

A similaridade entre duas instâncias é, então, calculada por meio da similaridade do cosseno entre esses vetores, resultando em um *score* normalizado entre 0 (nenhuma similaridade) e 1 (similaridade total). A métrica permite, assim, avaliar não apenas a sobreposição direta de *frames*, mas também a proximidade conceitual entre eles, oferecendo uma medida interpretável da **aderência semântica** entre os modos.

Os resultados do experimento, sintetizados na Tabela 17.4, compararam os resultados entre áudio original (OA), audiodescrição (AD) e conteúdo de vídeo (CV) a partir de: (i) unidades lexicais (ULs); (ii) *frames* e elementos de *frame* (EFs) e (iii) todos os componentes anotados em conjunto (UL + *frame* + EFs), conforme apresentado na Tabela 17.4.

Tabela 17.4: Similaridade de cosseno entre representações semânticas baseadas em frame de OA, AD e VC de conteúdo visual dinâmico no *dataset Audition*

	pairs	avg	var	stdev
AD LU	323	0.42	0.05	0.22
AD Frame and FE	323	0.38	0.05	0.22

¹⁹O acesso aos dados da primeira versão do *dataset* estão disponíveis em: <https://huggingface.co/dataset/s/FrameNetBrasil/Audition>.



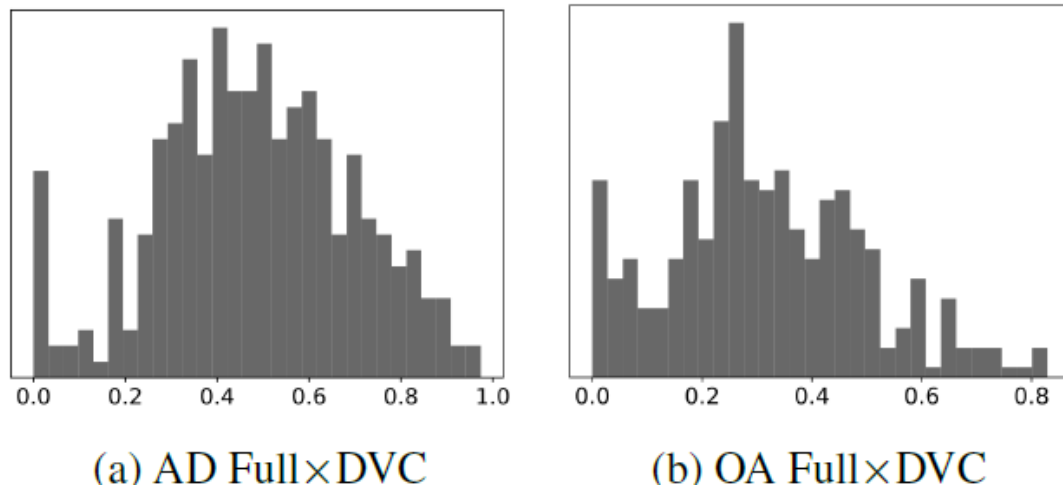
	pairs	avg	var	stdev
AD Full	323	0.49	0.05	0.21
OA LU	357	0.23	0.03	0.17
OA Frame and FE	357	0.29	0.04	0.19
OA Full	357	0.32	0.03	0.18

Fonte: (Gamonal et al., 2025) (no prelo)

As médias de similaridade por cosseno foram superiores entre as representações de AD e VC (máximo de 0.49), em relação às de OA e VC (máximo de 0.32). Esses resultados validam a hipótese de que a audiodescrição desempenha, de fato, a função de mediação semântica entre o conteúdo visual e o público-alvo, e demonstram que o *dataset* pode ser utilizado para mensurar o nível de similaridade semântico com rigor metodológico.

Essa tendência é reforçada pela distribuição dos valores de similaridade apresentados na Figura 17.16.

Figura 17.16: Distribuição de valores de similaridade entre AD full, AO Full e CVD



Fonte: (Gamonal et al., 2025) (no prelo)

A AD mantém maior convergência semântica com o vídeo do que o áudio original, como esperado. As comparações que incluem todos os componentes anotados (*full*) produziram as maiores médias de similaridade, o que reforça a validade da anotação semântica e do próprio *design* do *dataset* para finalidades analíticas.

Esses resultados indicam o potencial do *Audition* para servir a aplicações de PLN no campo da acessibilidade, como:

- **Geração automática de audiodescrição**, com base na associação entre o conteúdo visual segmentado por meio das *bounding boxes* e suas respectivas evocações semânticas;
- **Detecção de similaridade semântica entre modos**, útil para tarefas de reformulação, sumarização ou avaliação de traduções intersemióticas;

- **Classificação e extração de informações** em ambientes multimodais, como identificação de entidades visuais e suas funções em narrativas audiovisuais;
- **Avaliação de coerência narrativa** entre trilha sonora, falas, imagens e recursos de acessibilidade, especialmente em contextos educativos e culturais.

No campo da Tradução Audiovisual Acessível, o *Audition* representa um avanço metodológico por permitir a avaliação sistemática da construção de sentido entre modos de tradução intersemiótica. O estudo semântico por meio de *frames* entre modos – como AD, legenda, *closed caption* e *on-screen text* – permite avaliar a eficácia desses recursos e propor melhorias na sua elaboração e padronização.

Além disso, por ser um *dataset* anotado manualmente e estruturado a partir de uma teoria semântica atrelada aos estudos sociocognitivos da linguagem, o *Audition* caminha para se tornar um recurso padrão-ouro para o desenvolvimento de tecnologias voltadas à acessibilidade cultural. O alinhamento temporal e semântico entre os modos proporciona base para aplicações em *machine learning* supervisionado, incluindo o treinamento de modelos multimodais com foco em acessibilidade.

17.6 Considerações finais

Este capítulo buscou apresentar como a aplicação computacional da Semântica de *Frames*, a FrameNet Brasil, pode contribuir com aplicações transformadoras no campo do Processamento de Linguagem Natural no que tange a causas sociais. Para isso, discutiu-se o papel da anotação computacional em duas frentes – manual e automática – como estratégia central para a construção de sistemas computacionais mais interpretáveis e sensíveis aos contextos de uso, partindo de dois projetos: *DataToStopGBV* e *Audition*.

No que concerne ao projeto DataToStopVBG, os resultados evidenciam os benefícios do uso da metodologia da FrameNet para interpretar registros abertos nos sistemas de saúde, auxiliando na identificação de possíveis casos de VBG e possíveis padrões de violência. Ao avançar além das abordagens estatísticas tradicionais, o projeto revelou possíveis correlações entre violência e questões de saúde das vítimas, abrindo caminhos para investigações futuras e intervenções em parceria com profissionais da saúde. Dessa forma, os resultados revelam como a aplicação da FrameNet pode ser uma aliada no enfrentamento da violência.

Neste momento, duas vertentes do projeto estão em andamento. Primeiramente, o estudo mais profundo de possíveis padrões semânticos ligados à violência e a condições de saúde dentro desse contexto. Espera-se que esses dados sejam validados no âmbito da análise da FrameNet, em um trabalho conjunto com profissionais da saúde. Paralelamente a esse estudo, está sendo desenvolvido um painel com dados sobre a subnotificação de violências e padrões semânticos relacionados à violência de gênero, com o objetivo de gerar mapeamentos descritivos sobre tais casos no município de Recife. Por fim, o próximo passo envisioned para esta iniciativa consiste no estudo dos dados com grupos de foco mais específicos, tais como violência autoprovocada e infantil.

No âmbito do projeto *Audition*, os resultados não foram diferentes, apontando para as vantagens da aplicação da metodologia da FrameNet para tarefas de PLN com foco em acessibilidade. Os resultados da análise apontam que a audiodescrição tem maior convergência semântica com o conteúdo visual do que o áudio original, o que valida os benefícios da anotação multimodal empregada, além de indicar a possibilidade da aplicação

para a geração automática de audiodescrição. O *Audition*, então, se destaca como *dataset* promissor para tecnologias de tradução audiovisual acessível e referência no campo da acessibilidade cultural.

A próxima etapa central do projeto é a finalização do *dataset*. Uma versão inicial foi compilada como prova de conceito, permitindo validar o recurso por meio da medida de similaridade semântica. Paralelamente, estamos desenvolvendo um guia semântico para a audiodescrição fílmica, com o objetivo de apresentar as contribuições da tecnologia linguística para a experiência cinematográfica acessível. Essa iniciativa destaca a importância do uso de *frames* semânticos na estruturação de roteiros de audiodescrição, promovendo aderência ao conteúdo visual e o cuidado com a experiência fílmica do usuário. Espera-se, assim, que o *Audition* seja amplamente utilizado em tarefas voltadas à promoção da inclusão cultural por meio da tradução audiovisual acessível em português brasileiro.

Além da descrição dos projetos em desenvolvimento na FrameNet Brasil, este capítulo procurou evidenciar o valor da aplicação da Semântica de *Frames* como fundamento para o desenvolvimento de recursos linguísticos e metodologias voltadas a demandas sociais urgentes. Os estudos de caso apresentados ilustram caminhos distintos de aplicação da anotação semântica que demonstram a amplitude do alcance que a aplicação da FrameNet como recurso de PLN possui. Espera-se, então, que esta discussão contribua para a expansão do escopo de atuação em PLN no português brasileiro, especialmente em iniciativas que tragam benefícios concretos para a população.

Agradecimentos

O projeto DataToStopGBV foi apoiado pelo Programa de Aceleração da Fundação Patrick J. McGovern, pela Fundação José Luiz Setúbal e pelo Instituto Galo da Manhã. Também expressamos nossa gratidão aos nossos parceiros da Secretaria Municipal de Saúde do Recife — Luciana Caroline, Marcella Abath, Natalia Barros e Yana Lopes — pela valiosa colaboração e apoio contínuo ao longo deste projeto. O *dataset* *Audition* integra a ReINVenTA (Rede de Pesquisa e Inovação para Análise Visual e Textual de Objetos Multimodais), financiada pela FAPEMIG (RED-00106-21) e pelo CNPq (420945/2022-9), e contou também com apoio da CAPES (88887.936139/2024-00) e do CNPq (151361/2023-1). M. A. Gamonal é bolsista de pós-doutorado (CAPES – processo 88887.015648/2024-00). T. T. Torrent é bolsista de produtividade do CNPq (311241/2025-5). A rede ReIN-Venta é apoiada pelo Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação da Informação (INCT-TILDIAR), financiado pelo CNPq, processo nº 408490/2024-1.

Glossário deste capítulo

- **Semântica de *Frames*:** Modelo teórico proposto por Charles Fillmore, segundo o qual o significado de uma palavra está associado a um esquema conceitual (*frame*) que organiza papéis semânticos e elementos relacionados a uma cena ou situação.
- ***Frame*:** Estrutura conceitual que representa uma situação recorrente, com seus participantes, objetos e relações, ativada pelo uso de determinadas palavras ou expressões.



- **Elemento de *Frame* (EF):** Papel semântico associado a um *frame*, como Agente, Paciente, Causa etc.
- **Unidade Lexical (UL):** Combinação de um lema e um sentido específico que ativa um *frame*. Exemplo: em “Eu *comprei* a maçã da promoção”, “comprar.v” evoca Comércio_comprar, já em “Os moradores *compraram* a briga contra o projeto de lei”, o *frame* evocado é Encontro_hostil.
- **FrameNet Brasil (FN-Br):** Projeto que adapta e expande a FrameNet original (em inglês) para o português brasileiro, com adaptações teóricas e metodológicas, além de utilizar ferramentas computacionais próprias.
- **Relações *Frame-a-Frame*:** Conexões entre *frames* baseadas em relações, como Subframe, Herança, Perspectiva, Uso etc.
- **Relações Qualia Ternárias:** Conjunto de relações semânticas inspiradas na Teoria do Léxico Gerativo, proposta por James Pustejovsky e utilizada na FN-Br para modelar aspectos internos dos conceitos, como finalidade, composição e origem.
- **Anotação semântica baseada em *frames*:** Processo de rotulação de unidades lexicais, *frames* e elementos de *frame* em textos e outros modos com base na teoria da Semântica de *Frames*.
- **Anotação automática:** Análise feita por modelos computacionais supervisionados ou não supervisionados, que aplicam rótulos semânticos automaticamente a partir de dados previamente anotados.
- **Anotação manual:** Processo de anotação realizado por linguistas especialistas e/ou estudantes em formação em linguística, seguindo políticas rigorosas e validadas, com foco na interpretação precisa dos dados.
- **Padrão-ouro (*gold standard*):** *Dataset* anotado manualmente com alto grau de confiabilidade, usado como referência para treinamento e avaliação de sistemas automáticos.
- **Webtool:** Ferramenta web desenvolvida pela FN-Br para visualização e anotação de dados semânticos.
- **Lutma:** Ferramenta para não especialistas estruturada por meio de tutoriais com passo a passo para criação de *frames* semânticos.
- **LOME (*Large Ontology Multilingual Extraction*):** Modelo de rotulação semântica automática que opera com base na Semântica de *Frames*.
- ***Bounding box*:** Retângulo utilizado em visão computacional para delimitar objetos ou entidades visuais em imagens ou vídeos.
- **Similaridade semântica:** Medida computacional do grau de proximidade conceitual entre itens lexicais, expressões ou representações com base em sua estrutura semântica.
- ***Spread activation* (ativação propagada):** Técnica computacional que simula a ativação de nós conectados em uma rede semântica, permitindo capturar relações indiretas entre *frames*.



- **Similaridade do cosseno:** Métrica de comparação entre dois vetores, usada para mensurar o ângulo entre eles e, assim, sua similaridade conceitual.
- **Audiodescrição (AD) - *fílmica*:** Recurso de acessibilidade que consiste na narração de informações visuais relevantes para a experiência fílmica de produtos audiovisuais, voltado a pessoas cegas ou com baixa visão.
- **Tradução intersemiótica:** Tradução de significados entre sistemas semióticos diferentes, como de imagem para texto, ou de som para linguagem verbal.
- **Tradução Audiovisual Acessível:** Campo que estuda e desenvolve recursos de tradução que garantam acessibilidade a produtos audiovisuais, como audiodescrição, legendas, *closed caption* etc.
- **Violência baseada em gênero (VBG):** Qualquer forma de violência direcionada contra indivíduos com base em seu gênero, frequentemente associada a relações desiguais de poder.
- **DataToStopGBV:** Projeto que aplica anotação semântica baseada em *frames* e PLN a registros do SUS para apoiar o diagnóstico e a prevenção da violência baseada em gênero.
- **Audition:** *Dataset* multimodal anotado semanticamente, voltado à análise e geração de traduções acessíveis, como audiodescrição, com base na convergência entre modos comunicativos.

Capítulo 18

Redes complexas no processamento de língua natural do português

Vitor H. B. D. Santi
Lucas D. V. Figueiredo
Diego R. Amancio
Maria Cristina F. Oliveira
Oswaldo N. Oliveira Jr.

Publicado em: 16/04/2026

Este capítulo explora a aplicação de redes complexas no Processamento de Linguagem Natural (PLN) para o português. Redes complexas, estruturas compostas por nós (entidades) e arestas (conexões), servem de arcabouço para analisar relações linguísticas em tarefas como sumarização extrativa, análise de sentimentos e avaliação de qualidade textual, classificação de textos, detecção de notícias falsas, avaliação de traduções e diagnósticos médicos a partir de transcrições. São introduzidos conceitos fundamentais (como centralidade e coeficiente de aglomeração) e modelos de rede (aleatórias, livres de escala e mundo-pequeno). Um tópico central é a detecção de comunidades por algoritmos que identificam agrupamentos temáticos em textos. O capítulo destaca contribuições pioneiras de pesquisadores brasileiros, que utilizaram redes para tarefas desde tradução automática até diagnóstico médico. Para mapear a pesquisa em PLN para português, são empregadas redes de similaridade, identificando comunidades temáticas como inteligência artificial aplicada à educação, construção e utilização de *corpus* em tarefas de PLN, análise de sentimentos e desinformação. Conclui-se que, embora modelos de linguagem de larga escala (LLMs) apresentem desempenho superior em tarefas de PLN, as redes complexas continuam essenciais para tarefas de análise estrutural, interpretabilidade e visualização da literatura científica.

18.1 Introdução

Redes complexas são estruturas formadas por nós — que representam alguma entidade de interesse — e arestas — que estabelecem conexões entre pares de nós. Essas redes tipicamente descrevem sistemas com padrões de interconexão não triviais e oferecem um arcabouço para a análise de sistemas dinâmicos em diversos domínios, incluindo biologia, telecomunicações e ciências sociais (Estrada, 2011; Newman, 2010). Sua relevância se estende à área do Processamento de Linguagem Natural (PLN), em que o entendimento das relações entre palavras e conceitos é essencial para uma gama de tarefas como sumarização, análise de sentimento, modelagem de tópicos e recuperação de informação. No PLN, os



textos são frequentemente representados por métodos que focam no conteúdo, como o modelo *Bag-of-Words* (BoW), que desconsidera a ordem das palavras, ou por modelos de *embeddings* (como *Word2Vec* ou *BERT*), que capturam o significado semântico em um espaço vetorial. A abordagem de Redes Complexas (RC) é fundamentalmente diferente e complementar. Em vez de focar apenas na frequência ou na semântica de palavras isoladas, a RC modela explicitamente a estrutura relacional do texto. Ao representar palavras, sentenças ou conceitos como nós e suas conexões (sejam elas de coocorrência, adjacência ou relações sintáticas) como arestas, uma RC permite analisar a topologia do discurso. Essa perspectiva estrutural revela propriedades que outras representações não capturam diretamente. Com elas, é possível quantificar a coesão textual, identificar quais conceitos são mais centrais (ou “hubs”) em uma argumentação e detectar a organização temática através da detecção de comunidades, como veremos na Seção 3. Portanto, as redes complexas oferecem uma “lupa” sobre a organização e a dinâmica do texto, indo além do conjunto de palavras que ele contém.

Medidas extraídas de redes complexas, como coeficientes de aglomeração e métricas de centralidade de nós, são fundamentais para entender a sua dinâmica e comportamento (OAEPublish, 2025). As redes podem também apresentar propriedades – como serem livre-de-escala ou pequeno-mundo – que impactam como as informações relacionadas às entidades se propagam e se estruturam no sistema (SpringerOpen Applied Network Science, 2017). Há grandes desafios para modelar redes, sobretudo porque muitos sistemas apresentam propriedades de autodissimilaridade em múltiplas escalas, para os quais a elaboração de modelos realistas requer esforço significativo. Ademais, a natureza evolutiva das redes complexas exige técnicas de análise para capturar as suas dinâmicas, especialmente em aplicações do mundo real, em que a disponibilidade de dados pode ser uma limitação (ScienceDirect, 2025a). Essa combinação de um sólido arcabouço teórico com a possibilidade de aplicações práticas ressalta a relevância das redes complexas para a pesquisa em PLN e seu potencial para ampliar a nossa compreensão das interações linguísticas. Também são assunto de debates as implicações do uso de redes complexas em várias áreas, por exemplo, a necessidade de uso ético de dados e potenciais vieses em interpretações algorítmicas. À medida que pesquisadores se aprofundam no estudo dessas redes, ganha cada vez mais importância a discussão sobre sua relevância e as responsabilidades associadas ao seu uso.

Neste capítulo, buscamos apresentar uma visão sobre usos de redes complexas no processamento da língua portuguesa. O conteúdo do capítulo está organizado da seguinte maneira: na Seção 18.2 introduzimos os conceitos e a terminologia básica em Redes Complexas. Na Seção 18.3 introduzimos o problema de detecção de comunidades em redes e alguns algoritmos clássicos. Na Seção 18.4 discutimos algumas aplicações de redes complexas em PLN para o português, e na Seção 18.5 recorremos às redes complexas para apresentar um panorama sobre as pesquisas em PLN no âmbito da Língua Portuguesa, não só as que envolvem redes complexas. As considerações finais são apresentadas na Seção 18.6.

18.2 Redes complexas

Como o objetivo deste capítulo é ilustrar o uso de redes complexas em PLN, apresentaremos algumas informações essenciais para a compreensão das aplicações, sem preocupação de tratar o tema de forma abrangente ou exaustiva. Assim, além de introduzir conceitos fundamentais, discutiremos as métricas das redes usadas em PLN, os modelos de redes

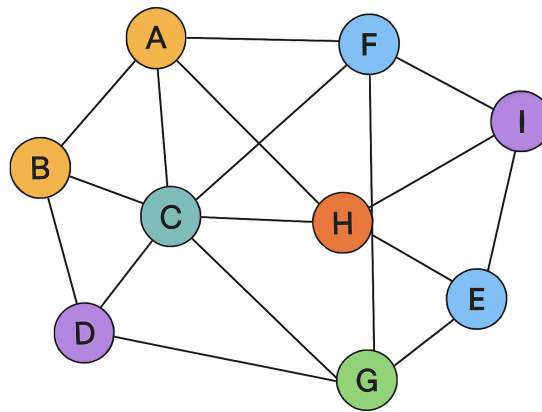


e a dinâmica ou evolução temporal de redes. Uma introdução ao tema de redes complexas, que aborda conceitos e algumas aplicações, pode ser encontrada na referência (Appel; Junior, 2011).

18.2.1 Conceitos fundamentais

Os elementos fundamentais de uma rede são os **nós**, também denominados vértices, que representam as entidades de interesse, e as conexões entre nós, denominadas **arestas**. Formalmente, uma rede (ou grafo) pode ser representada como um par $G = (V, E)$, em que V é o conjunto de nós (ou vértices) e E é o conjunto de arestas.

Figura 18.1: Representação visual de uma rede não direcionada, em que os círculos representam os nós e as linhas representam as arestas.



A Figura 18.1 ilustra uma representação visual de uma rede. Cada **nó** $v \in V$ pode ser associado a dados que registram atributos específicos da entidade correspondente, representados como pares chave-valor. Por exemplo, um nó que representa uma pessoa pode incluir chaves como “nome”, “idade”, “localização”, com seus valores correspondentes (Mousavi, 2024).

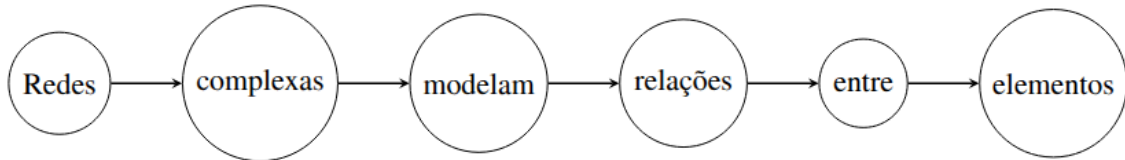
As **arestas** estabelecem as conexões entre as entidades. Cada aresta é um par ordenado ou não ordenado de nós, isto é, $e = (u, v)$ com $u, v \in V$. Quando o par é não ordenado, isto é, $(u, v) = (v, u)$, a rede é dita **não-direcionada**; quando o par é ordenado, a rede é **direcionada** e (u, v) representa uma conexão do nó u para o nó v . Além disso, as arestas podem ter pesos associados, definindo uma **rede ponderada**, em que cada aresta $e \in E$ tem um peso $w(e) \in \mathbb{R}$, representando a intensidade ou custo da relação.

Em uma rede social, uma aresta indicativa de uma relação de amizade pode ser bidirecional, enquanto que uma aresta representando uma relação de compra seria unidirecional – em ambos os casos, evidenciando o fluxo direcional da relação. Em redes sociais, um exemplo de relação unidirecional ocorre quando um usuário segue outro, mas esse segundo usuário não o segue de volta. Isso é comum em plataformas como o *X* ou o *Instagram*. A título de ilustração de uma representação textual simples, considere, por exemplo, a sentença “Redes complexas modelam relações entre elementos”. Podemos representá-la como uma rede em que cada palavra é um nó (conjunto de nós V) e as arestas conectam as palavras que aparecem consecutivamente (conjunto de arestas E) (Figura 18.2):

$$V = \{\text{Redes, complexas, modelam, relações, entre, elementos}\}$$

$$E = \{(\text{Redes, complexas}), (\text{complexas, modelam}), (\text{modelam, relações}), (\text{relações, entre}), (\text{entre, elementos})\}$$

Figura 18.2: Representação de uma sentença como uma rede, em que cada palavra é um nó e as arestas conectam palavras consecutivas.



A importância de um nó na rede é tipicamente refletida pelo seu **grau**. O grau de um vértice v , denotado por $\deg(v)$, é definido como o número de arestas incidentes a v . Em redes direcionadas, distinguimos o grau de entrada ($\deg^-(v)$) e o grau de saída ($\deg^+(v)$), correspondendo, respectivamente, ao número de arestas que chegam e ao número de arestas que saem de v . Um grau alto tipicamente está associado a um nó mais central ou influente na rede.

A representação de entidades e conexões em termos de nós e arestas define a essência das redes complexas, permitindo modelar sistemas complexos do mundo real, como redes sociais e padrões de comunicação na linguagem (ScienceDirect, 2025b). Formalmente, o estudo de redes complexas utiliza a teoria dos grafos. Por exemplo, podemos definir um grafo $G = (V, E)$ em que V representa palavras, sentenças ou documentos, e E reflete relacionamentos entre eles, como coocorrência em um *corpus* ou relacionamentos sintáticos em uma sentença.

18.2.2 Propriedades

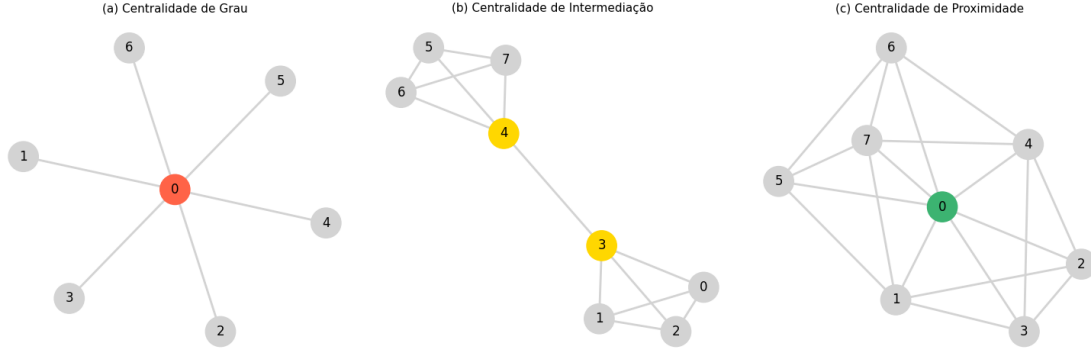
Várias propriedades exibidas pelas redes permitem entender a sua estrutura e comportamento, bem como analisar os mecanismos de propagação de informação na rede. **Medidas de centralidade** permitem quantificar e entender a importância de nós individuais. Além do **grau** (também conhecido como **Centralidade de Grau**), métricas como Centralidade de Intermediação e Centralidade de Proximidade fornecem indicações sobre como os nós estabelecem pontes ou estão estrategicamente posicionados dentro da rede. A **Centralidade de Intermediação** quantifica com que frequência um nó atua como uma ponte ao longo dos caminhos mais curtos entre outros nós – em PLN, por exemplo, isso poderia revelar palavras ou conceitos influentes em um texto. A **Centralidade de Proximidade** avalia quão rapidamente é possível alcançar todos os outros nós a partir de um dado nó, destacando os principais atores na distribuição de informações. A Figura 18.3 ilustra esses conceitos.

Para um grafo com N nós, as definições matemáticas de algumas dessas centralidades para um nó v são:

- **Centralidade de Grau (*Degree Centrality*)**: em redes direcionadas, distingue-se entre o grau de entrada e o de saída. A forma normalizada é dada por:
 - Grau de Entrada (*In-Degree*): $C_D^{\text{in}}(v) = \frac{k_{\text{in}}(v)}{N-1}$, onde $k_{\text{in}}(v)$ é o número de arestas que chegam a v .



Figura 18.3: Representação visual para as medidas de centralidade mencionadas. (a) centralidade de grau. O nó 0 é considerado central, pois tem grau 6 enquanto os demais têm grau 1. (b) centralidade de intermediação. Os nós 3 e 4 são centrais para a conexão entre os dois grupos de nós mostrados na figura. (c) centralidade de proximidade. O nó 0 tem alta centralidade de proximidade, pois através dele os demais nós podem ser acessados mais facilmente. Ele obviamente também tem centralidade de grau, por ser mais conectado que os demais.



- Grau de Saída (*Out-Degree*): $C_D^{\text{out}}(v) = \frac{k_{\text{out}}(v)}{N-1}$, onde $k_{\text{out}}(v)$ é o número de arestas que saem de v .

- **Centralidade de Intermediação (*Betweenness Centrality*):**

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

onde σ_{st} é o número total de caminhos mais curtos do nó s para o nó t , e $\sigma_{st}(v)$ é o número desses caminhos que passam por v .

- **Centralidade de Proximidade (*Closeness Centrality*):**

$$C_C(v) = \frac{N-1}{\sum_{u \neq v} d(v, u)}$$

onde $d(v, u)$ é a distância do caminho mais curto entre os nós v e u .

Além de medidas de centralidade, várias estratégias foram propostas para a identificação de nós influentes na estrutura da rede, adaptadas a diferentes tipos de rede e cenários de aplicação. O **coeficiente de aglomeração** é uma medida de como os nós se agrupam localmente. Um alto coeficiente de aglomeração indica interconexões fortemente localizadas entre subgrupos de nós. No contexto de PLN, por exemplo, pode indicar a presença de conceitos ou palavras fortemente relacionados em um contexto particular ou tópico específico (Budell et al., 2023). Esta é uma propriedade particularmente relevante para análise semântica, em que é crucial entender os relacionamentos entre palavras ou frases ao implementar tarefas como desambiguação lexical ou reconhecimento de entidades nomeadas (Al-Rfou et al., 2014).

O **Coefficiente de Aglomeração Local** para um nó v é definido como:

$$C(v) = \frac{2E_v}{k_v(k_v - 1)}$$

onde k_v é o grau do nó v (o número de seus vizinhos) e E_v é o número de arestas existentes entre os vizinhos de v . O denominador $k_v(k_v - 1)/2$ representa o número máximo possível de arestas entre os vizinhos de v .

Uma revisão abrangente dessas estratégias as categoriza em abordagens tradicionais baseadas em topologia, para otimização, modelos de difusão de informação, métodos de aprendizado de máquina e técnicas mais recentes aplicáveis a redes dinâmicas (Chen et al., 2025). Ao contrário das medidas clássicas de centralidade, esses métodos consideram interações na rede ao longo do tempo, além das propriedades estáticas. Outra métrica relevante, denominada **Comprimento de Caminho Médio**, indica o número médio de passos ao longo dos caminhos mínimos entre todos os possíveis pares de nós. Em redes complexas, essa métrica tende a ter valores baixos, o que facilita a disseminação eficiente de informação na rede. O **Diâmetro**, por outro lado, é definido como o maior valor entre os comprimentos dos menores caminhos entre qualquer par de nós da rede.

As definições matemáticas para essas métricas são:

- **Comprimento de Caminho Médio (Average Path Length):**

$$L = \frac{1}{N(N-1)} \sum_{i \neq j} d(v_i, v_j)$$

onde $d(v_i, v_j)$ é a distância do caminho mais curto entre os nós v_i e v_j .

- **Diâmetro (Diameter):**

$$D = \max_{i,j} d(v_i, v_j)$$

que representa a maior distância de caminho mais curto entre qualquer par de nós (v_i, v_j) na rede.

18.2.3 Tipos de redes

As redes podem ser categorizadas com base em suas propriedades estruturais e comportamento, que de alguma forma refletem a natureza intrincada dos sistemas do mundo real. Muito conhecidas são as redes aleatórias, livres de escala e mundo-pequeno.

Redes aleatórias

Nesse tipo de rede, as conexões entre os nós são estabelecidas aleatoriamente. Pode-se citar, a título de exemplo, redes de sensores distribuídos aleatoriamente em ambientes físicos ou redes de coocorrência de palavras construídas a partir de textos embaralhados ou gerados aleatoriamente. A despeito de terem servido de base para estudos teóricos importantes na ciência de redes, elas não refletem o comportamento e os padrões de interconectividade observados em sistemas do mundo real. Ainda assim, a análise desse tipo de rede é importante para ilustrar os princípios fundamentais que governam redes de tipos mais complexos. O modelo canônico para redes aleatórias é o de **Erdős-Rényi (ER)**, $G(N, p)$. Sua formação se dá a partir de N nós, onde cada par de nós possível é



conectado por uma aresta com uma probabilidade p independente. A distribuição de grau $P(k)$ de uma rede aleatória segue uma distribuição binomial, que, para redes grandes e com baixa probabilidade de conexão, se aproxima de uma **distribuição de Poisson**:

$$P(k) \approx \frac{\lambda^k e^{-\lambda}}{k!}$$

onde λ é o grau médio da rede. A maioria dos nós tem um grau próximo da média. Em redes aleatórias do tipo ER, a distância típica entre vértices é relativamente pequena.

Redes livres de escala

Essas redes se caracterizam pelo fato de a distribuição de grau dos seus nós seguir uma lei de potência. Em outras palavras, um número reduzido de nós tem um altíssimo número de conexões (são os chamados “*hubs*”), enquanto a grande maioria dos demais é muito pouco conectada. Redes de citação entre artigos científicos e redes de coocorrência de palavras em grandes *corpora*, como a *Wikipedia* ou o *Google Books*, são exemplos de redes livres de escala (Steyvers; Tenenbaum, 2005). Em PLN, esse fenômeno está relacionado à **Lei de Zipf** (Zipf, 1949), um princípio fundamental da linguística. A Lei de Zipf postula que a frequência de uma palavra em um *corpus* é inversamente proporcional à sua posição em um *ranking* de frequência. Em uma rede de coocorrência de palavras, os termos mais frequentes (os *hubs*) naturalmente coocorrem com uma variedade muito maior do que outras palavras. Por outro lado, a maioria das palavras — aquelas de baixa frequência, que compõem a chamada “cauda longa” de Zipf — tende a apresentar poucas conexões.

A distribuição de grau em lei de potência observada nessas redes textuais, ilustrada na Figura 18.4, é, em grande parte, um reflexo da Lei de Zipf que governa a frequência das palavras. O surgimento desse tipo de rede pode ser explicado por mecanismos como crescimento e conexão preferencial, ilustrados no conhecido modelo Barabási-Albert (BA), proposto por Albert-László Barabási e Réka Albert em 1999 (Barabási; Albert, 1999). O padrão de distribuição de conexões segue uma **lei de potência**, em que a probabilidade $P(k)$ de um nó ter k conexões é dada por:

$$P(k) \sim k^{-\gamma}$$

em que γ é o expoente da distribuição, geralmente um valor entre 2 e 3. A formação pelo modelo **Barabási-Albert (BA)** se baseia no mecanismo de **conexão preferencial**: a probabilidade $\Pi(v_i)$ de um novo nó se conectar a um nó existente v_i é proporcional ao seu grau k_i :

$$\Pi(v_i) = \frac{k_i}{\sum_j k_j}$$

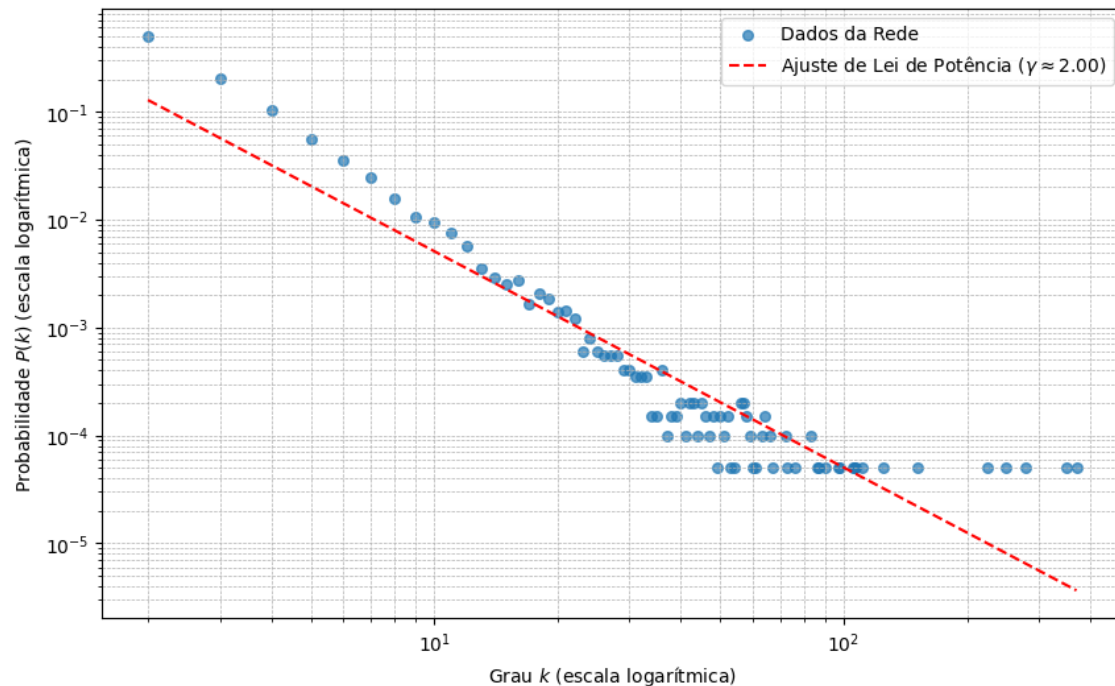
A inserção de novos nós com probabilidade de conexão proporcional ao grau dos nós já existentes resulta no fenômeno conhecido como “rico-fica-mais-rico”, que induz à formação de *hubs*. Embora o modelo BA reproduza a característica de existência de *hubs*, algumas características de redes reais ainda não são representadas por este modelo. Uma delas é a possibilidade de surgimento de novas arestas entre vértices antigos ou o desaparecimento de arestas.

Redes mundo-pequeno

Essas redes são caracterizadas por comprimentos de caminho curtos e coeficientes de aglomeração altos, o que permite que a maioria dos nós seja acessível a partir de qualquer nó



Figura 18.4: Distribuição de probabilidade para um grau k (eixo Y, logarítmico) versus o grau k (eixo X, logarítmico) em uma rede livre de escala. Os dados são relativos a uma rede sintética obtida usando um algoritmo que simula o modelo de Barabási-Albert. Foram gerados 20.000 nós, sendo que cada novo nó inserido na rede se conecta a dois nós pré-existent. Os pontos seguem aproximadamente a linha tracejada vermelha indicativa da lei de potência, que é a “assinatura” visual dessa distribuição.



com um número pequeno de passos. Essa propriedade é frequentemente exemplificada pelo conceito dos “seis graus de separação”. Segundo esse conceito, quaisquer duas pessoas no mundo estariam — em média — a seis conexões entre si. A rede de atores de Hollywood que atuaram juntos exhibe propriedades de mundo pequeno, assim como a *WordNet* e redes sintáticas de línguas naturais (Steyvers; Tenenbaum, 2005).

O modelo mundo-pequeno, introduzido em 1998 por Duncan J. Watts e Steven Strogatz (Watts; Strogatz, 1998), ilustra como uma rede pode transicionar de um formato de grade regular para uma estrutura mais aleatória enquanto mantém suas características de mundo-pequeno. Isso ocorre por meio de uma reatribuição aleatória de arestas, o que enfatiza as conexões sem alterar significativamente a estrutura global. O modelo de **Watts-Strogatz (WS)** descreve a formação dessas redes a partir de uma rede regular (anel) em que cada aresta é religada aleatoriamente com uma probabilidade p . O padrão de distribuição não é definido por uma fórmula única de grau, mas pela coexistência de duas propriedades:

1. **Alto Coeficiente de Aglomeração (C):** Similar ao de uma rede regular e muito maior que o de uma rede aleatória ($C \gg C_{\text{aleatória}}$).
2. **Baixo Comprimento de Caminho Médio (L):** Similar ao de uma rede aleatória, crescendo logaritmicamente com o número de nós ($L \sim \log(N)$).

Redes complexas possuem várias características definidoras, incluindo heterogeneidade na distribuição do grau dos nós, alta aglomeração e modularidade. A diversidade nas conexões leva à formação de grupos coesos (comunidades) na rede, com conexões mais esparsas mantidas entre diferentes grupos. Essa modularidade é crucial para entender como as redes operam e evoluem.

18.2.4 Dinâmica

O estudo da dinâmica de redes complexas permite determinar como elas evoluem ao longo do tempo e influenciam os vários processos que nelas ocorrem, o que é importante para entender fenômenos como a propagação de doenças e a difusão de informações. As redes não são estáticas; elas mudam e se adaptam à medida que novos nós e links são adicionados ou removidos. Essa dinâmica pode ser entendida utilizando diferentes modelos teóricos, permitindo aos pesquisadores prever estados futuros das redes e compreender os mecanismos subjacentes que impulsionam as mudanças. Esses modelos podem descrever como um estado ou informação se propaga de nó para nó. Dois exemplos clássicos são os modelos de epidemia, como o SIR (Suscetível-Infetado-Recuperado), e os modelos de difusão de informação, como o Modelo de Limiar Linear:

- **Modelo SIR (Suscetível-Infetado-Recuperado):** usado para modelar a propagação de doenças; nesse modelo, cada nó pode estar em um de três estados. A dinâmica pode ser descrita por um sistema de equações diferenciais (em uma aproximação de campo médio):

$$\begin{aligned}\frac{ds}{dt} &= -\beta \langle k \rangle s(t) i(t) \\ \frac{di}{dt} &= \beta \langle k \rangle s(t) i(t) - \delta i(t) \\ \frac{dr}{dt} &= \delta i(t)\end{aligned}$$

Aqui, $s(t)$, $i(t)$ e $r(t)$ são as frações de nós *suscetíveis*, *infetados* e *recuperados*, respectivamente. β é a taxa de transmissão, δ é a taxa de recuperação e $\langle k \rangle$ é o grau médio da rede. A condição para uma epidemia ocorrer é que o número básico de reprodução, $R_0 = \frac{\beta \langle k \rangle}{\delta}$, seja maior do que 1.

- **Modelo de Limiar Linear (*Linear Threshold Model*):** usado para modelar a difusão de inovações ou comportamentos. Cada nó v possui um limiar de ativação $\theta_v \in [0, 1]$. Um nó inativo v torna-se ativo no instante $t + 1$ se a soma das influências de seus vizinhos já ativos, ponderada por pesos b_{uv} , atinge seu limiar:

$$\sum_{u \in \text{vizinhos\_ativos}(v)} b_{uv} \geq \theta_v$$

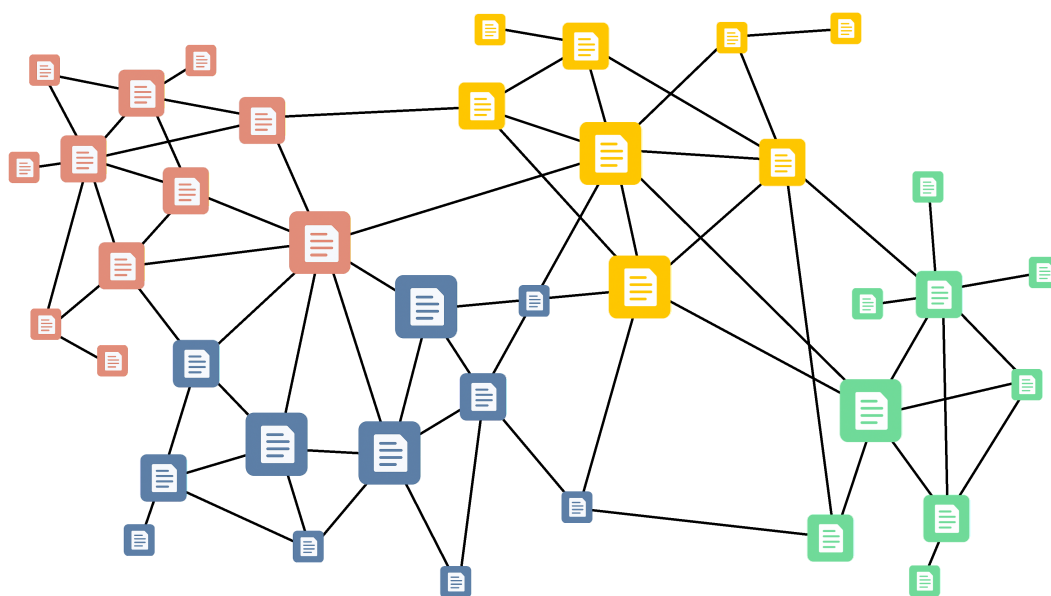
onde $\sum_u b_{uv} \leq 1$. O processo continua até que nenhum novo nó possa ser ativado.



18.3 Detecção de comunidades

No contexto do PLN, o conceito de “comunidade” estabelece um elo essencial entre a estrutura da rede e o significado linguístico. Na prática, uma comunidade identifica um agrupamento temático coeso ou um campo semântico em um *corpus*. De fato, em redes de linguagem, principalmente em redes semânticas (Steyvers; Tenenbaum, 2005), é comum observar grupos de nós fortemente conectados entre si e menos conectados com elementos externos ao grupo. No contexto de redes complexas, tais grupos são denominados comunidades. A Figura 18.5 ilustra uma rede hipotética e sua organização em comunidades, indicadas pelas diferentes cores. Em tarefas de processamento de linguagem e em diversas outras aplicações que envolvem a análise de redes, é comum a realização de uma etapa de identificação de comunidades.

Figura 18.5: Representação visual das comunidades detectadas em uma rede não direcionada.



Considere, por exemplo, uma rede construída a partir de um grande conjunto de notícias, em que os nós correspondem a palavras e as arestas conectam palavras que aparecem frequentemente juntas. Nessa rede, esperaríamos que palavras como “futebol”, “campeonato”, “jogador” e “partida” formassem um agrupamento muito denso, representando o tópico “Esportes” (ou especificamente “Futebol”). Da mesma forma, palavras como “governo”, “eleição”, “congresso” e “partido” formariam outra comunidade (“Política”). Identificar essas estruturas pode ser útil para revelar a organização temática latente do texto e constitui uma etapa crucial para tarefas como a modelagem de tópicos (em que comunidades podem indicar tópicos), a sumarização (ajudando a identificar os conceitos centrais de um texto) ou a desambiguação de palavras (entendendo os diferentes contextos em que uma palavra é usada).

Muitos algoritmos são usados para identificar comunidades em redes complexas, em que se busca maximizar a coesão entre os vértices de um mesmo grupo. Esses algoritmos são aplicados em vários domínios além do PLN, incluindo análise de redes sociais, biologia,

detecção de fraudes e telecomunicações. Métodos proeminentes incluem o *Algoritmo de Propagação de Rótulos* (LPA, *Label Propagation Algorithm*) (Raghavan et al., 2007); o algoritmo *Infomap* (Rosvall; Bergstrom, 2008); o algoritmo *Girvan-Newman* (Newman; Girvan, 2004), e o algoritmo *Louvain* (Blondel et al., 2008).

Eles variam significativamente quanto à abordagem (Palacio-Niño; Berzal, 2024) (Singh; Garg, 2025). O algoritmo de *Girvan-Newman* (Newman; Girvan, 2004), um dos pioneiros, opera pela remoção iterativa de arestas: ele calcula a centralidade de intermediação das arestas e remove aquelas com maior centralidade até que a rede seja segmentada em comunidades distintas. Essa abordagem enfatiza a importância da conectividade das arestas. O LPA também adota uma estratégia simples, atribuindo rótulos a nós com base no voto da maioria de seus vizinhos. Inicialmente, cada nó recebe um rótulo único e, em cada iteração, os rótulos dos nós são atualizados para refletir o rótulo mais frequente entre seus vizinhos. Esse processo iterativo continua até que os rótulos se estabilizem, resultando em comunidades distintas. O LPA é eficiente e escalável, adequado para grandes redes. No entanto, seus resultados podem ser sensíveis à atribuição dos rótulos iniciais e à ordem de processamento dos nós, muitas vezes exigindo múltiplas execuções para alcançar resultados consistentes. Uma abordagem de consenso entre partições distintas pode também ser necessária para melhorar a qualidade das partições (Lancichinetti; Fortunato, 2012).

Já o algoritmo *Infomap* adota princípios da teoria da informação e passeios aleatórios. Um passeio aleatório (*random walk*) é um processo estocástico que descreve a movimentação de um “agente” que se desloca de nó em nó em uma rede seguindo regras probabilísticas. O algoritmo busca minimizar o comprimento do caminho de um passeador aleatório através da rede, capturando o fluxo de informação. Ao conceituar a rede como um mapa, o *Infomap* identifica agrupamentos de nós que são frequentemente visitados pelo passeador aleatório. Este algoritmo é reconhecido por sua robustez e estabilidade, sendo uma opção confiável quando for essencial entender o fluxo de informação na rede.

O *Louvain* é um algoritmo heurístico introduzido por Blondel et al. (Blondel et al., 2008) para detecção de comunidades em grandes redes. O algoritmo otimiza uma métrica de modularidade, uma função de qualidade que mede quão bem uma partição de uma rede se para conexões densas intra-comunidade de conexões esparsas inter-comunidade. Ele opera em duas fases principais que são repetidas iterativamente: (i) Otimização local de modularidade, em que cada nó é inicialmente atribuído à sua própria comunidade, e nós são movidos para comunidades vizinhas se isso resultar em um ganho de modularidade; e (ii) Agregação de comunidades, a partir do momento que não há ganhos de modularidade adicional no nível local, cada comunidade detectada é colapsada em um único nó, produzindo uma “super-rede” menor. A primeira fase é então reaplicada. Esse processo hierárquico e multinível prossegue até que nenhuma melhoria na modularidade possa ser alcançada. O algoritmo é eficiente, mostrando-se adequado para redes muito grandes, com milhões de nós, e hierárquico, produzindo uma hierarquia de comunidades em diferentes níveis de resolução. Também é estocástico (i.e., não determinístico), uma vez que a ordem em que os nós são processados ao decidir mover um nó para uma comunidade vizinha afeta o resultado. Mais recentemente, uma extensão do *Louvain* foi criada de forma a garantir que as comunidades encontradas sejam comunidades conectadas, além de aumentar a velocidade do método (Traag et al., 2019).

A despeito dos avanços significativos, a detecção de comunidades segue com desafios em aberto. Uma das questões fundamentais é a ausência de métodos universalmente aplicáveis a diferentes estruturas de rede. Os algoritmos tradicionais geralmente pressupõem que as comunidades são grupos distintos e densos de nós, o que pode levar a imprecisões

em cenários em que as comunidades se sobrepõem, i.e., existem nós que pertencem a múltiplas comunidades. Além disso, é difícil avaliar objetivamente a qualidade do resultado, em vista da ausência de uma única métrica aplicável aos diferentes métodos. Métricas como o *Índice Rand Ajustado* (ARI), o *Índice Omega* e a *Modularidade* não capturam adequadamente a complexidade de estruturas de comunidades sobrepostas. Assim, muitas vezes é necessário adaptar os métodos de avaliação às características específicas das redes e comunidades analisadas. Outro desafio é identificar corretamente a granularidade (ou o nível de resolução) das comunidades detectadas, i.e., se estamos buscando identificar numerosas comunidades menores ou identificar um número menor de comunidades, mas maiores. A escolha afeta a interpretabilidade e a usabilidade da estrutura de comunidades obtida. Além disso, a análise das propriedades estruturais locais, que poderia aprimorar o processo de detecção, é muitas vezes limitada pela escalabilidade dos algoritmos existentes e pela ausência de dados de referência (*ground truth*) para validação.

A complexidade e a variabilidade inerentes às redes ressaltam a necessidade de abordagens customizadas, e pesquisas seguem na busca por ampliar a escalabilidade, o desempenho e a aplicabilidade dos algoritmos. Tendências incluem a integração de técnicas de aprendizado de máquina, o desenvolvimento de métodos de detecção de comunidades dinâmicas para análise em tempo real e algoritmos para lidar com comunidades sobrepostas.

18.4 Aplicações de redes complexas em PLN para o português

O desenvolvimento de aplicações de redes complexas em PLN tem tido participação significativa de pesquisadores brasileiros. A título de ilustração, dos cerca de 100 artigos indexados na *Web of Science* obtidos em uma busca expressa como (“natural language processing” and “complex network*”), realizada em setembro de 2025, aproximadamente um quarto são de autores brasileiros. Entretanto, essa busca não conseguiu capturar um conjunto representativo de artigos na área, como percebemos ao fazer um levantamento das contribuições de brasileiros da área. Possivelmente ocorreu o mesmo com autores de outros países. Aparentemente, a área está sub-representada na *Web of Science* e, portanto, o número real de artigos deve ser consideravelmente maior, provavelmente da ordem de algumas centenas. De todo modo, é notável a porcentagem de trabalhos brasileiros no tópico.

Em vista da grande flexibilidade das redes complexas como modelo de representação, sua diversidade de aplicações é enorme. A vantagem dessa abordagem é explorar a topologia da rede textual, que é complementar às análises de conteúdo. Por exemplo, em tarefas de sumarização extrativa, nos moldes das discutidas no Capítulo 5, a centralidade de um nó (representando uma sentença) pode indicar sua importância para o resumo, independentemente do peso semântico de suas palavras isoladas. Em análise de sentimentos, por exemplo, em textos de postagens em redes sociais (ver Capítulo 16), a forma como os conceitos positivos e negativos se agrupam na rede pode revelar a polaridade de um texto de opinião com mais robustez. Na avaliação de traduções (ver Capítulo 4), métricas de rede mostraram-se capazes de distinguir a qualidade e a fluidez de textos traduzidos por máquinas ou por humanos.

Apresentamos, nesta seção, um breve resumo de algumas contribuições de autores brasileiros, que foram pioneiros em demonstrar como essas métricas estruturais capturam aspectos complexos da qualidade textual, da simplificação e até mesmo do diagnóstico médico baseado em transcrições de fala.

Parcela significativa das contribuições descreve aplicações que abordam a classificação



de textos modelados como redes complexas. A avaliação da qualidade de redações é uma das aplicações de interesse (ver Capítulo 7). Um dos primeiros artigos a utilizar redes complexas em PLN (Antiqueira et al., 2007) identificou uma alta correlação de métricas de redes com a qualidade dos textos em redações de vestibular. As redações foram modeladas como redes de coocorrência em que os nós representam as palavras. As notas atribuídas por professores aos itens de coesão/coerência, correção gramatical e adequação ao tema diminuíram com o aumento dos valores de métricas como grau de saída, coeficiente de aglomeração e desvio da dependência linear no crescimento da rede. A maior correlação foi observada para o quesito coesão/coerência, indicando que as métricas da rede permitiam capturar esse aspecto qualitativo do texto.

Como um exemplo, considere o seguinte trecho de texto que apresenta boa coesão: *O pesquisador instalou gravadores na floresta. Os sons captados revelaram a presença de aves raras.* Nesse caso, uma rede de coocorrência de palavras revelaria conexões claras entre termos como *pesquisador*, *gravadores*, *floresta*, *sons* e *aves*, formando uma estrutura interligada com caminhos curtos entre os nós. A palavra *sons*, por exemplo, funcionaria como um elo entre a ação de gravar e a descoberta das aves, contribuindo para a coesão semântica do texto. Agora, observe um texto menos coeso: *O pesquisador instalou gravadores na floresta. A cidade estava movimentada com o festival de música.* Embora ambas as frases sejam gramaticalmente corretas, elas pertencem a contextos distintos e não compartilham elementos semânticos relevantes. A rede de coocorrência resultante apresentaria dois componentes desconexos, refletindo a ausência de continuidade temática. Métricas como o coeficiente de aglomeração reduzido ou um maior diâmetro da rede poderiam ser utilizadas para indicar essa quebra de coesão.

A topologia e a dinâmica de redes revelam características estruturais de um texto, o que permite avaliar questões linguísticas inerentes à tradução manual e automática (abordada no Capítulo 4). Amancio et al. (Amancio et al., 2011a) mostraram que é possível distinguir diferentes qualidades de traduções geradas por ferramentas de tradução automática de suas correspondentes manuais utilizando métricas como graus de entrada (ID) e de saída (OD), coeficiente de agrupamento (CC) e caminhos mínimos (SP). Os autores verificaram, por exemplo, que em redes de traduções automáticas o OD médio excede consistentemente os valores obtidos para traduções manuais, e que os valores de CC dos textos originais não são preservados em traduções manuais, mas o são em boas traduções automáticas. Isso provavelmente reflete os rearranjos textuais realizados por humanos durante a tradução manual.

Considere a frase original em inglês: *The biologist recorded bird songs at sunrise in the rainforest.* Essa sentença formaria uma rede de coocorrência coesa, conectando termos como *biologist*, *recorded*, *bird songs*, *sunrise*, e *rainforest*, todos relacionados ao contexto de pesquisa ecológica. Agora, compare duas possíveis traduções para o português:

- Tradução automática: O biólogo registrou sons de pássaros na manhã na floresta.
- Tradução humana: O biólogo gravou cantos de aves ao amanhecer na floresta tropical.

No exemplo, a tradução automática gerou estruturas mais genéricas e menos conectadas semanticamente. Termos como *registrou*, *sons*, *manhã* e *floresta* são corretos, mas formam uma rede com menor densidade e maior diâmetro, refletindo uma coesão temática mais fraca. Já a tradução humana preservou nuances do original, como *cantos de aves*



e *amanhecer*, e introduziu termos mais específicos, como *floresta tropical*, resultando em uma rede mais compacta e interligada.

Ao comparar diferentes redes, métricas como o coeficiente de aglomeração, a densidade de conexões e o número de componentes podem indicar maior fidelidade semântica e coesão na tradução humana. Esse tipo de análise é especialmente útil para avaliar sistemas de tradução automática anteriores à era dos modelos que empregam aprendizado de máquina, que frequentemente apresentavam limitações na preservação de contexto e fluidez textual. O estudo citado (Amancio et al., 2011a) mostrou que as redes de palavras desses dois tipos de texto, ou seja, gerados por traduções automáticas ou manuais, possuem “assinaturas” topológicas distintas. Por exemplo, métricas como o grau médio das palavras e o coeficiente de agrupamento diferem. Isso permitiu treinar classificadores para automaticamente distinguir traduções humanas de traduções automáticas de baixa qualidade.

Ressalte-se que o artigo em questão foi publicado há mais de 15 anos, quando o desempenho dos tradutores automáticos era muito inferior ao atual. Abordagem similar foi aplicada para avaliar a qualidade da tradução automática, inclusive comparando traduções do português para o espanhol e para o inglês (Amancio et al., 2011a). As métricas de redes foram usadas como entrada em algoritmos de aprendizado de máquina que conseguiram distinguir textos oriundos de tradução humana, tradução automática de alta qualidade e tradução automática de baixa qualidade. Um resultado importante foi que a acurácia na classificação aumentou quando foram consideradas métricas de diferentes níveis hierárquicos na rede. Concluiu-se que a possível captura de um contexto mais amplo com os níveis hierárquicos pode ser útil para aumentar a acurácia de classificação com aprendizado de máquina (Amancio et al., 2011a). É interessante fazer uma analogia com o que se faz hoje com aprendizado de máquina nos modelos de língua de larga escala (LLMs, de *large language models* em inglês), em que os *embeddings* e o mecanismo de atenção fornecem contexto amplo, como discutido na Seção 4.2.6.

Três artigos abordam a adequação de redes complexas para tarefas de sumarização extrativa em português (discutidas no Capítulo 5 deste livro). No primeiro, foram empregadas métricas de redes complexas para selecionar sentenças em um resumo extrativo (Antiqueira et al., 2009). A rede que representa um texto era constituída por nós correspondentes às sentenças, enquanto arestas conectavam sentenças que compartilhavam substantivos considerados significativos. Um conjunto de 14 sumarizadores foi produzido empregando diferentes métricas, como o grau dos nós e o comprimento de caminhos mínimos. Algumas versões apresentaram desempenho superior ao de sumarizadores que não empregavam conhecimento linguístico profundo. Em alguns casos, os resultados foram comparáveis aos dos sumarizadores de ponta baseados em recursos linguísticos de alto custo.

Para ilustrar a abordagem de sumarização baseada em redes de coocorrência semântica, considere o seguinte minitexto composto por três sentenças:

- **S1:** A inteligência artificial avança rapidamente no Brasil.
- **S2:** O avanço da inteligência artificial impacta diversos setores da sociedade.
- **S3:** O clima no Brasil é predominantemente tropical.

O modelo constrói uma rede em que os nós representam sentenças e as arestas conectam sentenças que compartilham conceitos ou termos relevantes (por exemplo, *inteligência artificial*, *Brasil*, *avanço*). Nesse cenário:



- **S1** compartilha os termos *inteligência artificial* e *Brasil* com **S2** e **S3**, respectivamente.
- **S2** conecta-se a **S1** por meio dos termos *inteligência artificial* e *avanço*.
- **S3** conecta-se a **S1** por meio do termo *Brasil*, mas não compartilha termos com **S2**.

A estrutura resultante é uma rede em que **S1** atua como um nó central, com maior grau de conectividade. De acordo com métricas de centralidade (como *centralidade de grau* ou *PageRank*), **S1** seria considerada a sentença mais representativa do conteúdo global do texto e, portanto, uma forte candidata à composição de um resumo extrativo. Esse tipo de abordagem é particularmente útil para sumarização automática, pois permite identificar sentenças que funcionam como “pontes semânticas” entre diferentes partes do texto, mesmo em *corpora* maiores e mais complexos.

Em outro trabalho (Tohalino; Amancio, 2018), redes também com sentenças representadas por nós foram usadas para identificar as sentenças mais relevantes para sumarização em vários documentos simultaneamente, em português e em inglês. No terceiro artigo (Amancio et al., 2012a), as redes eram formadas por coocorrência de palavras, ou seja, os nós eram constituídos por palavras em português. Mostrou-se que a incorporação de conhecimento linguístico foi capaz de melhorar, de maneira modesta, o desempenho do sumarizador automático. A conclusão principal foi quanto à utilidade de utilizar métricas de redes complexas para a tarefa de sumarização.

A simplificação de textos é essencial para garantir acessibilidade a leitores com habilidade de leitura limitada, o que tem sido objeto de pesquisa para o português do Brasil, como discutido no Capítulo 6 deste livro. Um desafio é identificar o nível de simplificação adequado para uma audiência específica. Em um trabalho com textos representados por redes de coocorrência, verificou-se que a regularidade topológica se correlaciona negativamente com a complexidade textual (Amancio et al., 2012b). Além disso, a distância entre conceitos, representados como nós, tende a diminuir em textos menos complexos. As métricas das redes complexas foram tratadas com técnicas multivariadas de reconhecimento de padrões, o que permitiu distinguir entre textos originais e suas versões simplificadas. Para cada texto original, duas versões simplificadas foram geradas manualmente com um número crescente de operações de simplificação. Como esperado, a distinção foi mais fácil para as versões fortemente simplificadas, em que as métricas mais relevantes foram força do nó, caminhos mínimos e diversidade. Além disso, a discriminação de textos complexos foi aprimorada com métricas hierárquicas mais altas da rede, apontando, assim, para a utilidade de considerar contextos mais amplos em torno dos conceitos.

A distinção de textos com opiniões positivas ou negativas sobre um tema (problema abordado no Capítulo 16) também pode ser feita classificando-se redes complexas que representam textos. No artigo de Amancio et al. (Amancio et al., 2011b), foram analisados pares de artigos de opinião publicados em um jornal de grande circulação, em que articulistas argumentavam contra ou a favor de um determinado tema. A distinção foi possível empregando várias métricas, incluindo graus, coeficiente de agrupamento, caminhos mínimos, eficiência global, centralidade de proximidade e acessibilidade.

A título de exemplo, considere duas resenhas curtas de um filme, uma positiva e outra negativa:

- Resenha positiva: O filme é ótimo. As atuações são ótimas e o roteiro também é ótimo.



- Resenha negativa: O filme é lento. As atuações são previsíveis e o roteiro é muito fraco.

A rede de coocorrência no primeiro caso, da resenha positiva, seria pequena e altamente conectada. O termo *ótimo* aparece em todas as sentenças e se conecta diretamente a *filme*, *atuação* e *roteiro*, formando um nó central (*hub*) com grau elevado. Essa estrutura indica forte polarização positiva e alta coesão semântica, o que facilita a detecção automática do sentimento positivo. No caso da resenha negativa, o resultado seria uma rede mais esparsa, com termos como *lento*, *previsível* e *fraco* distribuídos entre diferentes aspectos do filme. Não há um único termo dominante que conecte todos os elementos, resultando em uma rede com vários nós periféricos e menor densidade. Essa dispersão semântica reflete uma crítica fragmentada e facilita a identificação do sentimento negativo. Ao comparar essas redes, métricas como grau médio, coeficiente de aglomeração, diâmetro da rede e centralidade de termos afetivos podem ser utilizadas para inferir automaticamente o sentimento predominante. Redes densas com *hubs* positivos tendem a indicar avaliações favoráveis, enquanto redes dispersas com múltiplos termos negativos sugerem críticas desfavoráveis.

O estudo (Amancio et al., 2011b) mostrou que essa diferença na estrutura da rede, capturada por métricas como o coeficiente de agrupamento e os caminhos mínimos, é suficiente para classificar automaticamente a polaridade do texto. O conjunto de dados multidimensional foi mapeado em um espaço bidimensional por meio da análise de componentes principais. A distinção foi quantificada utilizando algoritmos de aprendizado de máquina, com acerto de 84,4% na discriminação automática das opiniões negativas.

Há muitas aplicações de PLN em saúde, como discutido nos Capítulos 9 e 10 deste livro. O diagnóstico médico é essencialmente uma tarefa de classificação, razão pela qual atualmente se empregam diferentes algoritmos de aprendizado de máquina para a análise de imagens médicas e exames clínicos. Esse tipo de aplicação pode também ser feito com texto, como foi o caso da análise de textos transcritos a partir de falas de pacientes e voluntários para diagnosticar o comprometimento cognitivo leve (CCL), que normalmente antecede os sintomas de demências (Santos et al., 2017b). Como os textos eram curtos, as redes complexas foram enriquecidas utilizando-se de *embeddings* de palavras. Ao enriquecer os nós (palavras) com *embeddings*, o modelo consegue capturar nuances semânticas que, combinadas com as métricas da rede, ajudam o classificador a detectar o comprometimento cognitivo. A maior acurácia na classificação binária do CCL foi obtida com o algoritmo Máquina de Vetores de Suporte (*Support Vector Machine*) (Santos et al., 2017b).

18.5 Exemplo de aplicação: Panorama de PLN em língua portuguesa

Estabelecer um panorama para qualquer área de pesquisa exige hoje um grande esforço, em vista do enorme número de publicações. Uma estratégia introduzida há alguns anos (Silva et al., 2016) para obter tais panoramas consiste em combinar redes complexas com PLN, partindo-se de redes de citações ou de similaridade formadas com artigos extraídos de buscas em repositórios como a *Web of Science* ou o *OpenAlex*. Empregando algoritmos de identificação de comunidades, determinam-se as comunidades (agrupamentos) correspondentes a tópicos ou subtópicos da área sob estudo. A análise da rede, tanto em termos de seu formato e topologia quanto do conteúdo das comunidades, permite obter uma visão geral da área. Nesta seção, descrevemos os resultados de empregar essa estratégia para a



área de PLN em português. A metodologia é detalhada nas subseções 18.5.1 e 18.5.2, e os resultados são apresentados na subseção 18.5.3.

18.5.1 Redes de citações e similaridade

Uma rede de citações pode ser obtida a partir de um *corpus* de artigos, estabelecendo uma conexão entre dois artigos do *corpus* (representados como nós da rede) quando existe uma citação de um para o outro. Essa conexão é, em princípio, direcionada, i.e., sai do nó que representa o artigo que cita para o nó que representa o artigo citado. Entretanto, há situações em que se pode optar por ignorar o sentido das arestas, caso o objetivo seja simplesmente representar um relacionamento desse tipo entre os dois artigos.

Se o *corpus* é representativo de uma área de conhecimento, ou de um tópico de pesquisa, a rede de citações permite capturar informações sobre a estrutura do conhecimento nessa área ou tópico. É possível identificar artigos altamente influentes (muito citados), grupos temáticos, artigos que atuam como pontes entre diferentes grupos temáticos, e como ideias se propagaram ao longo do tempo. Artigos centrais muitas vezes são precursores de novas subáreas ou pontos de referência para a comunidade de pesquisa representada na rede. Já artigos que ligam grupos temáticos podem indicar trabalhos com alto grau de interdisciplinaridade, pelo potencial de aproximar comunidades de pesquisa que antes não dialogavam entre si. Esse tipo de análise permite responder perguntas como: quais artigos deram origem a um novo tema de pesquisa? Que áreas e subáreas estão crescendo rapidamente em número de citações? Existe uma fragmentação em comunidades?

As redes de citações evidenciam os relacionamentos entre os artigos, mas não capturam necessariamente o seu conteúdo semântico. Dois artigos podem abordar temas próximos ou relacionados sem que haja uma relação de citação entre eles. Para explorar relações de natureza semântica, pode-se recorrer às redes de similaridade textual. Nessa abordagem, uma conexão entre dois artigos representados como nós da rede é estabelecida quando seus conteúdos são considerados semanticamente próximos, ou seja, foi detectada alguma similaridade de conteúdo. Para quantificar essa proximidade semântica, utiliza-se uma representação vetorial conhecida como *embedding* (Pennington et al., 2014). Para mais informações sobre similaridade textual e *embeddings* sugerimos ao leitor consultar o Capítulo Modelos de linguagem deste livro.

Existem diversos modelos para gerar *embeddings*, que podem vetorizar desde palavras isoladas até textos completos. Na abordagem apresentada na Seção 18.5.3, cada resumo é representado por um vetor numérico usando o modelo *SentenceTransformer*, que pode mapear textos inteiros em um espaço multidimensional. A proximidade semântica dos textos é determinada pelo cálculo de similaridade do cosseno entre todos os pares de vetores. A expressão matemática para essa similaridade é dada por:

$$S_{cos}(A, B) = \frac{A \cdot B}{\|A\| \|B\|}$$

onde $A \cdot B$ é o produto escalar entre os vetores A e B , e $\|A\|$ e $\|B\|$ indicam as respectivas normas dos vetores. Esse produto escalar varia entre -1 e 1 ; no entanto, no contexto de *embeddings* textuais, considera-se apenas o módulo. Assim, uma similaridade próxima de 0 indica que os dois textos são totalmente distintos, enquanto uma similaridade próxima a 1 revela que os textos são semanticamente próximos.

Para a construção da rede de similaridade nessa abordagem, considera-se ainda um limiar de similaridade, que estabelece o valor mínimo necessário para que uma conexão seja

adicionada à rede. Ou seja, utilizando um limiar de 70% ($S_{cos}(A, B) = 0.7$), apenas os pares de artigos cuja similaridade exceder este valor estarão conectados por uma aresta. Essa etapa evita que se crie uma rede excessivamente densa e garante que o modelo gerado ofereça uma visão estruturada da proximidade temática entre os artigos, servindo de base para a identificação de padrões na literatura. Artigos que compartilham temas centrais tendem a compor agrupamentos temáticos, revelando subáreas de estudo ou tópicos emergentes. Conexões entre agrupamentos podem indicar tópicos interdisciplinares ou tendências de convergência de temas de pesquisa. Essas redes permitem identificar artigos que abordam temas semelhantes, além de evidenciar padrões ocultos não evidenciados nas relações de citação. As redes de similaridade são então complementares às redes de citações. São particularmente úteis para pequenos grupos de artigos, como em uma área emergente na qual ainda não se identificam trabalhos centrais com alto número de citações.

18.5.2 Identificação de tópicos em redes de pesquisa

O método introduzido por Silva et al. (Silva et al., 2016) estabelece uma estratégia para identificar os principais temas (tópicos) associados a uma área de pesquisa, dada uma rede criada a partir de um *corpus* textual de artigos científicos representativo da área. A estratégia associa um algoritmo de detecção de comunidades na rede com a identificação dos tópicos abordados pelos artigos pertencentes às diferentes comunidades. Assim, comunidades são caracterizadas por conjuntos de termos (palavras-chave - unigramas e bigramas) extraídos do *corpus* a partir da quantificação da importância semântica dos termos, no contexto da sua ocorrência naquela comunidade. Em particular, os termos elencados são extraídos do título e resumo dos artigos que compõem a rede (seja ela de citação ou de similaridade), cujos conteúdos contêm informações suficientes para uma boa caracterização dos termos mais representativos.

A importância dos termos na rede é quantificada por meio de dois índices de frequência relativa. Inicialmente, é determinada a frequência de um termo dentro de sua própria comunidade e, em seguida, sua frequência no restante da rede. Por exemplo, para determinar a frequência de um termo (w) em uma comunidade (α), conta-se o total de artigos $n_\alpha(w)$ dessa comunidade que contêm o termo (w). Com isso, a frequência do termo relativa dentro da comunidade $F_\alpha^{in}(w)$ é:

$$F_\alpha^{in}(w) = \frac{n_\alpha(w)}{|\alpha|}$$

em que $|\alpha|$ é o número de artigos pertencentes à comunidade (α). Analogamente, definimos a frequência relativa fora da comunidade (α) como:

$$F_\alpha^{out}(w) = \sum_{\gamma \neq \alpha} \frac{n_\gamma(w)}{N - |\alpha|}$$

Neste caso, (γ) representa uma comunidade qualquer, diferente de (α), e N é o número total de artigos da rede. Ou seja, $F_\alpha^{out}(w)$ se refere à frequência relativa do termo (w) fora da comunidade (α).

Com as relações de frequência interna e externa à comunidade (α) estabelecidas, a importância de um termo $I(w)$ é definida como a maior diferença entre $F_\alpha^{in}(w)$ e $F_\alpha^{out}(w)$, ou seja:



$$I(w) = \max_{\alpha} [F_{\alpha}^{in}(w) - F_{\alpha}^{out}(w)]$$

Os termos-chave são agrupados com base na distância topológica média entre os artigos que os contêm, ou seja, é utilizado o comprimento médio do caminho mais curto $\langle l \rangle_{uv}$ entre um par de termos-chave (u, v) . O algoritmo inicialmente determina a menor distância (l_{ij}) entre cada par de artigos (i, j) na rede e então, para cada par de termos-chave (u, v) , calcula a média dos menores caminhos entre todos os pares de resumos (A_i, A_j) dos artigos cujos termos-chave u e v estão presentes. A representação matemática desses conceitos é dada na equação abaixo:

$$\langle l \rangle_{uv} = \sum_{(u,v) \in (A_i \times A_j)} \frac{l_{ij}}{|(u,v) \in (A_i \times A_j)|}$$

Em termos gerais, significa que conjuntos de termos-chave são agrupados segundo a distância topológica média entre eles. Ressalte-se que os unigramas e bigramas são classificados de acordo com a mesma medida. Quando um bigrama possui alto $I(w)$, há grande probabilidade de que seus unigramas compostos também tenham alto $I(w)$, gerando redundância. Esse problema é resolvido com a remoção dos unigramas do conjunto de palavras-chave que fizerem parte de qualquer outro bigrama no conjunto. Essa prática proporciona uma maior prioridade a termos-chave (bigramas) mais específicos. Ao final, o método gera uma rede complexa composta por conjuntos densamente conectados (agrupamentos), que são caracterizados por seus termos-chave com maior índice de importância.

18.5.3 Pesquisa em PLN para o português

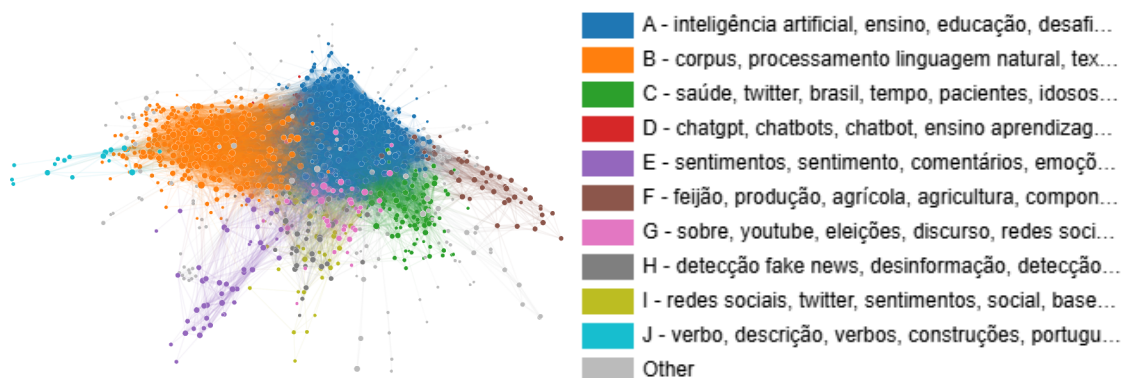
O desafio inicial consistiu em localizar as contribuições em PLN em português na literatura. Buscamos identificar artigos científicos publicados sobre o PLN para a língua portuguesa, escritos nas línguas portuguesa ou inglesa. Bases de dados como *Web of Science* e *OpenAlex* incluem textos em português, mas as estratégias de busca podem ser falhas, ou pelo menos mais difíceis de otimizar do que para textos em inglês. Há também o problema da menor visibilidade dos fóruns de publicação. Artigos em fóruns relevantes, como o PROPOR, podem não estar indexados. A baixa visibilidade representa, obviamente, uma limitação para o estudo proposto, assim como para a valorização das contribuições científicas de países lusófonos em PLN. Diante dessas limitações, não há como garantir que o panorama apresentado a seguir seja totalmente fiel ao panorama da literatura produzida para o PLN em português, ao contrário do que se pode afirmar para estudos anteriores em outros tópicos (Oliveira; Oliveira, 2025; Oliveira et al., 2025). Nestes últimos trabalhos não havia dúvidas de que os artigos mais relevantes para as respectivas áreas foram recuperados. Críticas sobre o panorama apresentado serão muito bem-vindas para garantir estudos mais precisos no futuro e para que a comunidade do PLN em português busque estratégias para ampliar a visibilidade dos trabalhos na área.

O número de artigos que abordam PLN para a língua portuguesa encontrados em diferentes buscas, na *Web of Science* e *OpenAlex*, foi sempre inferior a 5.000. Esse número reduzido inviabiliza a construção de redes de citações, que seriam esparsamente conectadas. A alternativa foi construir redes de similaridade, como descrito na Seção 18.5.1, para o que também houve algumas dificuldades. Usado em estudos anteriores, o SciBERT (Beltagy et al., 2019) foi treinado com artigos e textos científicos em inglês e mostrou-se inadequado para textos em português. Isso motivou a adoção do *SentenceTransformer*



multilíngue, um modelo BERT capaz de lidar com outras línguas. Apesar de não ser especializado em artigos científicos, foi a opção com melhores resultados. A busca no *OpenAlex* com os termos “PLN” OR “NLP” OR “processamento de linguagem natural” OR “natural language processing” retornou 3.303 artigos, dos quais 1.766 compõem a maior componente conectada da rede de similaridade, construída considerando um limiar de similaridade igual a 70%. A maior componente da rede resultante é exibida na Figura 18.6.

Figura 18.6: Rede de similaridade para publicações sobre Processamento de Linguagem Natural em português, mostrando as principais comunidades temáticas e suas inter-relações.



Link para visualização interativa da rede: [Rede de Similaridade](#)

O Quadro 18.1 lista as 10 maiores comunidades identificadas na rede da Figura 18.6, incluindo o número de artigos por comunidade e uma descrição gerada pelo ChatGPT a partir dos termos-chave associados a cada comunidade. Estas descrições foram geradas com um comando “Faça uma descrição dos tópicos abordados pelo cluster X, com base nas palavras-chave que o caracterizam”. A maior comunidade (A, em azul, com 748 artigos) é dedicada à inteligência artificial, principalmente ao seu uso em diferentes domínios, como o jurídico e o educacional. Uma inspeção (sem valor estatístico) nos títulos dos artigos indica que muitos deles se referem a textos, alguns de fato envolvendo PLN, embora outros talvez não. Mencione-se aqui uma possível dificuldade de identificação de comunidades devido às ambiguidades de nomenclatura. O termo inteligência artificial (IA) é usado em contextos diversos, nem sempre com precisão na terminologia. É comum o uso do termo “aprendizado de máquina” de forma intercambiável com IA, como se fossem sinônimos, ignorando que aprendizado de máquina é subárea de IA. PLN é outra subárea de IA, mas suas comunidades de pesquisadores nem sempre se identificam com grupos de IA que atuam em outras subáreas. Parece, portanto, inevitável que a busca por trabalhos em PLN encontre artigos de IA, mas que não necessariamente abordam PLN. De qualquer forma, é surpreendente que quase metade dos artigos da rede estejam na comunidade A, em que o foco está em usos de IA - e não em PLN.

Já a segunda maior comunidade, B, em laranja, com 377 artigos, é de fato de artigos em PLN. Ela inclui artigos relacionados à construção e organização de *corpora*, classificação automatizada de textos e uso de linguística computacional para tarefas de análise e categorização. Artigos relacionados a aplicações em saúde aparecem na terceira comunidade, C, em verde, com 110 itens. Essa comunidade inclui trabalhos em temas variados, como análise sistemática da literatura em saúde e análise de sentimentos ou opinião relacionada

à saúde. Parte desses artigos pode abordar tópicos de PLN, mas provavelmente há outros que envolvem algum tipo de processamento de texto, não necessariamente PLN. As demais comunidades são consideravelmente menores, com algumas dezenas de artigos. São, em geral, bastante específicas e abordam temas em PLN, com uma possível exceção. Trata-se da comunidade F (em marrom, com 46 artigos), composta por artigos sobre agricultura. Dos poucos artigos efetivamente de PLN que podem ser identificados, destacam-se os de análise de literatura no tópico. A comunidade D (em vermelho, 53 artigos) inclui artigos sobre assistentes de conversação (*chatbots*), principalmente o ChatGPT. São, portanto, artigos recentes. A comunidade E (em roxo, com 47 artigos) é dedicada à análise de sentimentos. Eleições e política são os temas principais na comunidade G (rosa, com 41 artigos), com trabalhos de PLN principalmente dedicados à análise de texto em redes sociais. Outra comunidade típica de PLN é a H, em cinza-escuro, com 38 artigos que versam sobre detecção de notícias falsas ou desinformação. A comunidade I, em verde-claro, com 37 artigos, é vizinha à comunidade E na rede da Figura 18.6, pois também trata de análise de sentimentos, com foco no Twitter. A última comunidade, J, em verde-musgo, com 18 artigos, inclui estudos linguísticos, como análise semântico-discursiva utilizando PLN. Alguns poucos artigos que não foram categorizados como pertencendo a alguma comunidade aparecem em cinza-claro na Figura 18.6.

O formato da rede pode ser inspecionado cuidadosamente a partir do modelo tridimensional interativo acessado no link informado acima e também merece comentários. A rede é dominada por duas comunidades, A e B, que são bastante conectadas entre si. A comunidade J, de estudos linguísticos com e para PLN, é conectada apenas com a comunidade B, que engloba a maioria dos estudos em PLN. Pode-se afirmar, então, que essas duas são comunidades de pesquisadores de fato atuantes em PLN. O argumento se justifica porque o panorama que se depreende da rede da Figura 18.6 talvez permita distinguir trabalhos de grupos de pesquisadores que efetivamente atuam em PLN de trabalhos de pesquisadores não necessariamente identificados com essa área, mas que passaram a atuar em processamento de texto como consequência dos avanços tecnológicos em aprendizado de máquina e, mais recentemente, a IA gerativa. O que se observa é que as demais comunidades são sempre conectadas à comunidade A; seus artigos têm maior similaridade com aqueles classificados mais genericamente como IA do que com PLN. Apenas a comunidade E (em roxo), relacionada à análise de sentimentos, está próxima da comunidade B, embora ainda se mostre mais conectada à comunidade A.

Quadro 18.1: Comunidades na rede de PLN.

Comunidade	Nós	Palavras-chaves	Descrição
A	748	inteligência artificial, ensino, educação, desafios, aprendizado, tecnologias, científica, máquina, papel, ia	Aplicações de inteligência artificial no contexto educacional; tecnologias e estratégias para ensino, aprendizagem e desenvolvimento científico.



B	377	corpus, processamento língua natural, textos, automática, classificação, linguística, português brasileiro, textual, língua, modelos	Processamento de linguagem natural e linguística computacional; análise automática de textos, classificação textual, <i>corpora</i> e modelos linguísticos em português brasileiro.
C	110	saúde, twitter, brasil, tempo, pacientes, idosos, in, pandemia covid-19, tratamento, revisão	Uso de redes sociais para monitoramento de saúde pública; análise da pandemia de COVID-19, impacto em pacientes e idosos, e revisão de tratamentos no Brasil.
D	53	chatgpt, chatbots, chatbot, ensino-aprendizagem, sobre, revisão, ferramenta, auxiliar, assistente, sistemática	Desenvolvimento e aplicação de chatbots e assistentes virtuais; ferramentas para ensino e aprendizagem e revisões sistemáticas de tecnologias educacionais.
E	47	sentimentos, sentimento, comentários, emoções, português, métricas, mídias sociais, mineração, rede, mapeamento	Análise de sentimentos em mídias sociais; mineração de dados textuais, métricas emocionais e mapeamento de interações em português.
F	46	feijão, produção, agrícola, agricultura, componentes, rendimento, produtividade, cultura soja, seleção, gesso	Produção e produtividade agrícola; técnicas de cultivo, seleção de componentes e otimização de rendimento em culturas como feijão e soja.
G	41	sobre, youtube, eleições, discurso, redes sociais, deputados, discursos, ódio, sentimento, twitter	Análise de discursos e sentimentos em redes sociais durante eleições; estudo de conteúdo em YouTube e Twitter, incluindo discursos de deputados e manifestações de ódio.
H	38	detecção fake news, desinformação, detecção automática, notícias falsas, jornalismo, utilizando, produção, covid-19 brasil, language, avaliação	Detecção automática de fake news e desinformação; análise de notícias falsas e produção jornalística durante a pandemia de COVID-19 no Brasil.
I	37	redes sociais, twitter, sentimentos, social, baseado, tweet sobre, eventos, classificação, identificação, caso	Análise de sentimentos e classificação de eventos em redes sociais; identificação de padrões sociais e comportamentais no Twitter.



J	18	verbo, descrição, verbos, construções, português, linguagem natural, processamento, base, verbal, semântico-discursiva	Estudos linguísticos em português; análise de verbos, construções verbais e descrições semântico-discursivas utilizando processamento de linguagem natural.

Uma análise manual permite identificar os tópicos mais relevantes nas diferentes comunidades, mas é inviável para um estudo pormenorizado, em vista do número de artigos da rede (mais de 1.700). É possível, entretanto, buscar auxílio das ferramentas com modelos de língua de larga escala (LLMs, do inglês *large-language models*), que podem processar texto a uma velocidade muito superior à de um ser humano. Em setembro de 2025, não há garantias de que tais ferramentas forneçam análises ou resumos fiéis aos artigos sob estudo, principalmente porque a janela de contexto disponibilizada para um usuário pode não ser suficiente para cobrir todo o *corpus*. Há, entretanto, evidências empíricas de que essas ferramentas conseguem fornecer uma análise superficial adequada. Com essa expectativa, solicitamos ao ChatGPT uma expansão na descrição das comunidades da Figura 18.6.

Com os termos-chave de cada comunidade e suas respectivas descrições iniciais fornecidas pelo ChatGPT, foi solicitado um levantamento mais detalhado dos tópicos apontados por ele. Para isso, foi utilizado o seguinte comando “*Faça uma expansão para a descrição do cluster X, se aprofundando nos tópicos apresentados por ele de acordo com suas palavras-chave e a busca central utilizada para produzir a rede*”. Os textos produzidos, individualmente para cada comunidade, apresentaram um aprofundamento dos tópicos descritos na coluna “Descrição” do Quadro 18.1. Os textos foram editados, por concisão, e suas análises e conclusões estão descritas a seguir.

Já comentamos a partir da análise manual que a maior comunidade, A, foca na aplicação de técnicas de IA e PLN na educação, revelando um movimento crescente de integração de tecnologias inteligentes em processos de ensino e aprendizagem, com destaque para sistemas de tutoria inteligente, plataformas adaptativas, ferramentas de correção automática de textos e *chatbots* educacionais. Os trabalhos também abordam de forma crítica os desafios pedagógicos e técnicos que acompanham essa transformação. Entre eles estão a dificuldade de adaptar modelos de PLN para diferentes contextos culturais, a necessidade de preparar professores para integrar essas ferramentas em suas práticas e os riscos de dependência tecnológica. Além disso, surgem preocupações sobre vieses algorítmicos, transparência na tomada de decisão das máquinas e privacidade dos dados estudantis, questões que exigem debates éticos e regulatórios. A literatura coberta nessa comunidade não se restringe às aplicações imediatas, mas também reflete sobre o impacto estrutural da adoção da IA na educação. Muitos artigos discutem como essas tecnologias estão redefinindo o papel do professor, não como substituto, mas como mediador em um ecossistema de aprendizagem híbrido, onde humanos e máquinas colaboram. Ao mesmo tempo, destacam-se análises empíricas sobre a eficácia dessas ferramentas na melhoria do desempenho acadêmico, no engajamento dos alunos e na inclusão de grupos tradicionalmente marginalizados no acesso ao ensino.

Como já mencionado, a comunidade B contém literatura fundamental de PLN, concentrando-se nos fundamentos linguísticos e computacionais com ênfase em *corpus*, classificação automática e análise textual. Os estudos desse grupo envolvem a construção, curadoria e utilização de *corpora*. Fica evidenciado o papel central dos *corpora* para



treinar, avaliar e comparar algoritmos de PLN em diferentes tarefas, como análise de sentimentos, categorização de documentos e reconhecimento de entidades nomeadas. São abordados também os desafios metodológicos relacionados à anotação, padronização e representatividade dos dados (questões abordadas no Capítulo [Conjunto de dados, dataset e corpus](#) deste livro), uma vez que a qualidade e a diversidade dos *corpus* impactam o desempenho dos modelos. Nesse sentido, surgem discussões sobre métodos automáticos e semiautomáticos de anotação, bem como sobre a integração de abordagens linguísticas tradicionais com técnicas de aprendizado de máquina. Além disso, chama-se a atenção para as particularidades do português, idioma menos contemplado na pesquisa internacional de PLN do que o inglês. Esse foco evidencia um esforço da comunidade científica em reduzir desigualdades linguísticas e promover avanços na adaptação de modelos globais a contextos locais. Os trabalhos também revelam preocupações quanto à dimensão sociotécnica do PLN, discutindo o risco de vieses reproduzidos por *corpora* desbalanceados ou pouco representativos. A literatura aponta ainda para a necessidade de ferramentas linguísticas acessíveis, que possam servir tanto à pesquisa acadêmica quanto a aplicações práticas em educação, saúde e setores industriais.

A comunidade C explora a aplicação do PLN em saúde, com ênfase em redes sociais como o Twitter e em contextos marcados pela pandemia da COVID-19. Os trabalhos desse grupo investigam como técnicas de PLN podem ser empregadas para monitorar, analisar e prever fenômenos relacionados à saúde pública, desde a identificação de sintomas relatados por pacientes até a avaliação de percepções sociais sobre políticas sanitárias. Essa vertente evidencia o papel estratégico do PLN na compreensão de grandes volumes de dados não estruturados, capazes de complementar fontes tradicionais de informação médica e epidemiológica. Os estudos também abordam a utilização do Twitter como fonte de dados em tempo real, permitindo a detecção de surtos, o acompanhamento de discussões sobre vacinas e tratamentos e a análise das emoções e preocupações expressas pela população. O caráter dinâmico e global das redes sociais potencializa o PLN como ferramenta de vigilância em saúde, mas também levanta desafios metodológicos, como a necessidade de filtrar ruídos, lidar com ambiguidade textual e diferenciar informações confiáveis de conteúdos desinformativos. Além disso, a pandemia de COVID-19 aparece em muitos artigos, em um período em que se empregou PLN para analisar notícias relacionadas e comentários em redes sociais. Outro ponto relevante é o impacto da análise automatizada na prática clínica e na formulação de políticas de saúde. Pesquisas avaliam o potencial de modelos de PLN para apoiar diagnósticos, identificar grupos de risco, prever demandas hospitalares e oferecer suporte a revisões sistemáticas de literatura biomédica. Paralelamente, surgem reflexões sobre os limites éticos e legais do uso de dados provenientes de redes sociais, sobretudo em relação à privacidade e ao consentimento dos usuários.

Os *chatbots* são o assunto central da comunidade D, principalmente o ChatGPT. A literatura agrupa estudos sobre o desenvolvimento de agentes conversacionais voltados para suporte pedagógico — desde *chatbots* que oferecem retorno imediato em tarefas escritas até assistentes que guiam o estudante passo a passo em resoluções de problemas. Muitos trabalhos relatam experimentos com plataformas educacionais integradas a *chatbots*, avaliando como esses agentes podem personalizar explicações, fornecer exercícios adaptativos e atuar como tutores suplementares em disciplinas diversas. Alguns trabalhos nessa comunidade combinam revisões sistemáticas com estudos empíricos para testar protótipos em sala de aula ou em ambientes controlados. As avaliações frequentemente medem ganhos de aprendizagem, engajamento, aceitabilidade por parte de alunos e professores, além de métricas de usabilidade e qualidade da conversação (coerência, relevância das respostas,



taxa de intervenção humana). Há também trabalhos técnicos que abordam arquiteturas (agentes baseados em regras vs. modelos gerativos), adaptação para domínios educacionais, técnicas de *prompting* e estratégias para integrar *chatbots* a sistemas de gestão de aprendizagem (para saber mais sobre o tema, sugerimos consultar o Capítulo [ChatGPT, MariTalk e outros agentes de conversação](#) deste livro).

A análise de sentimentos é o tópico dominante nas comunidades E e I, com destaque para dados provenientes de mídias sociais e comentários online em português. Os trabalhos desses grupos exploram como técnicas de PLN podem ser aplicadas para identificar, classificar e quantificar sentimentos positivos, negativos ou neutros, bem como emoções mais complexas, em grandes volumes de dados textuais. Essa linha de pesquisa é relevante para entender o comportamento dos usuários, monitorar percepções públicas sobre produtos, serviços ou eventos, fornecendo subsídios para áreas como publicidade, comunicação e políticas públicas. Os estudos evidenciam o desenvolvimento e a aplicação de métricas específicas para a análise de sentimentos, incluindo métodos baseados em léxicos, aprendizado de máquina e modelos de linguagem pré-treinados. Além disso, os trabalhos abordam questões metodológicas e práticas, incluindo desafios relacionados à ambiguidade textual, ironia, sarcasmo e diversidade linguística no português brasileiro.

A comunidade F pareceu ser a menos aderente ao PLN na análise manual, pois os títulos dos artigos normalmente não indicavam uso de PLN. Entretanto, segundo a descrição do *ChatGPT*, o PLN está presente na análise textual em contextos agrícolas, com ênfase em culturas como feijão e soja. Ainda segundo o *ChatGPT*, a comunidade contém literatura na interface entre tecnologia e prática agrícola, demonstrando como a análise de textos e a extração de conhecimento podem contribuir para avanços concretos na produção de alimentos. A comunidade G é dedicada ao estudo de eleições e políticas, concentrando-se na análise de discursos políticos, eventos eleitorais e interações em redes sociais, principalmente no Twitter e no *YouTube*. Há artigos explorando técnicas de PLN para identificar padrões de comunicação, detectar discursos de ódio, analisar sentimentos e compreender o impacto das mensagens veiculadas por políticos, influenciadores e cidadãos em contextos eleitorais. Essa abordagem permite mapear tanto a propagação de opiniões quanto a dinâmica de engajamento do público em diferentes plataformas digitais. Além disso, investigam estratégias de comunicação política, incluindo a construção de narrativas, o uso de linguagem persuasiva e a disseminação de mensagens polarizadoras.

Artigos sobre detecção de notícias falsas e desinformação predominam na comunidade H, em que há ênfase no uso de PLN para identificar padrões linguísticos, sinais de veracidade e estratégias de manipulação em notícias e conteúdos compartilhados em redes sociais. O objetivo central é desenvolver métodos que permitam avaliar a confiabilidade das informações, detectando conteúdos enganosos e prevenindo a disseminação de desinformação em larga escala. Os estudos nessa comunidade abordam múltiplas abordagens metodológicas, incluindo modelos baseados em aprendizado de máquina supervisionado e não supervisionado, análise de métricas linguísticas, redes semânticas e técnicas de classificação de textos. Além disso, a literatura enfatiza a importância de *corpus* específico para o português brasileiro, permitindo que os algoritmos lidem com variações linguísticas, regionalismos, gírias e construções sintáticas típicas do idioma. Muitas pesquisas também relacionam a detecção de notícias falsas sobre a pandemia de COVID-19, analisando a propagação de informações incorretas e seu impacto sobre comportamentos de saúde e percepção pública.

A pequena comunidade J concentra-se no estudo de verbos, construções verbais e aspectos semântico-discursivos do português. Há ênfase na aplicação de técnicas de PLN para

analisar padrões gramaticais, categorias verbais e relações semânticas, contribuindo para o desenvolvimento de modelos linguísticos mais precisos e adaptados às especificidades do idioma. Também aparecem pesquisas associadas à anotação e classificação de verbos, construção de bases de dados linguísticos e análise de estruturas sintáticas e semânticas em textos. Esses estudos permitem compreender como os verbos e suas construções funcionam em diferentes contextos comunicativos, servindo para tarefas de PLN como análise sintática, geração de linguagem natural e tradução automática.

Para fins de estudo e transparência, foi disponibilizado um arquivo complementar contendo uma lista dos artigos que compõem as dez maiores comunidades da rede de similaridade. O arquivo está organizado em abas que correspondem às comunidades identificadas; cada aba apresenta uma tabela com os títulos e respectivos links de acesso aos artigos no repositório de origem.

Link de acesso: [Arquivo Complementar](#)

18.6 Considerações finais

O sucesso do uso de redes complexas para o PLN parece ser indiscutível, a julgar pela gama de aplicações comprovadas. A expectativa foi confirmada de que é possível capturar características semânticas, sintáticas e mesmo pragmáticas ao representar um texto como uma rede. A literatura da área, entretanto, não é vasta e o número de artigos deve ser da ordem de algumas poucas centenas. Essa produção tem contribuição relevante de autores brasileiros, que estiveram entre os primeiros a aplicar conceitos e métodos de redes complexas em PLN. Embora os trabalhos dos autores brasileiros sejam majoritariamente focados no processamento de textos em inglês, há pelo menos algumas dezenas de artigos que abordam PLN para o português. Neste capítulo, apresentamos um breve resumo de algumas dessas contribuições, em sumarização, desambiguação lexical, tradução automática, análise de sentimentos e diagnóstico médico baseado em texto.

As perspectivas quanto aos usos de redes complexas para aplicações em PLN foram abordadas em um comentário (Amancio, 2014) que enfatizou a necessidade de maior rigor na interpretação dos fatores que contribuem para o sucesso dos modelos de rede. Por outro lado, é interessante fazer uma comparação ou analogia com modelos de língua de larga escala (LLMs), uma vez que muitas das tarefas de PLN realizadas com redes complexas podem atualmente ser feitas utilizando LLMs. Pode-se especular que as informações advindas das características estruturais e dinâmicas das redes representando textos sejam, de alguma maneira, inferidas por meio dos mecanismos de *embedding* e de atenção dos LLMs. A grande capacidade de processamento e a janela de contexto dos LLMs tendem a resultar em desempenho significativamente superior nas tarefas de PLN, em comparação ao uso de redes. De fato, em pelo menos um trabalho ainda não publicado de pesquisadores do Núcleo Interinstitucional de Linguística Computacional (NILC) mostrou-se que o uso de modelos de língua gera resultados consideravelmente melhores. Referimo-nos à capacidade de sistemas inteligentes de estimar a pontuação de redações de ENEM, em que modelos usando BERT (“transformer”) mostraram-se mais eficientes do que o uso de métricas de redes complexas. A comparação não é necessariamente justa, pois o desempenho das redes não foi otimizado variando-se as métricas exaustivamente.

De todo modo, acreditamos que a maioria das tarefas de PLN será mais bem executada com LLMs, o que poderia indicar que não há muito futuro para o uso de redes complexas em PLN. Entretanto, deve-se considerar que as redes continuarão sendo adequadas para verificar modelos e estudar a importância de aspectos linguísticos nas diferentes tarefas.



As informações estruturais capturadas pelos modelos podem contribuir para a melhor compreensão dos processos linguísticos envolvidos, ou seja, garantir mais interpretabilidade para os resultados do processamento de textos. Redes também serão importantes para arquiteturas baseadas em grafos e como fontes de informação em tarefas de visualização exploratória de itens da literatura científica, como no panorama apresentado neste capítulo para o PLN em português.



Referências

- ABDIN, M. et al. **Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone.**, 2024. Disponível em: <<https://arxiv.org/abs/2404.14219>>
- ABONIZIO, H. Q. et al. **Language-Independent Fake News Detection: English, Portuguese, and Spanish Mutual Features.** **Future Internet**, v. 12, n. 5, 2020.
- ABREU, S. C. DE; VIEIRA, R. Relp: Portuguese open relation extraction. **KO KNOWLEDGE ORGANIZATION**, v. 44, n. 3, p. 163–177, 2017.
- ACKEL, A. Abordagens digitais para estudos de Paleografia: desafios, atualidade, desdobramentos. **LaborHistórico**, v. 7, n. 3, p. 100–120, 2021.
- AEJAS, B.; BELHI, A.; BOURAS, A. **Smart Contracts Auto-generation for Supply Chain Contexts.** (F. Noël et al., Eds.)Product Lifecycle Management. PLM in Transition Times: The Place of Humans and Transformative Technologies. **Anais...Cham: Springer Nature Switzerland**, 2023.
- AGGARWAL, C. C. **Machine Learning for Text.** [s.l.] Springer, 2018.
- AGICHTEIN, E.; GRAVANO, L. **Snowball: Extracting relations from large plain-text collections.** Proceedings of the fifth ACM conference on Digital libraries. **Anais...2000.**
- AGNOLONI, T. et al. **Making Italian Parliamentary Records Machine-Actionable: the Construction of the ParlaMint-IT corpus.** Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference. **Anais...Marseille, France: European Language Resources Association**, jun. 2022. Disponível em: <<https://aclanthology.org/2022.parlaclarin-1.17>>
- AHA, D. W.; KIBLER, D.; ALBERT, M. K. **Instance-based learning algorithms.** **Machine Learning**, v. 6, n. 1, p. 37–66, 1 jan. 1991.
- AI@META. **Llama 3 Model Card.** 2024.
- AJAY, H. B.; TILLET, P.; PAGE, E. B. **Analysis of essays by computer (AEC-II).** Storrs, CT: Univeristy of Connecticut, 1973.
- AKBIK, A. et al. **Multilingual information extraction with PolyglotIE.** Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations. **Anais...2016.** Disponível em: <<https://aclanthology.org/C16-2056/>>



ALBERTO, T. C.; LOCHTER, J. V.; ALMEIDA, T. A. Post or Block? Advances in Automatically Filtering Undesired Comments. **Journal of Intelligent and Robotic Systems**, v. 80, p. 245–259, 2015.

ALBUQUERQUE, G. et al. Applying event classification to reveal the Estado da Índia. **Proceedings of the International Conference on the Computational treatment of Portuguese, PROPOR**, a2024.

ALBUQUERQUE, H. et al. **UlyssesNERQ: Expanding Queries from Brazilian Portuguese Legislative Documents through Named Entity Recognition**. (P. Gamallo et al., Eds.) Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1. **Anais...**Santiago de Compostela, Galicia/Spain: Association for Computational Linguistics, mar. b2024. Disponível em: <<https://aclanthology.org/2024.propor-1.35>>

ALBUQUERQUE, H. O. et al. **UlyssesNER-Br: A Corpus of Brazilian Legislative Documents for Named Entity Recognition**. (V. Pinheiro et al., Eds.) Computational Processing of the Portuguese Language. **Anais...**Cham: Springer International Publishing, 2022. Disponível em: <<https://github.com/ulysses-camara/>>

ALBUQUERQUE, H. O. et al. **Named entity recognition: a survey for the portuguese language**. **Procesamiento del Lenguaje Natural**, b2023.

ALBUQUERQUE, H. O. et al. **On the Assessment of Deep Learning Models for Named Entity Recognition of Brazilian Legal Documents**. (N. Moniz et al., Eds.) Progress in Artificial Intelligence. **Anais...**Cham: Springer Nature Switzerland, a2023.

ALBUQUERQUE, H. O. et al. **UlyssesLegalNER-Br: from Legislative to Legal, a comprehensive corpus of Brazilian legal documents for Named Entity Recognition**. 17th Computational Processing of the Portuguese Language (PROPOR 2026). **Anais...**2026.

ALEIXO, P.; PARDO, T. A. S. **Uma Ferramenta Semi-automática para Anotação de Córpus pela Teoria Discursiva Multidocumento CST**. [s.l.] Instituto de Ciências Matemáticas e de Computação, 2008.

ALIGULIYEV, R. M. et al. COSUM: Text summarization based on clustering and optimization. **Expert Systems**, v. 36, n. 1, p. e12340, 2019.

ALIKANIOTIS, D.; YANNAKOUDAKIS, H.; REI, M. **Automatic Text Scoring Using Neural Networks**. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. **Anais...**Association for Computational Linguistics, 2016.

ALLES, V. J. **Construção de um corpus para extrair entidades nomeadas do Diário Oficial da União utilizando aprendizado supervisionado**. mathesis—[s.l.] Master's thesis, Universidade Federal de Brasília, 2018.



ALMEIDA, G. DE. **Translating the post-editor: an investigation of post-editing changes and correlations with professional experience across two Romance languages**. 2013. Disponível em: <<https://api.semanticscholar.org/CorpusID:60255248>>

ALMEIDA, P. G. R. **Uma jornada para um Parlamento inteligente: Câmara dos Deputados do Brasil**. *Red Información*, v. 24, 2021.

ALMEIDA, T. A. et al. Text normalization and semantic indexing to enhance Instant Messaging and SMS spam filtering. *Knowledge-Based Systems*, v. 108, p. 25–32, 2016.

ALMEIDA, T. A.; ALMEIDA, J.; YAMAKAMI, A. Spam filtering: how the dimensionality reduction affects the accuracy of Naive Bayes classifiers. *Journal of Internet Services and Applications*, v. 1, n. 3, p. 183–200, 2011.

ALMEIDA, T. A.; YAMAKAMI, A. Facing the spammers: A very effective approach to avoid junk e-mails. *Expert Systems with Applications*, v. 39, p. 6557–6561, 2012.

ALMOUZINI, S.; KHEMAKHEM, M.; ALAGEEL, A. **Detecting Arabic Depressed Users from Twitter Data**. *Procedia Computer Science*, v. 163, p. 257–265, 2019.

ALONSO, M. A. et al. **Sentiment Analysis for Fake News Detection**. *Electronics*, v. 10, n. 11, 2021.

AL-RFOU, R. et al. **Collective named entity disambiguation using graph-based ranking**. Proceedings of the 23rd International Conference on World Wide Web (WWW '14). *Anais...ACM*, 2014. Disponível em: <<https://doi.org/10.1145/2566486.2568010>>

ALUÍSIO, S. M. et al. Detecting Respiratory Insufficiency Via Voice Analysis: The SPIRA Project. Em: **Practical Machine Learning for Developing Countries on the Tenth International Conference on Learning Representations**. *Proceeding*. [s.l.] ICLR, 2022.

ALUÍSIO, S.; GASPERIN, C. **Fostering Digital Inclusion and Accessibility: The PorSimples project for Simplification of Portuguese Texts**. (T. Solorio, T. Pedersen, Eds.) Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas. *Anais...Los Angeles, California: Association for Computational Linguistics*, jun. 2010. Disponível em: <<https://aclanthology.org/W10-1607>>

ALVARES, R. V.; GARCIA, A. C. B.; FERRAZ, I. **STEMBR: A stemming algorithm for the Brazilian Portuguese language**. Portuguese conference on artificial intelligence. *Anais...Springer*, 2005.

AMANCIO, D. R. et al. **Using metrics from complex networks to evaluate machine translation**. *Physica A: Statistical Mechanics and its Applications*, v. 390, n. 1, p. 131–142, 2011.



AMANCIO, D. R. et al. **Opinion Discrimination Using Complex Network Features**. Complex Networks: Second International Workshop on Complex Networks (Com-pleNet 2010). **Anais...**: Communications em Computer e Information Science.Springer, b2011. Disponível em: <https://doi.org/10.1007/978-3-642-25501-4_16>

AMANCIO, D. R. et al. **Extractive summarization using complex networks and syntactic dependency**. **Physica A: Statistical Mechanics and its Applications**, v. 391, n. 4, p. 1855–1864, a2012.

AMANCIO, D. R. et al. **Complex networks analysis of language complexity**. **Europhysics Letters (EPL)**, v. 100, n. 5, p. 58002, b2012.

AMANCIO, D. R. **A perspective on the advancement of natural language processing tasks via topological analysis of complex networks**. **Physics of Life Reviews**, v. 11, n. 4, p. 641–643, 2014.

AMARAL, D.; VIEIRA, R. Nerp-crf: uma ferramenta para o reconhecimento de entidades nomeadas por meio de conditional random fields. **Linguamática (Braga)**, 2014.

AMERICAN PSYCHIATRIC ASSOCIATION. **Diagnostic and Statistical Manual of Mental Disorders 5th edition**. Arlington, VA: American Psychiatric Association, 2013.

AMORIM, E.; CANÇADO, M.; VELOSO, A. **Automated Essay Scoring in the Presence of Biased Ratings**. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. **Anais...**Association for Computational Linguistics, 2018.

AMORIM, E.; VELOSO, A. **A Multi-aspect Analysis of Automatic Essay Scoring for Brazilian Portuguese**. Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics. **Anais...**Valencia, Spain: Association for Computational Linguistics, abr. 2017.

ANANIADOU, S.; MCNAUGHT, J. **Text Mining for Biology And Biomedicine**. Norwood, MA, USA: Artech House, Inc., 2005.

ANCHIÊTA, R. T. et al. **PiLN IDPT 2021: Irony Detection in Portuguese Texts with Superficial Features and Embeddings**. Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain, September, 2021. **Anais...**2021.

ANDERSEN, P. M. et al. **Automatic extraction of facts from press releases to generate news stories**. Third Conference on Applied Natural Language Processing. **Anais...**1992.

ANDREW, J. J.; TANNIER, X. **Automatic Extraction of Entities and Relation from Legal Documents**. Proceedings of the Seventh Named Entities Workshop.



- Anais...**Melbourne, Australia: Association for Computational Linguistics, jul. 2018.
- ANGELIDIS, I.; CHALKIDIS, I.; KOUBARAKIS, M. **Named entity recognition, linking and generation for greek legislation**. Legal Knowledge and Information Systems. **Anais...**IOS Press, 2018.
- ANSARI, L.; JI, S. Ensemble hybrid learning methods for automated depression detection. **IEEE Transactions on computational Social Systems**, 2022.
- ANTIQUEIRA, L. et al. **Strong correlations between text quality and complex networks features**. **Physica A: Statistical Mechanics and its Applications**, v. 373, p. 811–820, 2007.
- ANTIQUEIRA, L. et al. **A complex network approach to text summarization**. **Information Sciences**, v. 179, n. 5, p. 584–599, 2009.
- APPEL, A. P.; JUNIOR, E. R. H. Por dentro das redes complexas – detectando grupos e prevendo ligações. Em: VILAIN, P.; ROESLER, V. (Eds.). **Tópicos em Banco de Dados, Multimídia e Web: Minicursos do XXVI SBBD e do XVII WebMedia**. Porto Alegre: Sociedade Brasileira de Computação, 2011. p. 123–147.
- ARAGÓN, M. E. et al. **Detecting Depression in Social Media using Fine-Grained Emotions**. NAACL-2019. **Anais...**Minneapolis, USA: Association for Computational Linguistics, 2019.
- ARAUJO, P. H. L. DE et al. **LeNER-Br: A Dataset for Named Entity Recognition in Brazilian Legal Text**. (A. Villavicencio et al., Eds.)Computational Processing of the Portuguese Language. **Anais...**Cham: Springer International Publishing, 2018. Disponível em: <<https://github.com/peluz/lener-br>>
- ARDILA, R. et al. **Common Voice: A Massively-Multilingual Speech Corpus**. (N. Calzolari et al., Eds.)Proceedings of the Twelfth Language Resources and Evaluation Conference. **Anais...**Marseille, France: European Language Resources Association, 2020. Disponível em: <<https://aclanthology.org/2020.lrec-1.520/>>
- AREVALO, E. M.; FONTEYN, L. **MacBERTh: Development and Evaluation of a Historically Pre-trained Language Model for English (1450-1950)**. ICON Workshop on Natural Language Processing for Digital Humanities. **Anais...**2021.
- ARFÉ, B.; MASON, L.; FAJARDO, I. Simplifying informational text structure for struggling readers. **Read Writ (2018) Volume 31, Issue 9**, p. 2191–2210, 2018.
- ARORA, A. K. A. A. S. A. A. **Anxious Depression Prediction in Real-time Social Data**. International Conference on Advances in Engineering Science Management & Technology. **Anais...**Dehradun, India: 2019.
- ASAHARA, M.; MATSUMOTO, Y. **Japanese named entity extraction with redundant morphological analysis**. Proceedings of the 2003 human language technology



conference of the North American chapter of the association for computational linguistics. **Anais...**2003.

ASCHBRENNER, J. A. N. A. A. B. A. J. T. A. K. A. **Social Media and Mental Health: Benefits, Risks, and Opportunities for Research and Practice**. **Journal of Technology in Behavioral Science**, v. 5, p. 245–257, 2020.

ASCHBRENNER, K. A. et al. A survey of online and mobile technology use at peer support agencies. **Psychiatric Quarterly**, p. 1–10, 2018.

ASSI, F. M. et al. **UFSCar's Team at ABSAPT 2022: Using Syntax, Semantics and Context for Solving the Tasks**. Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2022), A Coruña, Spain, September 20, 2022. **Anais...**2022.

AVANÇO, L. V.; NUNES, M. DAS G. V. **Lexicon-Based Sentiment Analysis for Reviews of Products in Brazilian Portuguese**. Proceedings of the 2014 Brazilian Conference on Intelligent Systems. **Anais...**2014.

AVERINA, M.; LEVANOVA, O.; KASATKINA, N. **Named Entity Recognition for Russian Judicial Rulings Text**. 2022 32nd Conference of Open Innovations Association (FRUCT). **Anais...**2022.

AZIZ, W.; SPECIA, L. **Fully Automatic Compilation of a Portuguese-English Parallel Corpus for Statistical Machine Translation**. STIL 2011. **Anais...**Cuiabá, MT: 2011.

AZZIMONTI, M.; FERNANDES, M. **Social media networks, fake news, and polarization**. **European Journal of Political Economy**, v. 76, p. 102256, 2023.

BAADE, A.; PENG, P.; HARWATH, D. **MAE-AST: Masked Autoencoding Audio Spectrogram Transformer**. (H. Ko, J. H. L. Hansen, Eds.)Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022. **Anais...**ISCA, 2022. Disponível em: <<https://doi.org/10.21437/Interspeech.2022-10961>>

BABU, A. et al. XLS-R: Self-supervised cross-lingual speech representation learning at scale. **arXiv preprint arXiv:2111.09296**, 2021.

BACH, N. X. et al. **Reference Extraction from Vietnamese Legal Documents**. Proceedings of the 10th International Symposium on Information and Communication Technology. **Anais...**: SoICT '19. New York, NY, USA: Association for Computing Machinery, 2019.

BAEVSKI, A. et al. **wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations.**, 2020. Disponível em: <<https://arxiv.org/abs/2006.11477>>

BAEZA-YATES, R. A.; RIBEIRO-NETO, B. A. Modern Information Retrieval-the con-



cepts and technology behind search. 2011.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Recuperação de Informação-: Conceitos e Tecnologia das Máquinas de Busca**. [s.l.] Bookman Editora, 2013.

BAHDANAU, D.; CHO, K.; BENGIO, Y. **Neural Machine Translation by Jointly Learning to Align and Translate**. (Y. Bengio, Y. LeCun, Eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. **Anais...**San Diego, California.: 2015. Disponível em: <<http://arxiv.org/abs/1409.0473>>

BAI, J. et al. Qwen technical report. **arXiv preprint arXiv:2309.16609**, 2023.

BALAGE FILHO, P. P.; PARDO, T. A. S.; ALUÍSIO, S. M. **An Evaluation of the Brazilian Portuguese LIWC Dictionary for Sentiment Analysis**. Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology. **Anais...**2013.

BALAGE FILHO, P. P.; PARDO, T. A. S.; NUNES, M. DAS G. V. **Sumarização automática de textos científicos: Estudo de caso com o sistema gistsumm**. [s.l.] Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo, 2007.

BALAMURALI, B. T. et al. **Deep Neural Network-Based Respiratory Pathology Classification Using Cough Sounds**. **Sensors**, v. 21, n. 16, 2021.

BANARESCU, L. et al. **Abstract Meaning Representation for Sembanking**. Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse. **Anais...**Sofia, Bulgaria: Association for Computational Linguistics, 2013. Disponível em: <<http://aclweb.org/anthology/W13-2322>>

BANERJEE, S.; LAVIE, A. **METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments**. (J. Goldstein et al., Eds.) Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. **Anais...**Ann Arbor, Michigan: Association for Computational Linguistics, jun. 2005. Disponível em: <<https://aclanthology.org/W05-0909>>

BANKO, M. et al. **Open Information Extraction from the Web**. Proceedings of the 20th International Joint Conference on Artificial Intelligence. **Anais...** IJCAI'07.San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2007. Disponível em: <<http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=9909B5C03DA1A3CCFFF4263898B69100?doi=10.1.1.74.5174&rep=rep1&type=pdf>>

BANZA, A. P. **A edição digital da História do Futuro, de António Vieira: arquivo e ferramentas**. Actas da Jornada de Humanidades Digitais do CIDEHUS (to appear). **Anais...**2022.

BARABÁSI, A.-L.; ALBERT, R. Emergence of scaling in random networks. **Science**, v. 286, n. 5439, p. 509–512, 1999.



BARACHO, J.; LISBOA, L.; LOPES, R. **Levantamento e Análise Qualitativa de Bases de Dados de Fake News em Português**. Anais do VI Workshop sobre as Implicações da Computação na Sociedade. **Anais...**Porto Alegre, RS, Brasil: SBC, 2025. Disponível em: <<https://sol.sbc.org.br/index.php/wics/article/view/35944>>

BARBOSA DE LIMA, T. et al. **Avaliação Automática de Redação: Uma revisão sistemática**. **Revista Brasileira de Informática na Educação**, v. 31, p. 205–221, 2023.

BARBOSA, G. C. G.; GLAUBER, R.; CLARO, D. B. **Classificação de Relações Abertas Utilizando Features Independentes do Idioma**. Proceedings of the 4th Symposium on Knowledge Discovery, Mining and Learning (KDMiLe). **Anais...**SBC, 2016.

BARRAULT, L. et al. **Findings of the 2019 Conference on Machine Translation (WMT19)**. Proceedings of WMT. **Anais...**Florence, Italy: 2019.

BARRAULT, L. et al. **Findings of the 2020 Conference on Machine Translation (WMT20)**. Proceedings of the Fifth Conference on Machine Translation. **Anais...**Online: Association for Computational Linguistics, nov. 2020. Disponível em: <<https://www.aclweb.org/anthology/2020.wmt-1.1>>

BARRETT, M. J.; AGIC, Z.; SØGAARD, A. The Dundee Treebank. **Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories: TLT14**, p. 242–248, 2015.

BARRIERE, V.; FOURET, A. **May I Check Again? — A simple but efficient way to generate and use contextual dictionaries for Named Entity Recognition. Application to French Legal Texts**. Proceedings of the 22nd Nordic Conference on Computational Linguistics. **Anais...**Turku, Finland: Linköping University Electronic Press, 2019. Disponível em: <<https://aclanthology.org/W19-6136>>

BARROS, T. S. **Um modelo BERT para sumarização extrativa de textos em documentos da Polícia Federal**. mathesis—[s.l.] (Mestrado em Ciências da Computação) - Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande, 2022.

BARZILAY, R.; ELHADAD, N.; MCKEOWN, K. **Sentence ordering in mult-document summarization**. Proceedings of the first international conference on Human language technology research. **Anais...**2001.

BASILE, V. et al. **We Need to Consider Disagreement in Evaluation**. (K. Church, M. Liberman, V. Kordoni, Eds.) Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future. **Anais...**Online: Association for Computational Linguistics, ago. 2021. Disponível em: <<https://aclanthology.org/2021.bppf-1.3/>>

BATISTA, H. H. et al. **A comparative analysis of text embedding approach to extract named entities in Portuguese legal documents**. Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional. **Anais...**SBC, 2021.



BAXENDALE, P. B. Machine-made index for technical literature—an experiment. **IBM Journal of research and development**, v. 2, n. 4, p. 354–361, 1958.

BELCAVELLO, F. et al. **Frame-Based Annotation of Multimodal Corpora: Tracking (A)Synchronies in Meaning Construction**. (T. T. Torrent et al., Eds.) Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet. **Anais...**Marseille, France: European Language Resources Association, 2020. Disponível em: <<https://aclanthology.org/2020.framenet-1.4>>

BELCAVELLO, F. **FrameNet Annotation for Multimodal Corpora: devising a methodology for the semantic representation of text-image interactions in audiovisual productions**. Tese de Doutorado em Linguística—Juiz de Fora: Universidade Federal de Juiz de Fora, 2023.

BELCAVELLO, F. et al. **Frame2: A FrameNet-based Multimodal Dataset for Tackling Text-image Interactions in Video**. (N. Calzolari et al., Eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). **Anais...**Torino, Italia: ELRA; ICCL, 2024. Disponível em: <<https://aclanthology.org/2024.lrec-main.655/>>

BELTAGY, I.; LO, K.; COHAN, A. **SciBERT: A Pretrained Language Model for Scientific Text**. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). **Anais...**Association for Computational Linguistics, 2019. Disponível em: <<https://aclanthology.org/D19-1371>>

BENDER, E. M. **Linguistically Naïve != Language Independent: Why NLP Needs Linguistic Typology**. Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous? **Anais...**Athens, Greece: Association for Computational Linguistics, mar. 2009. Disponível em: <<https://www.aclweb.org/anthology/W09-0106>>

BENGIO, Y.; COURVILLE, A.; VINCENT, P. Representation learning: A review and new perspectives. **IEEE transactions on pattern analysis and machine intelligence**, v. 35, n. 8, p. 1798–1828, 2013.

BERBER SARDINHA, T. **Linguística de Corpus**. Barueri: Manole, 2004.

BERTI, L. C. et al. Fundamental frequency related parameters in Brazilians with COVID-19. **The Journal of the Acoustical Society of America**, v. 153, n. 1, p. 576–585, 2023.

BERTI, L. C. et al. **Acoustic Characteristics of Voice and Speech in Post-COVID-19**. Healthcare. **Anais...**MDPI, 2025.

BHARDWAJ, S.; AGGARWAL, S.; MAUSAM, M. **CaRB: A crowdsourced benchmark for open IE**. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Pro-



cessing (EMNLP-IJCNLP). **Anais...**2019.

BHATTACHARYA, D. et al. Coswara: A respiratory sounds and symptoms dataset for remote screening of SARS-CoV-2 infection. **Scientific data**, v. 10, n. 1, p. 397, 2023.

BIBER, D. **Representativeness in Corpus Design**. **Literary and Linguistic Computing**, v. 8, n. 4, p. 243–257, jan. 1993.

BIBER, D. **Register: Overview**. Em: BROWN, K. (Ed.). **Encyclopedia of Language & Linguistics (Second Edition)**. Second Edition ed. Oxford: Elsevier, 2006. p. 476–482.

BICK, E. **The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework**. tese de doutorado—[s.l.] Aarhus University Press, Denmark; University of Arhus, 2000.

BIKEL, D. M.; SCHWARTZ, R.; WEISCHEDEL, R. M. An algorithm that learns what's in a name. **Machine learning**, v. 34, p. 211–231, 1999.

BIRD, S. **NLTK: the natural language toolkit**. Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions. **Anais...**2006.

BIRNBAUM, M. L. et al. Role of social media and the Internet in pathways to care for adolescents and young adults with psychotic disorders and nonpsychotic mood disorders. **Early Intervention in Psychiatry**, v. 11, n. 4, p. 290–295, 2017.

BITTENCOURT JR., J. A. S. **Avaliação automática de redação em língua portuguesa empregando redes neurais profundas**. mathesis—[s.l.] Universidade Federal de Goiás, 2020.

BITTENCOURT, M. M.; SILVA, R. M.; ALMEIDA, T. A. ML-MDLText: An efficient and lightweight multilabel text classifier with incremental learning. **APPLIED SOFT COMPUTING**, v. 96, p. 1–15, 2020.

BLEI, D. M.; MORENO, P. J. **Topic Segmentation with an Aspect Hidden Markov Model**. Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. **Anais...**New York, NY, USA: Association for Computing Machinery, 2001.

BLONDEL, V. D. et al. **Fast unfolding of communities in large networks**. **Journal of Statistical Mechanics: Theory and Experiment**, v. 2008, n. 10, p. P10008, 2008.

BOITO, M. Z. Simplificação lexical de substantivos e multiword expressions. **Salão de Iniciação Científica (26. : 2014 out. 20-24 : UFRGS, Porto Alegre, RS)**, 2014.

BOJANOWSKI, P. et al. Enriching Word Vectors with Subword Information. **Transactions of the Association for Computational Linguistics**, v. 5, p. 135–146, 2017.

BOJAR, O. et al. **Findings of the 2016 Conference on Machine Translation**.



Proceedings of the First Conference on Machine Translation. **Anais...**Berlin, Germany: Association for Computational Linguistics, 2016.

BOND JR., C. F.; DEPAULO, B. M. Accuracy of Deception Judgments. **Personality and Social Psychology Review**, v. 10, n. 3, p. 214–234, 2006.

BONIFACIO, L. H. et al. **A Study on the Impact of Intradomain Finetuning of Deep Language Models for Legal Named Entity Recognition in Portuguese.** (R. Cerri, R. C. Prati, Eds.)Intelligent Systems. **Anais...**Cham: Springer International Publishing, 2020.

BONIFACIO, L. H. et al. **mMARCO: A Multilingual Version of MS MARCO Passage Ranking Dataset.**, 2021. Disponível em: <<https://arxiv.org/abs/2108.13897>>

BONIFACIO, L. H. N. **Modelos Profundos de Linguagem para Reconhecimento de Entidades Nomeadas em Domínio Jurídico.** mathesis—[s.l.] Master's thesis, Universidade Federal de Mato Grosso do Sul, 2020.

BORDINO, I. et al. **Garnlp: a natural language processing pipeline for garnishment documents.** **Information Systems Frontiers**, v. 23, p. 101–114, 2021.

BORIN, E.; DONATO, F. Financial Sustainability of Digitizing Cultural Heritage: The International Platform Europeana. **Journal of Risk and Financial Management**, v. 16, n. 10, p. 421, 2023.

BOWKER, L.; PEARSON, J. **Working with Specialized Language: A Practical Guide to Using Corpora.** 1. ed. London: Routledge, 2002.

BRANDT, M. B. **Modelagem da informação legislativa: arquitetura da informação para o processo legislativo brasileiro.** tese de doutorado—[s.l.] Universidade Estadual Paulista (Unesp), 2020.

BRAUN, H. I. Understanding Scoring Reliability: Experiments in Calibrating Essay Readers. **Journal of Educational Statistics**, v. 13, n. 1, p. 1–18, 1988.

BREIMAN, L. Random forests. **Machine learning**, v. 45, n. 1, p. 5–32, 2001.

BREITFELLER, L. et al. **Finding Microaggressions in the Wild: A Case for Locating Elusive Phenomena in Social Media Posts.** Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). **Anais...**2019.

BRICIU, A.; LUPEA, M. **Studying the language of mental illness in Romanian social media.** IEEE 14th International Conference on Intelligent Computer Communication and Processing (ICCP). **Anais...**2018.

BRIDGEMAN, B. Handbook of automated essay evaluation: Current applications and new directions. Em: SHERMIS, M. D.; BURSTEIN, J. (Eds.). [s.l.] Routledge/Taylor &



Francis Group, 2013. p. 221–232.

BRIGHAM, E. O.; MORROW, R. The fast Fourier transform. **IEEE spectrum**, v. 4, n. 12, p. 63–70, 1967.

BRIN, S. **Extracting patterns and relations from the world wide web**. International workshop on the world wide web and databases. **Anais...**Springer, 1998.

BRITO, M. et al. **CDJUR-BR - Uma Coleção Dourada do Judiciário Brasileiro com Entidades Nomeadas Refinadas**. Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. **Anais...**Porto Alegre, RS, Brasil: SBC, 2023. Disponível em: <<https://github.com/mauriciobritojr/CDJUR-BR>>

BRITTO, H.; FINGER, M.; GALVES, C. Computational and linguistic aspects of the construction of The Tycho Brahe Parsed Corpus of Historical Portuguese. **Romanistische Korpuslinguistik, Korpora und gesprochene Sprache, Romance Corpus Linguistics, Corpora and Spoken Language, ScriptOralia**, v. 126., 2002.

BROWN, C. et al. **Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data**. Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. **Anais...**a2020.

BROWN, P. et al. **A statistical approach to language translation**. Proceedings of the 12th conference on Computational linguistics -. **Anais...**Budapest, Hungary: Association for Computational Linguistics, 1988. Disponível em: <<http://portal.acm.org/citation.cfm?doid=991635.991651>>. Acesso em: 10 jun. 2020

BROWN, T. et al. Language models are few-shot learners. **Advances in neural information processing systems**, v. 33, p. 1877–1901, b2020.

BROWN, T. B. et al. **Language Models are Few-Shot Learners**. (H. Larochelle et al., Eds.)Advances in Neural Information Processing Systems. **Anais...**Curran Associates, Inc., c2020. Disponível em: <<https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfbcb4967418bfb8ac142f64a-Abstract.html>>

BRUM, H.; NUNES, M. DAS G. V. **Building a Sentiment Corpus of Tweets in Brazilian Portuguese**. (N. C. (Conference chair) et al., Eds.)Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). **Anais...**Miyazaki, Japan: European Language Resources Association (ELRA), mar. 2018.

BRUNETTE, M. et al. Use of smartphones, computers and social media among people with SMI: opportunity for intervention. **Community Mental Health Journal**, p. 1–6, 2019.

BUCCI, S.; SCHWANNAUER, M.; BERRY, N. The digital revolution and its impact on mental health care. **Psychology and Psychotherapy: Theory, Research and Practice**, v. 92, n. 2, p. 277–297, 2019.



BUCKLEY, C.; VOORHEES, E. M. **Evaluating evaluation measure stability**. ACM SIGIR Forum. **Anais...**ACM New York, NY, USA, 2017.

BUDEL, G. et al. **Topological properties and organizing principles of semantic networks**. arXiv preprint arXiv:2304.12940, 2023.

BUENO, R. O. et al. **Overview of the Task on Irony Detection in Spanish Variants**. Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing. **Anais...**2019.

BURDISO, S. G.; ERRECALDE, M.; MONTES-Y-GÓMEZ, M. **t-SS3: a text classifier with dynamic n-grams for early risk detection over text streams**. **Pattern Recognition Letters**, v. 138, p. 130–137, 2020.

BURRISS, L. L. Attribution in network radio news: A cross-network analysis. **Journalism Quarterly**, v. 65, n. 3, p. 690–694, 1988.

BURSTEIN, J. **Opportunities for Natural Language Processing Research in Education**. (A. Gelbukh, Ed.) Computational Linguistics and Intelligent Text Processing. **Anais...**Berlin, Heidelberg: Springer Berlin Heidelberg, 2009.

CABRAL, B.; SOUZA, M.; CLARO, D. B. **PortNOIE: A Neural Framework for Open Information Extraction for the Portuguese Language**. International Conference on Computational Processing of the Portuguese Language. **Anais...**Springer, 2022.

CABRAL, L. et al. **FakeWhastApp.BR: NLP and Machine Learning Techniques for Misinformation Detection in Brazilian Portuguese WhatsApp Messages**. Proceedings of the 23rd International Conference on Enterprise Information Systems (ICEIS 2021) - Volume 1. **Anais...**2021.

CABRÉ, M. T. A Terminologia, uma disciplina em evolução: passado, presente e alguns elementos de futuro. **Debate Terminológico**. ISSN: 1813-1867, n. 01, 2005.

CABRERA-DIEGO, L. A.; GHEEWALA, A. **Jus Mundi at SemEval-2023 Task 6: Using a Frustratingly Easy Domain Adaption for a Legal Named Entity Recognition System**. Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023). **Anais...**Toronto, Canada: Association for Computational Linguistics, jul. 2023.

CAI, Y. et al. **Rank-Then-Score: Enhancing Large Language Models for Automated Essay Scoring**, 2025. Disponível em: <<https://arxiv.org/abs/2504.05736>>

CAMERON, H. F.; GONÇALVES, M. F.; QUARESMA, P. **Linguistic and orthographical classic Portuguese variants Challenges for NLP**. Proceedings of the 14th International Conference on the Computational Processing of Portuguese. **Anais...**2020.

CAMERON, H.; OLIVAL, F.; VIEIRA, R. **Planear a normalização automática: tipologia de variação gráfica do corpus das Memórias Paroquiais (1758)**. **LaborHistórico**, v. 9, n.



1, p. 52234, 2023.

CANAVILHAS, J. Webjornalismo: Da pirâmide invertida à pirâmide deitada. **Aprender**, n. 32, p. 58–65, 2012.

CANDIDO JUNIOR, A. **Análise bidirecional da língua na simplificação sintática em textos de português voltada à acessibilidade digital**. ICMC - USP São Carlos: Biblioteca Digital USP, 2013.

CANDIDO-JUNIOR, A.; OLIVEIRA, M. DE; ALUÍSIO, S. M. Simplifica: um Sistema Web de Autoria de Textos Simplificados. **Simpósio Brasileiro de Sistemas Multimídia e Web (Webmedia 2009) v.2**, p. 55–58, 2009.

CARDELLINO, C. et al. **Legal NERC with ontologies, Wikipedia and curriculum learning**. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. **Anais...Valencia, Spain: Association for Computational Linguistics**, abr. 2017. Disponível em: <<https://aclanthology.org/E17-2041>>

CARDOSO, E. F.; SILVA, R. M.; ALMEIDA, T. A. Towards automatic filtering of fake reviews. **Neurocomputing**, v. 309, p. 106–116, 2018.

CARDOSO, P. C. F. et al. **CSTNews-a discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese**. Proceedings of the 3rd RST Brazilian Meeting. **Anais...2011**.

CARDOSO, P. C. F. **Exploração de métodos de sumarização automática multi-documento com base em conhecimento semântico-discursivo**. tese de doutorado—[s.l.] (Doutorado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2014.

CARL, M.; WAY, A. (EDS.). **Recent Advances in Example-Based Machine Translation**. [s.l.] Springer Netherlands, 2003.

CARMO, D. et al. **PTT5: Pretraining and validating the T5 model on Brazilian Portuguese data**. **CoRR**, v. abs/2008.09144, b2020.

CARMO, D. et al. **PTT5: Pretraining and validating the T5 model on Brazilian Portuguese data**. **arXiv preprint arXiv:2008.09144**, a2020.

CARNEIRO, F. C. A. D. F. A. F. J. N. A. V. **Early Detection of Depression: Social Network Analysis and Random Forest Techniques**. **J Med Internet Res**, v. 21, n. 6, p. e12554, 2019.

CARPINETO, C.; ROMANO, G. A survey of automatic query expansion in information retrieval. **Acm Computing Surveys (CSUR)**, v. 44, n. 1, p. 1–50, 2012.

CARROLL, J. et al. **Practical Simplification of English Newspaper Text to Assist**



Aphasic Readers. In Proc. of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology. **Anais...**1998.

CARVALHO, F.; SANTOS, G. DOS; GUEDES, G. P. **AffectPT-br: an Affective Lexicon based on LIWC 2015.** Proceedings of the 37th International Conference of the Chilean Computer Science Society. **Anais...**2018.

CARVALHO, P. et al. **Clues for Detecting Irony in User-Generated Contents: Oh...!! It's "so Easy";-).** Proceedings of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion. **Anais...**2009.

CARVALHO, P.; SILVA, M. J. **SentiLex-PT 02.** <https://b2share.eudat.eu>, 2017. Disponível em: <<https://b2share.eudat.eu/records/93ab120efdaa4662baec6adee8e7585f>>

CASANOVA, E. et al. **Transfer Learning and Data Augmentation Techniques to the COVID-19 Identification Tasks in ComParE 2021.** Proc. Interspeech 2021. **Anais...**a2021.

CASANOVA, E. et al. **Deep Learning against COVID-19: Respiratory Insufficiency Detection in Brazilian Portuguese Speech.** Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. **Anais...**Online: Association for Computational Linguistics, ago. b2021.

CASELI, H. DE M. et al. Building a Brazilian Portuguese parallel corpus of original and simplified texts. **Advances in Computational Linguistics, Research in Computer Science (CICLing-2009)**, v. 41, p. 59–70, 2009.

CASELI, H. DE M.; INÁCIO, M. **NMT and PBSMT Error Analyses in English to Brazilian Portuguese Automatic Translations.** Proceedings of the Twelfth Language Resources and Evaluation Conference. **Anais...**Marseille, France: European Language Resources Association, 2020. Disponível em: <<https://aclanthology.org/2020.lrec-1.446>>

CÁSSIA ALVES, V. DE et al. **College students-in-the-loop for their mental health: a case of AI and humans working together to support well-being.** **Interaction Design and Architecture(s)**, n. 59, p. 79–94, 2024.

CASTANO, A.; CASACUBERTA, F. **A connectionist approach to machine translation.** 5th European Conference on Speech Communication and Technology (Eurospeech 1997). **Anais...**ISCA, set. 1997. Disponível em: <<http://dx.doi.org/10.21437/eurospeech.1997-50>>

CASTILHO, S. et al. **Does post-editing increase usability? A study with Brazilian Portuguese as Target Language.** Proceedings of the 17th annual conference of the European association for machine translation. **Anais...**2014.

CASTILHO, S. et al. **A comparative quality evaluation of PBSMT and NMT using professional translators.** Proceedings of Machine Translation Summit XVI: Research Track. **Anais...**a2017.



CASTILHO, S. et al. **Is Neural Machine Translation the New State of the Art? The Prague Bulletin of Mathematical Linguistics**, v. 108, n. 1, p. 109–120, jun. b2017.

CASTILHO, S. et al. Approaches to Human and Machine Translation Quality Assessment. Em: **Translation Quality Assessment: From Principles to Practice**. Machine Translation: Technologies e Applications. [s.l.] Springer International Publishing, 2018. v. 1p. 9–38.

CASTILHO, S. et al. **Editors' foreword to the special issue on human factors in neural machine translation**. **Machine Translation**, v. 33, n. 1–2, p. 1–7, maio a2019.

CASTILHO, S. **On the Same Page? Comparing IAA in Sentence and Document Level Human MT Evaluation**. Proceedings of the Fifth Conference on Machine Translation. **Anais...Association for Computational Linguistics**, nov. 2020. Disponível em: <<https://www.aclweb.org/anthology/2020.wmt-1.137>>

CASTILHO, S. **Towards Document-Level Human MT Evaluation: On the Issues of Annotator Agreement, Effort and Misevaluation**. Proceedings of the Workshop on Human Evaluation of NLP Systems. **Anais...Association for Computational Linguistics**, abr. 2021. Disponível em: <<https://www.aclweb.org/anthology/2021.humeval-1.4>>

CASTILHO, S. **How Much Context Span is Enough? Examining Context-Related Issues for Document-level MT**. Proceedings of the Language Resources and Evaluation Conference. **Anais...Marseille, France: European Language Resources Association**, 2022. Disponível em: <<https://aclanthology.org/2022.lrec-1.323>>

CASTILHO, S. et al. **Translation Systems Care for Context? What About a GPT Model?** Proceedings of the 24th Annual Conference of the European Association for Machine Translation. **Anais...Tampere, Finland: EAMT**, 2023. Disponível em: <<https://events.tuni.fi/uploads/2023/06/11678752-proceedings-eamt2023.pdf>>

CASTILHO, S.; RESENDE, N. Post-Editese in Literary Translations. **Information**, v. 13, n. 2, p. 66, 2022.

CASTILHO, S.; RESENDE, N.; MITKOV, R. **What Influences the Features of Post-editese? A Preliminary Study**. Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019). **Anais...Varna, Bulgaria: Incoma Ltd., Shoumen, Bulgaria**, set. b2019. Disponível em: <<https://aclanthology.org/W19-8703>>

CASTRO, P. **Aprendizagem profunda para reconhecimento de entidades nomeadas em domínio jurídico**. mathesis—[s.l.] Master's thesis, Universidade Federal de Goiás, 2019.

CASTRO, P. V. Q. DE; SILVA, N. F. F. DA; SOARES, A. DA S. **Portuguese Named Entity Recognition Using LSTM-CRF**. (A. Villavicencio et al., Eds.) Proceedings of the 13th International Conference on the Computational Processing of the Portuguese Language. **Anais...2018**.



ÇETINDAĞ, C.; YAZICIOĞLU, B.; KOÇ, A. **Named-entity recognition in Turkish legal texts**. *Natural Language Engineering*, p. 1–28, 2022.

CHAE, Y.; DAVIDSON, T. Large language models for text classification: from zero-shot learning to instruction-tuning. *Sociological Methods & Research*, p. 00491241251325243, 2025.

CHAKRABORTY, A. et al. **Stop Clickbait: Detecting and preventing clickbaits in online news media**. 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). *Anais...*2016.

CHALKIDIS, I. et al. **LEGAL-BERT: The Muppets straight out of Law School**. Findings of the Association for Computational Linguistics: EMNLP 2020. *Anais...*Online: Association for Computational Linguistics, nov. 2020.

CHALKIDIS, I. et al. **Regulatory Compliance through Doc2Doc Information Retrieval: A case study in EU/UK legislation where text similarity has limitations**. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. *Anais...*Online: Association for Computational Linguistics, abr. 2021.

CHALKIDIS, I.; ANDROUTSOPOULOS, I. **A deep learning approach to contract element extraction**. Em: **Legal knowledge and information systems**. [s.l.] IOS Press, 2017. p. 155–164.

CHALKIDIS, I.; ANDROUTSOPOULOS, I.; ALETRAS, N. **Neural Legal Judgment Prediction in English**. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. *Anais...*Association for Computational Linguistics, 2019.

CHALKIDIS, I.; ANDROUTSOPOULOS, I.; MICHOS, A. **Extracting Contract Elements**. Proceedings of the 16th Edition of the International Conference on Artificial Intelligence and Law. *Anais...* ICAIL '17. New York, NY, USA: Association for Computing Machinery, 2017.

CHALL, J. S.; DALE, E. **Readability revisited: the new Dale-Chall readability formula**. [s.l.] Brookline Books, 1995.

CHALMERS, D. J. Syntactic transformations on distributed representations. **Connectionist Natural Language Processing: Readings from Connection Science**, p. 46–55, 1992.

CHANDRASEKAR, R.; DORAN, C.; SRINIVAS, B. Motivations and methods for text simplification. **Proceedings of the 16th International Conference on Computational Linguistics (COLING)**, p. 1041–1044, 1996.

CHARLES, A. C.; RUBACK, L.; OLIVEIRA, J. **Fakepedia Corpus: A Flexible Fake News Corpus in Portuguese**. Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23,



2022, Proceedings. **Anais...**Berlin, Heidelberg: Springer-Verlag, 2022. Disponível em: <https://doi.org/10.1007/978-3-030-98305-5_4>

CHAVARRO, J. et al. **FakeTrueBR: Um corpus brasileiro de notícias falsas**. Anais da XVIII Escola Regional de Banco de Dados. **Anais...**Porto Alegre, RS, Brasil: SBC, 2023. Disponível em: <<https://sol.sbc.org.br/index.php/erbd/article/view/24352>>

CHEN, A.; CHEN, D. O. Simulation of a machine learning enabled learning health system for risk prediction using synthetic patient data. **Scientific Reports**, v. 12, n. 1, p. 17917, out. 2022.

CHEN, D. et al. **Critical nodes identification in complex networks: a survey**. **Complex Engineering Systems**, v. 5, n. 11, p. –, 2025.

CHEN, P. P. **The Entity-Relationship Model - Toward a Unified View of Data**. **ACM Trans. Database Syst.**, v. 1, n. 1, p. 9–36, 1976.

CHEN, T.; GUESTRIN, C. **XGBoost: A scalable tree boosting system**. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. **Anais...**ACM, 2016.

CHEN, Y. et al. **Joint Entity and Relation Extraction for Legal Documents with Legal Feature Enhancement**. Proceedings of the 28th International Conference on Computational Linguistics. **Anais...**Barcelona, Spain (Online): International Committee on Computational Linguistics, dez. 2020.

CHEN, Y.; CONROY, N. J.; RUBIN, V. L. **Misleading Online Content: Recognizing Clickbait as "False News"**. Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection. **Anais...** WMDD '15. New York, NY, USA: Association for Computing Machinery, 2015. Disponível em: <<https://doi.org/10.1145/2823465.2823467>>

CHOUDHURY, M. D. et al. **Predicting Depression via Social Media**. International AAAI Conference on Web and Social Media (ICWSM). **Anais...**AAAI, 2013.

CHRISMAN, L. Learning recursive distributed representations for holistic computation. **Connection Science**, v. 3, n. 4, p. 345–366, 1991.

CIAMPAGLIA, G. L. et al. Computational fact checking from knowledge networks. **PloS one**, v. 10, n. 6, p. e0128193, 2015.

CIGNARELLA, A. T. et al. **Overview of the EVALITA 2018 Task on Irony Detection in Italian Tweets (IronITA)**. Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018). **Anais...**2018.

CLARK, K. et al. **ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators**. 8th International Conference on Learning Representations,



- ICLR 2020. **Anais...**Addis Ababa, Ethiopia: OpenReview.net, abr. 2020. Disponível em: <<https://openreview.net/forum?id=r1xMH1BtvB>>
- CLARKE, D. J. B. et al. **FAIRshake: Toolkit to Evaluate the FAIRness of Research Digital Resources**. *Cell Systems*, v. 9, n. 5, p. 417–421, 2019.
- CLEM, S. Post-Truth and Vices Opposed to Truth. v. 37, n. 2, p. 97–116, 2017.
- COELHO, G. et al. **Information Extraction in the Legal Domain: Traditional Supervised Learning vs. ChatGPT**. INSTICC; SciTePress, 2024.
- COELLO, J. M. A.; JUNQUEIRA, B. A. Automatic Analysis of Facebook Posts and Comments Written in Brazilian Portuguese. **International Journal for Innovation Education and Research**, 2019.
- COHAN, A. et al. **SMHD: a Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions**. COLING-2018. **Anais...**Santa Fe, USA: Association for Computational Linguistics, 2018.
- COHEN, J. P. et al. Problems in the deployment of machine-learned models in health care. **Cmaj**, v. 193, n. 35, p. E1391–E1394, 2021.
- COLEMAN, M.; LIAU, T. L. A computer readability formula designed for machine scoring. **Journal of Applied Psychology**, v. 60, p. 283–284, 1975.
- COLLOVINI, S. et al. **IberLEF 2019 Portuguese Named Entity Recognition and Relation Extraction Tasks**. Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing. **Anais...**2019. Disponível em: <http://ceur-ws.org/Vol-2421/NER/_Portuguese/_overview.pdf>
- CONCEIÇÃO, M. C.; ZANOLA, M. T. **Terminologia e mediação linguística: métodos, práticas e atividades**. Universidade do Algarve Editora, 2020.
- CONEGLIAN, C. S.; SANTAREM SEGUNDO, J. E. **Europeana no Linked Open Data: conceitos de Web Semântica na dimensão aplicada das Humanidades Digitais**. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, v. 22, n. 48, p. 88–99, 2017.
- CONNEAU, A. et al. **Unsupervised cross-lingual representation learning at scale**. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. **Anais...**2020.
- CONROY, N. K.; RUBIN, V. L.; CHEN, Y. **Automatic deception detection: Methods for finding fake news**. **Proceedings of the Association for Information Science and Technology**, v. 52, n. 1, p. 1–4, 2015.
- CONSOLI, B. S. et al. **Embeddings for Named Entity Recognition in Geoscience**



Portuguese Literature. Proceedings of The 12th Language Resources and Evaluation Conference. **Anais...**2020.

CONSORTIUM, L. D. ACE (Automatic Content Extraction) English Annotation Guidelines for Events. **Version**, n. 5.4.3, 2005.

COPPERSMITH, G. et al. **CLPsych 2015 Shared Task: Depression and PTSD on Twitter.** Second Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. **Anais...**Denver, USA: Association for Computational Linguistics, 2015.

CORDEIRO, P. R.; PINHEIRO, V. **Um corpus de notícias falsas do twitter e verificação automática de rumores em língua portuguesa.** Proceedings of the Symposium in Information and Human Language Technology. **Anais...**2019.

CORNU, G. **Linguistique juridique.** [s.l.: s.n.].

CORRÊA, U. B. **Análise de sentimento baseada em aspectos usando aprendizado profundo: uma proposta aplicada à língua portuguesa.** tese de doutorado—[s.l.] Universidade Federal de Pelotas, 2021.

CORREIA, F. A. et al. **Fine-grained legal entity annotation: A case study on the Brazilian Supreme Court.** **Information Processing & Management**, v. 59, n. 1, p. 102794, 2022.

CORTES, C.; VAPNIK, V. **Support-Vector Networks.** **Machine Learning**, v. 20, n. 3, p. 273–297, set. 1995.

CORTES, E. et al. **An Empirical Comparison of Question Classification Methods for Question Answering Systems.** (N. Calzolari et al., Eds.)Proceedings of the Twelfth Language Resources and Evaluation Conference. **Anais...**Marseille, France: European Language Resources Association, 2020. Disponível em: <<https://aclanthology.org/2020.lrec-1.665>>

COSTA, A. et al. **A linguistically motivated taxonomy for Machine Translation error analysis.** **Machine Translation**, v. 29, n. 2, p. 127–161, 2015.

COSTA, A. D. **A tradução por máquina enriquecida semanticamente com frames e papéis qualia.** Ph.D. thesis—Juiz de Fora: Universidade Federal de Juiz de Fora, 2020.

COSTA, P. B. DA et al. **BERTabaporu: assessing a genre-specific language model for Portuguese NLP.** Recent Advances in Natural Language Processing (RANLP-2023). **Anais...**Varna, Bulgaria: 2023.

COSTA, R. et al. **Expanding UlyssesNER-Br Named Entity Recognition Corpus with Informal User-Generated Text.** (G. Marreiros et al., Eds.)Progress in Artificial Intelligence. **Anais...**Cham: Springer International Publishing, 2022. Disponível em: <<https://github.com/ulysses-camara/>>



- COUTINHO, I.; MARTINS, B. **Transformer-based models for ICD-10 coding of death certificates with Portuguese text.** *Journal of Biomedical Informatics*, v. 136, p. 104232, 2022.
- COUTO, J. M. M.; REIS, J. C. S.; BENEVENUTO, F. **Can computer network attributes be useful for identifying low-credibility websites? A case study in Brazil.** *Social Network Analysis and Mining*, v. 14, n. 1, p. 153, 2024.
- COVER, T.; HART, P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, v. 13, n. 1, p. 21–27, 1967.
- COWIE, J. R. **Automatic analysis of descriptive texts.** First Conference on Applied Natural Language Processing. *Anais...*1983.
- COWIE, J.; LEHNERT, W. Information extraction. *Communications of the ACM*, v. 39, n. 1, p. 80–91, 1996.
- COX, A. N. A. M., Simon J. D. AND Gonzalez-Beltran. **Ten simple rules for making a vocabulary FAIR.** *PLOS Computational Biology*, v. 17, n. 6, p. 1–15, jun. 2021.
- COX, D. R. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, v. 20, n. 2, p. 215–232, 1958.
- CRESTANI, F. et al. **“Is this document relevant?... probably” a survey of probabilistic models in information retrieval.** *ACM Computing Surveys (CSUR)*, v. 30, n. 4, p. 528–552, 1998.
- CROFT, W. **Typology and Universals.** 2. ed. [s.l.] Cambridge University Press, 2002.
- CROFT, W. B.; METZLER, D.; STROHMAN, T. **Search engines: Information retrieval in practice.** [s.l.] Addison-Wesley, 2010. v. 520
- CSIKSZENTMIHALYI, M. **Flow: The Psychology of Optimal Experience.** [s.l.] Harper Perennial, 2008.
- CUCCHIARELLI, A.; VELARDI, P. Unsupervised named entity recognition using syntactic and semantic contextual evidence. *Computational Linguistics*, v. 27, n. 1, p. 123–131, 2001.
- CUI, H. et al. **Probabilistic query expansion using query logs.** Proceedings of the 11th international conference on World Wide Web. *Anais...*2002.
- CUI, L.; WEI, F.; ZHOU, M. **Neural Open Information Extraction.** Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). *Anais...*2018.
- CULOTTA, A.; MCCALLUM, A.; BETZ, J. **Integrating probabilistic extraction models and data mining to discover relations and patterns in text.** Procee-



dings of the Human Language Technology Conference of the NAACL, Main Conference. **Anais...**2006.

CUNHA, A. L. V. DA. **Coh-Metrix-Dementia: análise automática de distúrbios de linguagem nas demências utilizando Processamento de Línguas Naturais**. ICMC - USP São Carlos: Biblioteca Digital USP, 2015.

CUNHA, L. C. C. DA. **Um Corpus anotado de mensagens do WhatsApp em PT-BR para detecção automática de desinformação textual**. <https://github.com/cabrau/FakeWhatsApp.Br>, 2021.

DA SILVA JR., J. A. **Um avaliador automático de redações**. mathesis—[s.l.] Universidade Federal do Espírito Santo, 2021.

DADICO, C. M. **O Ódio Ancestral Como Elemento Constitutivo Do Estado Moderno e Seus Reflexos Na Compreensão dos Crimes De Ódio: Um Diálogo Entre o Direito Internacional e o Direito Brasileiro**. tese de doutorado—Porto Alegre, RS, Brazil: Programa de Pós-Graduação em Ciências Criminais da Escola de Direito da Pontifícia Universidade Católica do Rio Grande do Sul, 2020.

DAHL, Ö. **The Growth and Maintenance of Linguistic Complexity**. Amsterdam: John Benjamins Publishing Company, 2004. v. 71

DAI, E.; SUN, Y.; WANG, S. **Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository**. Proceedings of the International AAAI Conference on Web and Social Media. **Anais...**Atlanta, USA: 2020.

DALE, R.; MAZUR, P. **Handling Conjunctions in Named Entities**. (A. Gelbukh, Ed.)Computational Linguistics and Intelligent Text Processing. **Anais...**Berlin, Heidelberg: Springer Berlin Heidelberg, 2007.

DALIANIS, H. **Characteristics of Patient Records and Clinical Corpora**. Em: **Clinical Text Mining: Secondary Use of Electronic Patient Records**. Cham: Springer International Publishing, 2018. p. 21–34.

DARJI, H.; MITROVIĆ, J.; GRANITZER, M. **German BERT Model for Legal Named Entity Recognition**. Proceedings of the 15th International Conference on Agents and Artificial Intelligence. **Anais...**SCITEPRESS - Science; Technology Publications, 2023.

DARPA (ED.). **Proceedings of the 3rd Message Understanding Conference (MUC-3)**. San Diego, EUA: Morgan Kaufmann, 1991.

DE OLIVEIRA, J. M.; ANTUNES, R. S.; DA COSTA, C. A. **SOAP classifier for free-text clinical notes with domain-specific pre-trained language models**. **Expert Systems with Applications**, v. 245, p. 123046, 2024.

DE PAIVA, V.; RADEMAKER, A.; MELO, G. DE. **OpenWordNet-PT: An Open**



Brazilian Wordnet for Reasoning. Proceedings of COLING 2012: Demonstration Papers. **Anais...**2012.

DE SOUSA, S. C.; AZIZ, W.; SPECIA, L. **Assessing the post-editing effort for automatic and semi-automatic translations of DVD subtitles.** Proceedings of the International Conference Recent Advances in Natural Language Processing 2011. **Anais...**2011.

DE SOUZA, M. C. et al. **Keywords attention for fake news detection using few positive labels.** **Information Sciences**, v. 663, p. 120300, 2024.

DEERWESTER, S. et al. Indexing by latent semantic analysis. **Journal of the American society for information science**, v. 41, n. 6, p. 391–407, 1990.

DEJONG, G. Prediction and substantiation: A new approach to natural language processing. **Cognitive Science**, v. 3, n. 3, p. 251–273, 1979.

DEL CORRO, L.; GEMULLA, R. **Clausie: clause-based open information extraction.** Proceedings of the 22nd international conference on World Wide Web. **Anais...** WWW '13. New York, NY, USA: ACM; ACM, 2013. Disponível em: <<http://doi.acm.org/10.1145/2488388.2488420>>

DELL'ORLETTA, F.; MONTEMAGNI, S.; VENTURI, G. Read-it: Assessing readability of italian texts with a view to text simplification. **Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies**, p. 73–83, 2011.

DEMNER-FUSHMAN, D.; CHAPMAN, W. W.; MCDONALD, C. J. What can natural language processing do for clinical decision support? **J Biomed Inform**, v. 42, n. 5, p. 760–772, ago. 2009.

DESPOTOVIC, V. et al. **Detection of COVID-19 from voice, cough and breathing patterns: Dataset and preliminary results.** **Computers in Biology and Medicine**, v. 138, p. 104944, 2021.

DEVARAJU, A. et al. **FAIRsFAIR Data Object Assessment Metrics 0.5.** [s.l.] Research Data Alliance (RDA), out. 2020. Disponível em: <<https://zenodo.org/record/6461229>>.

DEVLIN, J. et al. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.** (J. Burstein, C. Doran, T. Solorio, Eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019. **Anais...**Minneapolis, MN, USA: Association for Computational Linguistics, 2019. Disponível em: <<https://doi.org/10.18653/v1/n19-1423>>

DIAS, M. S. et al. **A qualitative analysis of a corpus of opinion summaries based on aspects.** Proceedings of the 1st Workshop on Tools and Resources for Automatically Processing Portuguese and Spanish. **Anais...**2014.



DIAS-DA-SILVA, B. C.; MORALES, H. R. DE. A Construção de um Thesaurus Eletrônico para o Português. **Alfa**, 2003.

DIAZ, F.; MITRA, B.; CRASWELL, N. **Query Expansion with Locally-Trained Word Embeddings**. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. **Anais...**2016.

DODDINGTON, G. **Automatic Evaluation of Machine Translation Quality Using N-Gram Co-Occurrence Statistics**. Proceedings of the Second International Conference on Human Language Technology Research. **Anais...**: HLT '02.San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2002.

DODDINGTON, G. et al. **The Automatic Content Extraction (ACE) Program: Tasks, Data, and Evaluation**. (M. T. Lino et al., Eds.)Proceedings of LREC'2004, Fourth International Conference on Language resources and Evaluation (Lisboa, 26-28 May 2004). **Anais...**2004. Disponível em: <<http://www.lrec-conf.org/proceedings/lrec2004/pdf/5.pdf>>

DOHARE, S.; GUPTA, V.; KARNICK, H. **Unsupervised semantic abstractive summarization**. Proceedings of ACL 2018, Student Research Workshop. **Anais...**Melbourne, Australia: Association for Computational Linguistics (ACL), 2018.

DOHERTY, S. et al. **Mapping the industry I: Findings on translation technologies and quality assessment**. QTLaunchPad – Mapping the Industry I: Findings on Translation Technologies and Quality Assessment. **Anais...**GALA, 2013. Disponível em: <http://doras.dcu.ie/19474/1/Version_Participants_Final.pdf>. Acesso em: 11 nov. 2015

DOHERTY, S. et al. **On Education and Training in Translation Quality Assessment**. Em: MOORKENS, J. et al. (Eds.). **Translation Quality Assessment: From Principles to Practice**. Cham: Springer International Publishing, 2018. p. 95–106.

DONG, Z. et al. Exploring context window of large language models via decomposed positional vectors. **Advances in Neural Information Processing Systems**, v. 37, p. 10320–10347, 2024.

DORR, B. et al. Machine translation evaluation and optimization. Em: **Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation**. [s.l.] Springer, 2011. p. 745–843.

DOSOVITSKIY, A. et al. **An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale**. International Conference on Learning Representations. **Anais...**2021.

DOZIER, C. et al. **Named Entity Recognition and Resolution in Legal Text**. Em: FRANCESCONI, E. et al. (Eds.). **Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. p. 27–43.



DUBAY, W. *Robert Gunning's Fog Readability Formula*. **Plain Language At Work Newsletter**, v. 8, 2014.

DUBAY, W. H. **Smart Language: Readers, Readability, and the Grading of Text**. Costa Mesa, CA: Impact Information, 2007.

DUTRA, L. **Evaluating the Contribution of Framenet to Gender-Based Violence Identification: How semantic annotation can be used as a resource for identifying patterns of violence**. Dissertação (Mestrado em Tecnologia da Linguagem)—Gotemburgo: Universidade de Gotemburgo, 2024.

EARL, L. L. Experiments in automatic extracting and indexing. **Information Storage and Retrieval**, v. 6, n. 4, p. 313–330, 1970.

EDMUNDSON, H. P. New methods in automatic extracting. **Journal of the ACM (JACM)**, v. 16, n. 2, p. 264–285, 1969.

EFRON, B.; TIBSHIRANI, R. J. **An Introduction to the Bootstrap**. New York: Chapman & Hall/CRC, 1993.

EHRET, K. *An Information-Theoretic View on Language Complexity and Register Variation: Compressing Naturalistic Corpus Data*. **Corpus Linguistics and Linguistic Theory**, v. 17, n. 2, p. 383–410, out. 2021.

EHRET, K. et al. *Measuring Language Complexity: Challenges and Opportunities*. **Linguistics Vanguard**, v. 9, n. s1, p. 1–8, maio 2023.

EHRET, K.; SZMRECSANYI, B. *An Information-Theoretic Approach to Assess Linguistic Complexity*. Em: BAECHLER, R.; SEILER, G. (Eds.). **Complexity, Isolation, and Variation**. [s.l.] De Gruyter, 2016. p. 71–94.

EIGE. **European Institute for Gender Equality**. European Institute for Gender Equality, 2024. Disponível em: <<https://eige.europa.eu/gender-based-violence/what-is-gender-based-violence>>. Acesso em: 6 abr. 2024

EISENSTEIN, J. **Introduction to Natural Language Processing**. [s.l.] The MIT Press, 2019.

ELLIOT, N.; KLOBUCAR, A. Handbook of automated essay evaluation: Current applications and new directions. Em: SHERMIS, M. D.; BURSTEIN, J. (Eds.). [s.l.] Routledge/Taylor & Francis Group, 2013. p. 16–35.

ERDOĞAN, Y. E.; NARIN, A. COVID-19 detection with traditional and deep features on cough acoustic signals. **Computers in Biology and Medicine**, v. 136, p. 104765, 2021.

ERMAKOVA, L.; COSSU, J. V.; MOTHE, J. A survey on evaluation of summarization methods. **Information processing & management**, v. 56, n. 5, p. 1794–1814, 2019.



- ESTRADA, E. **The Structure of Complex Networks: Theory and Applications.** [s.l.] Oxford University Press, 2011.
- ESTRELLA, P.; POPESCU-BELIS, A.; KING, M. **The FEMTI guidelines for contextual MT evaluation: principles and resources.** Em: WALTER DAELEMANS; VÉRONIQUE HOSTE (Eds.). **Evaluation of translation Technology.** Linguistica Antverpiensia new Series- themes em Translation Studies. [s.l.: s.n.].
- ETZIONI, O. et al. Unsupervised named-entity extraction from the web: An experimental study. **Artificial intelligence**, v. 165, n. 1, p. 91–134, 2005.
- Euromatrix. Survey of Machine Translation Evaluation.** [s.l.] Statistical; Hybrid Machine Translation Between All European Languages. Euromatrix, dez. 2007.
- FACELI, K. et al. **Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina.** 3. ed. Rio de Janeiro, Brasil: LTC, 2025.
- FADER, A.; SODERLAND, S.; ETZIONI, O. **Identifying Relations for Open Information Extraction.** Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. **Anais...**Edinburgh, Scotland, UK.: Association for Computational Linguistics, jul. 2011. Disponível em: <<https://www.aclweb.org/anthology/D11-1142>>
- FAGHERAZZI, G. et al. Voice for health: the use of vocal biomarkers from research to clinical practice. **Digital biomarkers**, v. 5, n. 1, p. 78–88, 2021.
- FAIR DATA MATURITY MODEL WORKING GROUP RDA. **FAIR Data Maturity Model. Specification and Guidelines.** Research Data Alliance; Zenodo, 2020. Disponível em: <<https://doi.org/10.15497/rda00050>>
- FARHANGIAN, F.; CRUZ, R. M. O.; CAVALCANTI, G. D. C. **Fake news detection: Taxonomy and comparative study.** **Information Fusion**, v. 103, p. 102140, 2024.
- FARIAS, D. S. et al. **Opinion-Meter: A Framework for Aspect-Based Sentiment Analysis.** Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web. **Anais...**2016.
- FARZINDAR, A.; INKPEN, D. **Natural Language Processing for Social Media.** Second edition ed. [s.l.] Morgan; Claypool, 2018.
- FASEEH, M. et al. Hybrid approach to automated essay scoring: Integrating deep learning embeddings with handcrafted linguistic features for improved accuracy. **Mathematics**, v. 12, n. 21, p. 3416, 2024.
- FAUSTINI, P. H. A.; COVÕES, T. F. **Fake news detection in multiple platforms and languages.** **Expert Systems with Applications**, v. 158, p. 113503, 2020.
- FAUSTINI, P.; COVÕES, T. F. **Fake News Detection Using One-Class Classifica-**



tion. Proceedings of the 8th Brazilian Conference on Intelligent Systems (BRACIS'19). **Anais...**Salvador, BA, Brazil: IEEE, out. 2019.

FAWCETT, T. An introduction to ROC analysis. **Pattern recognition letters**, v. 27, n. 8, p. 861–874, 2006.

FEDERICO, M. et al. **Assessing the Impact of Translation Errors on Machine Translation Quality with Mixed-effects Models**. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). **Anais...**Doha, Qatar: Association for Computational Linguistics, out. 2014. Disponível em: <<https://aclanthology.org/D14-1172>>

FELLBAUM, C. **WordNet: An Electronic Lexical Database**. [s.l.] The MIT Press, 1998.

FELTRIM, V. D. et al. **A Construção de uma Ferramenta de Auxílio à Escrita de Resumos Acadêmicos em Português**. Anais do Encontro Nacional de Inteligência Artificial (ENIA'2003). **Anais...**SBC, 2003.

FELTRIN, G. R.; VIANNA, D.; SILVA, A. DA. **Um Estudo Sobre Métricas de Avaliação para Sumarização de Acórdãos**. Anais do XXXVIII Simpósio Brasileiro de Bancos de Dados. **Anais...**SBC, 2023.

FENNELLY, O. et al. Use of standardized terminologies in clinical practice: A scoping review. **Int J Med Inform**, v. 149, p. 104431, fev. 2021.

FERNANDES, J. M.; WON, M.; MARTINS, B. **Speechmaking and the Selectorate: Persuasion in Nonpreferential Electoral Systems**. **Comparative Political Studies**, v. 53, n. 5, p. 667–699, a2020.

FERNANDES, W. P. D. et al. **Appellate court modifications extraction for Portuguese**. **Artificial Intelligence and Law**, v. 28, n. 3, p. 327–360, b2020.

FERNANDES-SVARTMAN, F. et al. **Temporal prosodic cues for COVID-19 in Brazilian Portuguese speakers**. Proc. Speech Prosody 2022. **Anais...**2022.

FERRÁNDEZ, Ó. et al. Tackling HAREM's portuguese named entity recognition task with spanish resources. **Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área**. **Linguatca** (http://www.linguatca.pt/aval_conjunta/LivroHAREM/Cap11-SantosCardoso2007-Ferrandezetal.pdf), 2007.

FERREIRA, A. C. et al. Padrões linguísticos para detecção de ironia em múltiplos idiomas. **Revista Gestão & Tecnologia**, 2017.

FERREIRA MELLO, R. et al. **Towards automated content analysis of rhetorical structure of written essays using sequential content-independent features in Portuguese**. (A. F. Wise, R. Martinez-Maldonado, I. Hilliger, Eds.)LAK22 Conference



- Proceedings. **Anais...**United States of America: Association for Computing Machinery (ACM), 2022.
- FERREIRA, R. et al. Towards Automatic Content Analysis of Rhetorical Structure in Brazilian College Entrance Essays. Em: [s.l: s.n.]. p. 162–167.
- FERREIRA-PAIVA, L. et al. **A survey of data augmentation for audio classification**. Congresso Brasileiro de Automática-CBA. **Anais...**2022.
- FIAD, R. S. Reescrita, dialogismo e etnografia. **Linguagem em (Dis) curso**, v. 13, p. 463–480, 2013.
- FIELDS, J.; CHOVANEC, K.; MADIRAJU, P. A survey of text classification with transformers: How wide? how large? how long? how accurate? how expensive? how safe? **IEEE Access**, v. 12, p. 6518–6531, 2024.
- FILLMORE, C. J. Frame semantics. Em: KOREA, T. L. S. OF (Ed.). **Linguistics in the Morning Calm**. Seoul: Hanshin, 1982.
- FILLMORE, C. J. Frames and the Semantics of Understanding. **Quaderni di Semantica**, v. 6, n. 2, p. 222–254, 1985.
- FILLMORE, C. J.; JOHNSON, C. R.; PETRUCK, M. R. L. **Background to FrameNet**. **International Journal of Lexicography**, v. 16, n. 3, p. 235–250, 2003.
- FINATTO, M. J. B. Projeto PorPopular, frequência de verbos em português e no jornal popular brasileiro. Em: UFMS/LABORATÓRIO DE EDIÇÃO DA FALE-UFGM, E. DA (Ed.). **As Ciências do Léxico: lexicologia, lexicografia, terminologia**. 1. ed. [s.l.] Aparecida Negri Isquendo; Maria Cândida Trindade da Costa de Seabra, 2012. v. VI. p. 227–244.
- FINATTO, M. J. B. **Humanidades digitais e estudos históricos do léxico**. **Domínios de Linguagem**, v. 17, p. e1769, 2023.
- FINATTO, M. J. B.; ESTEVES, F. F.; VILLAR, G. S. **Construindo uma terminologia de raiz: textos legislativos sob exploração terminológica**. **Revista Platô**, v. 5, n. 9, 2022.
- FINATTO, M. J. B.; PARAGUASSU, L. B. Acessibilidade textual e terminológica. 2022.
- FINATTO, M. J.; GONÇALVES, M. F.; LAZZARI, R. Léxico e terminologia em um novo gênero textual do século XVIII: o manual para enfermeiros. In: **Natalia Terrón Vinagre & Jenny Brumme (orgs.) Emergencia de nuevos géneros textuales y terminología en la historia de los lenguajes de especialidad.**, 2023.
- FINE, K. Truthmaker semantics. **A Companion to the Philosophy of Language**, p. 556–577, 2017.
- FINGER, M. Técnicas de otimização da precisão empregadas no etiquetador Tycho Brahe.



Proceedings of the International Conference on the Computational treatment of Portuguese, PROPOR, 2000.

FINGER, M. Inteligência Artificial e os rumos do processamento do português brasileiro. **Estudos Avançados**, v. 35, p. 51–72, 2021.

FINGER, M. et al. **SPIRA-BM: Biomarkers for Respiratory Conditions by Audio Analysis via Artificial Intelligence**. Anais do XXV Simpósio Brasileiro de Computação Aplicada à Saúde. **Anais...**Porto Alegre, RS, Brasil: SBC, 2025. Disponível em: <<https://sol.sbc.org.br/index.php/sbcas/article/view/35566>>

FIRDAUS SOLIHIN, R. F. A., Indra Budi; MAKARIM, E. **Advancement of information extraction use in legal documents**. **International Review of Law, Computers & Technology**, v. 35, n. 3, p. 322–351, 2021.

FIRTH, J. R. A synopsis of linguistic theory, 1930-1955. Em: FIRTH, J. R. (Ed.). **Studies in Linguistic Analysis**. Oxford: Blackwell, 1957.

FISCHER, M. et al. **Identifying Fake News in Brazilian Portuguese**. Natural Language Processing and Information Systems: 27th International Conference on Applications of Natural Language to Information Systems, NLDB 2022, Valencia, Spain, June 15–17, 2022, Proceedings. **Anais...**Berlin, Heidelberg: Springer-Verlag, 2022. Disponível em: <https://doi.org/10.1007/978-3-031-08473-7_10>

FLORES, F. N.; MOREIRA, V. P.; HEUSER, C. A. **Assessing the impact of stemming accuracy on information retrieval**. International Conference on Computational Processing of the Portuguese Language. **Anais...**Springer, 2010.

FLORIAN, R. et al. **Named entity recognition through classifier combination**. Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003. **Anais...**2003.

FONSECA, A.; SADAT, F.; MASSÉ, A. B. **Identifying portuguese multiword expressions using different classification algorithms-a comparative analysis**. Proceedings of the 4th International Workshop on Computational Terminology (Computerm). **Anais...**2014.

FONSECA, E. R. et al. **Automatically Grading Brazilian Student Essays**. (A. Villavicencio et al., Eds.)Computational Processing of the Portuguese Language. **Anais...**Springer International Publishing, 2018.

FONT LLITJÓS, A.; CARBONELL, J. G.; LAVIE, A. **A framework for interactive and automatic refinement of transfer-based machine translation**. Proceedings of the 10th EAMT Conference: Practical applications of machine translation. **Anais...**Budapest, Hungary: European Association for Machine Translation, 2005. Disponível em: <<https://aclanthology.org/2005.eamt-1.13>>

FORCADA, M. L.; ÑECO, R. P. **Recursive hetero-associative memories for trans-**



- lation. International Work-Conference on Artificial Neural Networks. **Anais...**Springer, 1997.
- FORNACIARI, T.; POESIO, M. Automatic deception detection in Italian court cases. **Artif. Intell. Law**, v. 21, n. 3, p. 303–340, set. 2013.
- FORTUNA, P. et al. **A Hierarchically-Labeled Portuguese Hate Speech Dataset**. Proceedings of the Third Workshop on Abusive Language Online. **Anais...**2019.
- FORTUNA, P.; NUNES, S. A survey on automatic detection of hate speech in text. **ACM Computing Surveys (CSUR)**, 2018.
- FOVE. **Fove Eye Tracker.**, 2018. Disponível em: <<https://www.getfove.com/>>
- FREITAS ARAUJO, A. V. DE. **Informação, comunicação e documento**. Brasília, DF; Rio de Janeiro, RJ: CAPES: UAB; Departamento de Biblioteconomia, FACC/UFRJ, 2018. p. 96
- FREITAS, B. L.; SILVA, R. M.; ALMEIDA, T. A. Gaussian Mixture Descriptors Learner. **KNOWLEDGE-BASED SYSTEMS**, v. 188, p. 105039, 2019.
- FREITAS, C. et al. **Vampiro que brilha... rá! Desafios na anotação de opinião em um corpus de resenhas de livros**. Proceedings of XI Encontro de Linguística de Corpus. **Anais...**2012.
- FREITAS, C. Sobre a construção de um léxico da afetividade para o processamento computacional do português. **Revista Brasileira de Linguística Aplicada**, 2013.
- FREITAS, L. A. DE et al. **Pathways for irony detection in tweets**. Proceedings of the Symposium on Applied Computing (SAC). **Anais...**2014.
- FREITAS, L. A. DE. **Feature-level sentiment analysis applied to brazilian portuguese reviews**. tese de doutorado—[s.l.] Pontifícia Universidade Católica do Rio Grande do Sul, 2015.
- FREITAS, L. A. DE; SANTOS, L. DOS; DEON, D. Padrões linguísticos para detecção de ironia em múltiplos idiomas. **Revista Eletrônica de Iniciação Científica em Computação**, 2020.
- FULLER, C. et al. **An Analysis of Text-Based Deception Detection Tools**. Proceedings of the Twelfth Americas Conference on Information Systems. **Anais...**2006.
- GALHARDI, C. P. et al. **Fato ou Fake? Uma análise da desinformação frente à pandemia da COVID-19 no Brasil**. **Ciência & Saúde Coletiva**, v. 25, p. 4201–4210, out. 2020.
- GAMALLO, P.; GARCIA, M. **Multilingual open information extraction**. (F. Pereira et al., Eds.)Portuguese Conference on Artificial Intelligence. **Anais...**Cham: Springer; Springer International Publishing, 2015. Disponível em: <<https://doi.org/10.1007/978-3->



319-23485-4_72>

GAMALLO, P.; GARCIA, M.; FERNÁNDEZ-LANZA, S. **Dependency-based open information extraction**. Proceedings of the joint workshop on unsupervised and semi-supervised learning in NLP. *Anais...: ROBUS-UNSUP '12*. Stroudsburg, PA, USA: Association for Computational Linguistics; Association for Computational Linguistics, 2012. Disponível em: <<http://dl.acm.org/citation.cfm?id=2389961.2389963>>

GAMBHIR, M.; GUPTA, V. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, v. 47, p. 1–66, 2017.

GAMON, M. et al. Handbook of automated essay evaluation: Current applications and new directions. Em: SHERMIS, M. D.; BURSTEIN, J. (Eds.). [s.l.] Routledge/Taylor & Francis Group, 2013. p. 251–266.

GAMONAL, M. A. **Copa 2014 FrameNet Brasil: Diretrizes para a Constituição de um Dicionário Eletrônico Trilíngue a partir da Análise de Frames da Experiência Turística**. Dissertação de Mestrado em Linguística—[s.l.] Universidade Federal de Juiz de Fora, 2013.

GAMONAL, M. A. **Modelagem linguístico-computacional de metonímias na base de conhecimento multilíngue (m.knob) da FrameNet Brasil**. Tese de Doutorado—Juiz de Fora, Brasil: Universidade Federal de Juiz de Fora, 2017.

GAMONAL, M. A. et al. **Audition: A Frame-Annotated Multimodal Dataset for Accessible Audiovisual Content**. 2025.

GAO, M. et al. **Human-like Summarization Evaluation with ChatGPT.**, 2023. Disponível em: <<https://arxiv.org/abs/2304.02554>>

GARBIN, C. A. S. et al. Desafios do profissional de saúde na notificação da violência: obrigatoriedade, efetivação e encaminhamento. *Ciência & Saúde Coletiva*, v. 20, p. 1879–1890, 2015.

GARCIA, E. A. S. et al. **RoBERTaLexPT: A Legal RoBERTa Model pretrained with deduplication for Portuguese**. (P. Gamallo et al., Eds.) Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1. *Anais...Santiago de Compostela, Galicia/Spain: Association for Computational Linguistics*, mar. a2024. Disponível em: <<https://aclanthology.org/2024.propor-1.38>>

GARCIA, G. L. et al. **Text Summarization and Temporal Learning Models Applied to Portuguese Fake News Detection in a Novel Brazilian Corpus Dataset**. (P. Gamallo et al., Eds.) Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1. *Anais...Santiago de Compostela, Galicia/Spain: Association for Computational Linguistics*, mar. b2024. Disponível em: <<https://aclanthology.org/2024.propor-1.9>>

GARCIA, G. L.; AFONSO, L. C.; PAPA, J. P. **FakeRecogna: a new brazilian corpus**



- for fake news detection.** International Conference on Computational Processing of the Portuguese Language. **Anais...**Springer, 2022.
- GARCIA-MOLINA, H. et al. Challenges in data crowdsourcing. **IEEE Transactions on Knowledge and Data Engineering**, v. 28, n. 4, p. 901–911, 2016.
- GARCIA-MORENO, C.; WATTS, C. Violence against women: an urgent public health priority. **Bulletin of the World Health Organization**, v. 89, p. 2–2, 2011.
- GARIJO, D.; POVEDA-VILLALÓN, M. **Best Practices for Implementing FAIR Vocabularies and Ontologies on the Web.** **CoRR**, v. abs/2003.13084, 2020.
- GAUY, M. M. et al. **Discriminant Audio Properties In Deep Learning Based Respiratory Insufficiency Detection In Brazilian Portuguese.** Artificial Intelligence in Medicine: 21st International Conference on Artificial Intelligence in Medicine, AIME 2023, Portorož, Slovenia, June 12–15, 2023, Proceedings. **Anais...**Berlin, Heidelberg: Springer-Verlag, 2023. Disponível em: <https://doi.org/10.1007/978-3-031-34344-5_32>
- GAUY, M. M.; FINGER, M. Acoustic models of Brazilian Portuguese Speech based on Neural Transformers. **arXiv preprint arXiv:2312.09265**, 2023.
- GAUY, M.; FINGER, M. **Audio MFCC-gram Transformers for respiratory insufficiency detection in COVID-19.** Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. **Anais...**Porto Alegre, RS, Brasil: SBC, 2021. Disponível em: <<https://sol.sbc.org.br/index.php/stil/article/view/17793>>
- GAZZOLA, M.; LEAL, S. E.; ALUISIO, S. M. **Predição da Complexidade Textual de Recursos Educacionais Abertos em Português.** Proceedings of the Brazilian Symposium in Information and Human Language Technology. **Anais...**2019.
- GEMMEKE, J. F. et al. **Audio set: An ontology and human-labeled dataset for audio events.** 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). **Anais...**IEEE, 2017.
- GEORGE, J. F.; KEANE, B. T. **Deception Detection by Disinterested Third-Party Observers.** Proceedings of the Credibility Assessment and Information Quality in Government and Business Symposium, 39th Hawaii International Conference on System Sciences (HICSS). **Anais...**Kauai, HI: 2006.
- GEURGAS, R.; TESSLER, L. R. **Automatic detection of fake tweets about the COVID-19 Vaccine in Portuguese.** **Social Network Analysis and Mining**, v. 14, n. 1, p. 55, 8 mar. 2024.
- GHANEM, B. et al. **IDAT at FIRE2019: Overview of the Track on Irony Detection in Arabic Tweets.** Proceedings of the 11th Forum for Information Retrieval Evaluation. **Anais...**2019.
- GHOSH, A. et al. **SemEval-2015 Task 11: Sentiment Analysis of Figurative**



Language in Twitter. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). **Anais...**2015.

GIBBS, R. W.; COLSTON, H. L. **The Risks and Rewards of Ironic Communication.** Say not to say: new perspectives on miscommunication. **Anais...**2001. Disponível em: <<https://api.semanticscholar.org/CorpusID:12510370>>

GILLIN, N. **One-shot Prompt for Language Variety Identification.** Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024). **Anais...**2024.

GLAUBER, R. et al. **Challenges of an Annotation Task for Open Information Extraction in Portuguese.** (A. Villavicencio et al., Eds.)Computational Processing of the Portuguese Language. **Anais...**Cham: Springer International Publishing, 2018.

GLAUBER, R.; CLARO, D. B. **A systematic mapping study on open information extraction.** **Expert Systems with Applications**, v. 112, p. 372–387, 2018.

GLAUBER, R.; CLARO, D. B.; OLIVEIRA, L. S. **Dependency Parser on Open Information Extraction for Portuguese Texts - DptOIE and DependentIE on IberLEF.** Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) co-located with 35th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019). **Anais...**<http://ceur-ws.org/Vol-2421/>: CEUR Workshop Proceedings, a2019.

GLAUBER, R.; CLARO, D. B.; SENA, C. F. DE L. **Towards a Pragmatic Open Information Extraction for Portuguese Text - ICEIS17, InferPortOIE and PragmaticOIE on IberLEF.** Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) co-located with 35th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019). **Anais...**<http://ceur-ws.org/Vol-2421/>: CEUR Workshop Proceedings, b2019.

GÔLO, M. P. S. et al. **One-class learning for fake news detection through multimodal variational autoencoders.** **Engineering Applications of Artificial Intelligence**, v. 122, p. 106088, 2023.

GÔLO, M. P. S. et al. **On the use of Large Language Models to Detect Brazilian Politics Fake News.** Proceedings of the 21st National Meeting on Artificial and Computational Intelligence (ENIAC'2024). **Anais...**Belém, PA, Brazil: Brazilian Computer Society, nov. 2024.

GOMES, J. R. S. et al. **Deep Learning Brasil at ABSAPT 2022: Portuguese Transformer Ensemble Approaches.** Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2022), A Coruña, Spain, September 20, 2022. **Anais...**2022.

GONÇALO OLIVEIRA, H. et al. **Avaliação à medida no Segundo HAREM.** (C. Mota, D. Santos, Eds.)Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. **Anais...**Linguatca, 2008.



GONÇALVES, T. et al. Clinical Screening Prediction in the Portuguese National Health Service: Data Analysis, Machine Learning Models, Explainability and Meta-Evaluation. **Future Internet**, v. 15, n. 1, p. 26, 2023.

GONÇALVES, T.; QUARESMA, P. **A preliminary approach to the multilabel classification problem of Portuguese juridical documents**. Portuguese Conference on Artificial Intelligence. **Anais...**Springer, 2003.

GORDEEFF, E. M. **Avaliação sobre Animação e Cinema de vida real: semelhanças e diferenças**. **Diálogo com a Economia Criativa**, v. 8, n. 24, p. 50–63, 2023.

GORMAN, K.; BEDRICK, S. **We Need to Talk about Standard Splits**. (A. Korhonen, D. Traum, L. Màrquez, Eds.) Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. **Anais...**Florence, Italy: Association for Computational Linguistics, jul. 2019. Disponível em: <<https://aclanthology.org/P19-1267/>>

GOU, J. et al. Knowledge distillation: A survey. **International journal of computer vision**, v. 129, n. 6, p. 1789–1819, 2021.

GOURISARIA, M. K. et al. Comparative analysis of audio classification with MFCC and STFT features using machine learning techniques. **Discover Internet of Things**, v. 4, n. 1, p. 1, 2024.

GRAESSER, A. C. et al. Coh-Metrix: Analysis of text on cohesion and language. **Behavior Research Methods, Instruments, n Computer - Springer**, p. 193–202, 2004.

GRAESSER, A. C.; MCNAMARA, D. S.; KULIKOWICH, J. M. Coh-Metrix: Providing Multilevel Analyses of Text Characteristics. **Educational Researcher Vol. 40, N. 5**, p. 223–234, 2011.

GRAHAM, Y. et al. **Is all that Glitters in Machine Translation Quality Estimation really Gold?** Proceedings of COLING 2016: Technical Papers. **Anais...**Osaka, Japan: The COLING 2016 Organizing Committee, dez. 2016. Disponível em: <<https://www.aclweb.org/anthology/C16-1294>>

GRATTAFIORI, A. et al. **The Llama 3 Herd of Models.**, 2024. Disponível em: <<https://arxiv.org/abs/2407.21783>>

GRAVES, A.; MOHAMED, A.; HINTON, G. **Speech recognition with deep recurrent neural networks**. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. **Anais...**2013.

GREENBERG, J. H. Some universals of grammar with particular reference to the order of meaningful elements. Em: GREENBERG, J. H. (Ed.). **Universals of Grammar**. 2. ed. Cambridge, Mass.: MIT Press, 1966. p. 73–113.

GRICE, H. P. Logic and Conversation. Em: **Syntax and Semantics: Vol. 3: Speech Acts**. [s.l.] Academic Press, 1975.



GRISHMAN, R.; SUNDHEIM, B. **Message Understanding Conference- 6: A Brief History**. COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics. **Anais...**1996. Disponível em: <<https://aclanthology.org/C96-1079>>

GRUBER, A.; WEISS, Y.; ROSEN-ZVI, M. **Hidden Topic Markov Models**. Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics. **Anais...**: Proceedings of Machine Learning Research. San Juan, Puerto Rico: PMLR, mar. 2007.

GRUPPI, M.; HORNE, B. D.; ADALI, S. **NELA-GT-2019: A Large Multi-Labelled News Dataset for The Study of Misinformation in News Articles**. **CoRR**, v. abs/2003.08444, p. 1–5, 2020.

GRUPPI, M.; HORNE, B. D.; ADALI, S. **NELA-GT-2020: A Large Multi-Labelled News Dataset for The Study of Misinformation in News Articles**. **CoRR**, v. abs/2102.04567, p. 1–6, 2021.

GU, J. et al. **A Survey on LLM-as-a-Judge.**, 2025. Disponível em: <<https://arxiv.org/abs/2411.15594>>

GUARINO, N.; GUIZZARDI, G. **We need to Discuss the Relationship: Revisiting Relationships as Modeling Constructs**. Proceedings of the 27th International Conference on Advanced Information Systems Engineering (CAISE 2015). **Anais...**Springer-Verlag, 2015.

GUIMARÃES, G. M. C. et al. **Legal Document Segmentation and Labeling Through Named Entity Recognition Approaches**. **Journal of Information and Data Management**, v. 15, n. 1, 2024.

GUIMARÃES, J. A. C.; SANTOS, J. C. G. **A ementa jurisprudencial como resumo informativo em um domínio especializado: aspectos estruturais**. **Brazilian Journal of Information Science: research trends**, v. 10, n. 3, 2016.

GUIMARÃES, S. S. et al. **Characterizing Toxicity on Facebook Comments in Brazil**. Proceedings of the Brazilian Symposium on Multimedia and the Web. **Anais...**2020.

GUIZZARDI, G. **Ontology, Ontologies and the “I” of FAIR**. **Data Int.**, v. 2, n. 1-2, p. 181–191, 2020.

GULDEN, C. et al. **Extractive summarization of clinical trial descriptions**. **International Journal of Medical Informatics**, v. 129, p. 114–121, 2019.

GUMIEL, Y. B. et al. **Temporal Relation Extraction in Clinical Texts: A Systematic Review**. v. 54, n. 7, set. 2021.

GUPTA, S.; GUPTA, S. **Abstractive summarization: An overview of the state of the art**. **Expert Systems with Applications**, v. 121, p. 49–65, 2019.



- HABIBI, M. et al. Deep learning with word embeddings improves biomedical named entity recognition. **Bioinformatics**, v. 33, n. 14, p. i37–i48, 2017.
- HAENDCHEN FILHO, A. et al. **An approach to evaluate adherence to the theme and the argumentative structure of essays**. International Conference on Knowledge-Based Intelligent Information & Engineering Systems. **Anais...**2018.
- HAENDCHEN FILHO, A. et al. Imbalanced Learning Techniques for Improving the Performance of Statistical Models in Automated Essay Scoring. **Procedia Computer Science**, v. 159, p. 764–773, jan. 2019.
- HAENLEIN, M.; KAPLAN, A. A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. **California management review**, v. 61, n. 4, p. 5–14, 2019.
- HAILU, T. T.; YU, J.; FANTAYE, T. G. A framework for word embedding based automatic text summarization and evaluation. **Information**, v. 11, n. 2, p. 78, 2020.
- HAKUTA, K. Handbook of Automated Essay Evaluation: Current Applications and New Directions. Em: SHERMIS, M. D.; BURSTEIN, J. (Eds.). [s.l.] Routledge/Taylor & Francis Group, 2013. p. 347–353.
- HAN, J. et al. **Exploring automatic COVID-19 diagnosis via voice and symptoms from crowdsourced data**. ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). **Anais...IEEE**, 2021.
- HANLEY, J. A.; MCNEIL, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. **Radiology**, v. 143, n. 1, p. 29–36, 1982.
- HARTMANN, N. S. et al. **Portuguese word embeddings: Evaluating on word analogies and natural language tasks**. Proceedings of Symposium in Information and Human Language Technology. **Anais...**[S.l.: s.n.]: 2017.
- HARTMANN, N. S.; ALUÍSIO, S. M. **Adaptação Lexical Automática em Textos Informativos do Português Brasileiro para o Ensino Fundamental**. **Linguamática**, v. 12, n. 2, p. 3–27, dez. 2020.
- HARTMANN PEIXOTO, F. **Projeto Victor: relato do desenvolvimento da Inteligência Artificial na Repercussão Geral do Supremo Tribunal Federal**. **Revista Brasileira de Inteligência Artificial e Direito - RBIAD**, v. 1, n. 1, p. 1–22, 2020.
- HASEGAWA, T.; SEKINE, S.; GRISHMAN, R. **Discovering relations among named entities from large corpora**. Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (acl-04). **Anais...**2004.
- HASSAN, A.; SHAHIN, I.; ALSABEK, M. B. **COVID-19 detection system using recurrent neural networks**. 2020 International conference on communications, computing, cybersecurity, and informatics (CCCI). **Anais...IEEE**, 2020.



- HASSAN, H. et al. Achieving Human Parity on Automatic Chinese to English News Translation. **arXiv preprint 1803.05567**, 2018.
- HAUCH, V. et al. **Linguistic Cues to Deception Assessed by Computer Programs: A Meta-analysis**. Proceedings of the Workshop on Computational Approaches to Deception Detection. **Anais...**2012.
- HE, K. et al. **Masked autoencoders are scalable vision learners**. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. **Anais...**2022.
- HEARST, M. A. **Automatic acquisition of hyponyms from large text corpora**. Proceedings of the 14th conference on Computational linguistics-Volume 2. **Anais...**Association for Computational Linguistics, 1992.
- HEE, C. V.; LEFEVER, E.; HOSTE, V. **SemEval-2018 Task 3: Irony Detection in English Tweets**. Proceedings of the 12th International Workshop on Semantic Evaluation. **Anais...**2018.
- HEINRICH, T.; MARCHI, F. **TeamUFPR at ABSAPT 2022: Aspect Extraction with CRF and BERT**. Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2022), A Coruña, Spain, September 20, 2022. **Anais...**2022.
- HEUSDEN, R. VAN; KAMPS, J.; MARX, M. **Neural Coreference Resolution for Dutch Parliamentary Documents with the DutchParliament Dataset**. **Data**, v. 8, n. 2, 2023.
- HITZLER, A. H. Y. A. M. S. M. A. G. B. A. W. R. A. A. S. A. A. H. M. A. K. T. A. J. M. M. A. A. M. A. J. P. A. P. **Multimodal mental health analysis in social media**. **PLOS ONE**, v. 15, n. 4, p. 1–27, 2020.
- HOCHREITER, S.; SCHMIDHUBER, J. **Long Short-Term Memory**. **Neural Computation**, v. 9, n. 8, p. 1735–1780, nov. 1997.
- HOCKETT, C. F. **A Course in Modern Linguistics**. [s.l.] Macmillan, 1958.
- HOFMANN, T. **Probabilistic Latent Semantic Indexing**. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99). **Anais...**New York, NY, USA: Association for Computing Machinery, 1999.
- HOVY, E.; KING, M.; POPESCU-BELIS, A. **An introduction to MT evaluation**. Proceedings of Machine Translation Evaluation: Human Evaluators meet Automated Metrics. Workshop at the LREC 2002 Conference. Las Palmas, Spain. **Anais...**2002.
- HSU, W.-N. et al. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, v. 29, p. 3451–3460, 2021.



- HU, M.; LIU, B. **Mining Opinion Features in Customer Reviews**. Proceedings of the 19th National Conference on Artificial Intelligence. **Anais...**2004.
- HUA, Y.; DENG, Z.; MCKEOWN, K. **Improving Long Dialogue Summarization with Semantic Graph Representation**. (A. Rogers, J. Boyd-Graber, N. Okazaki, Eds.) Findings of the Association for Computational Linguistics: ACL 2023. **Anais...**Toronto, Canada: Association for Computational Linguistics, jul. 2023. Disponível em: <<https://aclanthology.org/2023.findings-acl.871/>>
- HUANG, P.-Y. et al. Masked autoencoders that listen. **Advances in Neural Information Processing Systems**, v. 35, p. 28708–28720, 2022.
- HUSSAIN, A. S.; THOMAS, A. **Large Language Models for Judicial Entity Extraction: A Comparative Study**., 2024. Disponível em: <<https://arxiv.org/abs/2407.05786>>
- HUTCHINS, J. **Towards a definition of example-based machine translation**., Proceedings of Second Workshop on Example-Based Machine Translation; **Anais...**2005.
- HUTCHINS, W. Machine Translation: A Concise History. **Journal of Translation Studies: Special Issue on The Teaching of Computer-aided Translation**, v. 13, p. 1–2, 2010.
- HUTCHINS, W. J. **Machine translation over fifty years**. **Histoire, Epistemologie, Langage**, v. XXII, n. 1, p. 7–31, 2001.
- IFTIKHAR, A.; UL QOUNAIN JAFFRY, S. W.; MALIK, M. K. **Information Mining From Criminal Judgments of Lahore High Court**. **IEEE Access**, v. 7, p. 59539–59547, 2019.
- IMOTIONS. **Eye Tracking - The Complete Pocket Guide**. [s.l.] www.imotions.com, 2017.
- INÁCIO, M. L.; CASELI, H. DE M. **Word Embeddings at Post-Editing**. (P. Quaresma et al., Eds.) Computational Processing of the Portuguese Language. **Anais...**Cham: Springer International Publishing, 2020.
- INDYK, P.; MOTWANI, R. **Approximate nearest neighbors: towards removing the curse of dimensionality**. Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing (STOC). **Anais...**Dallas, Texas, USA: ACM, 1998.
- IOFFE, S.; SZEGEDY, C. **Batch normalization: Accelerating deep network training by reducing internal covariate shift**. International conference on machine learning. **Anais...**PMLR, 2015.
- IPM. **INAF Brasil 2018: Indicador de Alfabetismo Funcional - Resultados Preliminares**. Instituto Paulo Montenegro, 2018.



- JACOBS, R. A. et al. Adaptive mixtures of local experts. **Neural computation**, v. 3, n. 1, p. 79–87, 1991.
- JACOBSEN, A. et al. **FAIR principles: interpretations and implementation considerations**. Data intelligence MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info ..., 2020.
- JAHAN, M. S.; OUSSALAH, M. A systematic review of hate speech automatic detection using natural language processing. **Neurocomputing**, 2023.
- JARGAS, A. M. **Expressões Regulares - 5a edição: Uma Abordagem Divertida**. [s.l.] Novatec Editora, 2016.
- JÄRVELIN, K.; KEKÄLÄINEN, J. Cumulated gain-based evaluation of IR techniques. **ACM Transactions on Information Systems (TOIS)**, v. 20, n. 4, p. 422–446, 2002.
- JERONIMO, C. et al. **Characterization of Fake News Based on Subjectivity Lexicons**. **Journal of Data Intelligence**, v. 1, p. 419–441, dez. 2020.
- JIANG, A. Q. et al. **Mistral 7B.**, 2023. Disponível em: <<https://arxiv.org/abs/2310.06825>>
- JIANG, S. et al. **Multi-Ontology Refined Embeddings (MORE): A hybrid multi-ontology and corpus-based semantic representation model for biomedical concepts**. **Journal of Biomedical Informatics**, v. 111, p. 103581, 2020.
- JIANG, S. et al. **Irony Detection in the Portuguese Language using BERT**. Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain, September, 2021. **Anais...**2021.
- JOHNSTONE, M. A.-M. A. T. **In an Absolute State: Elevated Use of Absolutist Words Is a Marker Specific to Anxiety, Depression, and Suicidal Ideation**. **Clinical Psychological Science**, v. 6, n. 4, p. 529–542, 2018.
- JONES, K. H. et al. Toward the Development of Data Governance Standards for Using Clinical Free-Text Data in Health Research: Position Paper. **J Med Internet Res**, v. 22, n. 6, p. e16760, jun. 2020.
- JONES, K. S. What might be in a summary? **Information retrieval**, v. 93, n. 1, p. 9–26, 1993.
- JONNALAGADDA, S.; GONZALEZ, G. Biosimplify: an open source sentence simplification engine to improve recall in automatic biomedical information extraction. **AMIA Annual Symposium Proceedings**, p. 351–356, 2010.
- JULIÃO, A. **Algoritmo do Google: Veja o impacto que tem no SEO**. Disponível



em: <<https://blog.ajestrategia.com.br/algorithm-do-google-veja-o-impacto-que-tem-no-seo/>>.

JUOLA, P. **Measuring Linguistic Complexity: The Morphological Tier**. *Journal of Quantitative Linguistics*, v. 5, n. 3, p. 206–213, dez. 1998.

JUOLA, P. **Assessing Linguistic Complexity**. Em: MIESTAMO, M.; SINNEMÄKI, K.; KARLSSON, F. (Eds.). *Studies in Language Companion Series*. Amsterdam: John Benjamins Publishing Company, 2008. v. 94p. 89–108.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition**. 3rd. ed. USA: Prentice Hall PTR, 2023.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition**. Em: 3rd. ed. [s.l.] Stanford University, 2026.

JUSTIÇA - CNJ, C. N. DE. **Conselho Nacional de Justiça — Justiça em Números**. <https://www.cnj.jus.br/pesquisas-judiciarias/justica-em-numeros/>, maio 24DC.

KAHLE, P. et al. **Transkribus-a service platform for transcription, recognition and retrieval of historical documents**. 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). *Anais...IEEE*, 2017.

KAMBHATLA, N. **Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction**. Proceedings of the ACL interactive poster and demonstration sessions. *Anais...2004*.

KAPETANIDIS, P. et al. Respiratory diseases diagnosis using audio analysis and artificial intelligence: a systematic review. *Sensors*, v. 24, n. 4, p. 1173, 2024.

KAPOOR, A. et al. **HLDC: Hindi Legal Documents Corpus**. Findings of the Association for Computational Linguistics: ACL 2022. *Anais...Association for Computational Linguistics*, 2022.

KARELITZ, J. L.; MICHAEL, V. C.; PERKINS, K. A. **Analysis of agreement between expired-air carbon monoxide monitors**. *Journal of smoking cessation*, v. 2, n. 12, p. 105–112, 2017.

KATCHAPAKIRIN, K. et al. **Facebook Social Media for Depression Detection in the Thai Community**. 15th International Joint Conference on Computer Science and Software Engineering (JCSSE). *Anais...2018*.

KHAYRALLAH, H.; KOEHN, P. **On the Impact of Various Types of Noise on Neural Machine Translation**. Proceedings of the 2nd Workshop on Neural Machine Translation and Generation. *Anais...Melbourne, Australia: Association for Computational Linguistics*, jul. 2018. Disponível em: <<https://aclanthology.org/W18-2709>>



KILGARRIFF, A.; AL., ET. PtTenTen: A Corpus for Portuguese Lexicography. Em: **Working with Portuguese Corpora**. [s.l: s.n.]. p. 111–130.

KIM, Y. **Convolutional Neural Networks for Sentence Classification**. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). **Anais...2014**.

KINCAID, J. P. et al. Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for Navy enlisted personnel. **Research Branch Report**, p. 8–75, 1975.

KIND, L. et al. Subnotificação e (in) visibilidade da violência contra mulheres na atenção primária à saúde. **Cadernos de Saúde Pública**, v. 29, p. 1805–1815, 2013.

KOEHN, P. et al. **Moses: Open Source Toolkit for Statistical Machine Translation**. Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions. **Anais...Prague, Czech Republic: Association for Computational Linguistics**, jun. 2007. Disponível em: <<https://aclanthology.org/P07-2045>>

KOEHN, P. **Statistical Machine Translation**. [s.l.] Cambridge University Press, 2009.

KOEHN, P. **Neural Machine Translation**. [s.l.] Cambridge University Press, 2020.

KOEHN, P.; OCH, F. J.; MARCU, D. **Statistical phrase-based translation**. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03. **Anais...Association for Computational Linguistics**, 2003. Disponível em: <<http://dx.doi.org/10.3115/1073445.1073462>>

KOHAU, R. **A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection**. Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI). **Anais...1995**.

KOHO, M. et al. **WarSampo Knowledge Graph: Finland in the Second World War as Linked Open Data**. **Semantic Web – Interoperability, Usability, Applicability**, v. 12, n. 2, p. 265–278, 2021.

KOJIMA, T. et al. Large language models are zero-shot reasoners. **Advances in neural information processing systems**, v. 35, p. 22199–22213, 2022.

KOLECK, T. A. et al. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. **J Am Med Inform Assoc**, v. 26, n. 4, p. 364–379, abr. 2019.

KONG, Q. et al. **Cross-Task Learning for Audio Tagging, Sound Event Detection and Spatial Localization: DCASE 2019 Baseline Systems**. [s.l.] DCASE2019 Challenge, 2019.



KONG, Q. et al. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, v. 28, p. 2880–2894, 2020.

KONSTANTINOVA, N. **Review of relation extraction methods: What is new out there?** Analysis of Images, Social Networks and Texts: Third International Conference, AIST 2014, Yekaterinburg, Russia, April 10-12, 2014, Revised Selected Papers 3. **Anais...**Springer, 2014.

KORNILOVA, A.; EIDELMAN, V. **BillSum: A Corpus for Automatic Summarization of US Legislation**. Proceedings of the 2nd Workshop on New Frontiers in Summarization. **Anais...**Association for Computational Linguistics, 2019.

KRINGS, H. P. **Repairing Texts: Empirical Investigations of Machine Translation Post-editing Processes**. [s.l.] Kent State University Press, 2001.

KULKARNI, M. et al. **Towards a Unified Multi-Domain Multilingual Named Entity Recognition Model**. Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. **Anais...**Dubrovnik, Croatia: Association for Computational Linguistics, 2023. Disponível em: <<https://aclanthology.org/2023.eacl-main.161>>

KUMAR, D. et al. **Understanding the Behaviors of Toxic Accounts on Reddit**. Proceedings of the ACM Web Conference 2023. **Anais...**2023.

KUZI, S.; SHTOK, A.; KURLAND, O. **Query expansion using word embeddings**. Proceedings of the 25th ACM international on conference on information and knowledge management. **Anais...**2016.

LAFFERTY, J. D.; MCCALLUM, A.; PEREIRA, F. C. N. **Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data**. Proceedings of the Eighteenth International Conference on Machine Learning. **Anais...** ICML '01.San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001. Disponível em: <<https://dl.acm.org/doi/abs/10.5555/645530.655813>>

LAGE-FREITAS, A. et al. **Predicting Brazilian court decisions**. arXiv, 2019. Disponível em: <<https://arxiv.org/abs/1905.10348>>

LAGUARTA, J.; HUETO, F.; SUBIRANA, B. **COVID-19 artificial intelligence diagnosis using only cough recordings**. **IEEE Open Journal of Engineering in Medicine and Biology**, v. 1, p. 275–281, 2020.

LAMPLE, G. et al. **Neural Architectures for Named Entity Recognition**. (K. Knight, A. Nenkova, O. Rambow, Eds.)Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. **Anais...**San Diego, California: Association for Computational Linguistics, jun. 2016.



- LANCICHINETTI, A.; FORTUNATO, S. Consensus clustering in complex networks. **Scientific reports**, v. 2, n. 1, p. 336, 2012.
- LANDRY, V. et al. Audio-based digital biomarkers in diagnosing and managing respiratory diseases: a systematic review and bibliometric analysis. **European Respiratory Review**, v. 34, n. 176, 2025.
- LÄUBLI, S. et al. A set of recommendations for assessing human-machine parity in language translation. **Journal of Artificial Intelligence Research**, v. 67, p. 653–672, 2020.
- LÄUBLI, S.; SENNRICH, R.; VOLK, M. **Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation**. Proceedings of EMNLP. **Anais...**Brussels, Belgium: 2018.
- LAZER, D. M. J. et al. The science of fake news. **Science**, v. 359, n. 6380, p. 1094–1096, 2018.
- LAZZARI, R. R.; FINATTO, M. J. B. Exame do vocabulário médico no Português no século XVIII: contribuições da lexicometria para o desenho de um dicionário histórico. **Mandinga-Revista de Estudos Linguísticos (ISSN: 2526-3455)**, v. 7, n. 1, p. 102–123, 2023.
- LE, Q.; MIKOLOV, T. **Distributed representations of sentences and documents**. International conference on machine learning. **Anais...**PMLR, 2014.
- LEACOCK, C. et al. **Automated Grammatical Error Detection for Language Learners**. [s.l.] Morgan; Claypool Publishers, 2010.
- LEAL, S. E. et al. **Avaliação automática da complexidade de sentenças do português brasileiro para o domínio rural**. Symposium in Information and Human Language Technology - STIL. **Anais...**SBC, 2019.
- LEAL, S. E. et al. **Using Eye-tracking Data to Predict the Readability of Brazilian Portuguese Sentences in Single-task, Multi-task and Sequential Transfer Learning Approaches**. Proceedings of the 28th International Conference on Computational Linguistics. **Anais...**Barcelona, Spain (Online): International Committee on Computational Linguistics, dez. 2020. Disponível em: <<https://www.aclweb.org/anthology/2020.coling-main.512>>
- LEAL, S. E. **Predição da complexidade sentencial do português brasileiro escrito, usando métricas linguísticas, psicolinguísticas e de rastreamento ocular**. tese de doutorado—[s.l.] Universidade de São Paulo, 2021.
- LEAL, S. E. et al. RastrOS Project: Natural Language Processing contributions to the development of an eye-tracking corpus with predictability norms for Brazilian Portuguese. **Language Resources and Evaluation**, p. 1333–1372, 2022.



- LEAL, S. E. et al. **NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese**. **Language Resources and Evaluation**, 2023.
- LEAL, S. E.; DURAN, M. S.; ALUÍSIO, S. M. **A Nontrivial Sentence Corpus for the Task of Sentence Readability Assessment in Portuguese**. Proceedings of the 27th International Conference on Computational Linguistics. **Anais...Association for Computational Linguistics**, ago. 2018.
- LÉCHELLE, W.; GOTTI, F.; LANGLAIS, P. WiRe57: A Fine-Grained Benchmark for Open Information Extraction. **arXiv preprint arXiv:1809.08962**, 2018.
- LEE, J. et al. **BioBERT: a pre-trained biomedical language representation model for biomedical text mining**. **Bioinformatics**, v. 36, n. 4, p. 1234–1240, set. 2019.
- LEE, S. et al. **A Survey on Evaluation Metrics for Machine Translation**. **Mathematics**, v. 11, n. 4, 2023.
- LEHNERT, W.; SUNDHEIM, B. A performance evaluation of text-analysis technologies. **AI magazine**, v. 12, n. 3, p. 81–81, 1991.
- LEITÃO, M. M.; RIBEIRO, A. J. C.; MAIA, M. Penalidade do Nome Repetido e Rastreamento Ocular em Português Brasileiro. **Revista Linguística**, v. v8 n2, 2012.
- LEITE, H. et al. **WRITEME: uma Ferramenta de Auxílio à Escrita de READMEs Baseada em Dados Abertos**. Anais do XVII Congresso Latino-Americano de Software Livre e Tecnologias Abertas. **Anais...Porto Alegre, RS, Brasil: SBC**, 2020.
- LEITNER, E.; REHM, G.; MORENO-SCHNEIDER, J. **Fine-Grained Named Entity Recognition in Legal Documents**. (M. Acosta et al., Eds.) Semantic Systems. The Power of AI and Knowledge Graphs. **Anais...Cham: Springer International Publishing**, 2019.
- LESK, M. **The seven ages of information retrieval.**, 1995. Disponível em: <<https://archive.ifla.org/VI/5/op/udtop5/udt-op5.pdf>>
- LÉTOFFÉ, O.; HUANG, X.; MARQUES-SILVA, J. **Towards trustable SHAP scores**. Proceedings of the AAAI Conference on Artificial Intelligence. **Anais...2025**.
- LEWIS, M. et al. **BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension**. (D. Jurafsky et al., Eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020. **Anais...Association for Computational Linguistics**, 2020. Disponível em: <<https://doi.org/10.18653/v1/2020.acl-main.703>>
- LI, Q.; JI, H. **Incremental joint extraction of entity mentions and relations**. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais...2014**.



- LI, Y. et al. **MAGE: Machine-generated Text Detection in the Wild**. (L.-W. Ku, A. Martins, V. Srikumar, Eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais...**Bangkok, Thailand: Association for Computational Linguistics, ago. 2024.
- LIKERT, R. **A Technique for the Measurement of Attitudes**. [s.l.] Archives of Psychology, 1932.
- LIMA, A. DA S.; BORGES, V. R. **Training and evaluating Named Entity Recognition Models using a Legal Corpus of publications from Government Gazettes**. 2022.
- LIMA, J. P.; COSTA, J. A.; ARAÚJO, D. C. **Comparison of Feature Extraction Methods for Brazilian Legal Documents Clustering**. 2021 IEEE Latin American Conference on Computational Intelligence (LA-CCI). **Anais...IEEE**, 2021. Disponível em: <<https://doi.org/10.1109/LA-CCI48322.2021.9769839>>
- LIMA, T. B. DE et al. Avaliação Automática de Redação: Uma revisão sistemática. **Revista Brasileira de Informática na Educação**, v. 31, p. 205–221, maio 2023.
- LIN, C. et al. SenseMood: Depression Detection on Social Media. Em: **2020 International Conference on Multimedia Retrieval**. New York, USA: Association for Computing Machinery, 2020a. p. 407–411.
- LIN, C.-Y. **ROUGE: A Package for Automatic Evaluation of Summaries**. Text Summarization Branches Out. **Anais...**Barcelona, Spain: Association for Computational Linguistics, jul. 2004. Disponível em: <<https://aclanthology.org/W04-1013>>
- LIN, H.; NG, V. **Abstractive Summarization: A Survey of the State of the Art**. **Proceedings of the AAAI Conference on Artificial Intelligence**, v. 33, n. 01, p. 9815–9822, 2019.
- LIN, J.; NOGUEIRA, R.; YATES, A. Pretrained Transformers for Text Ranking: BERT and Beyond. **arXiv preprint arXiv:2010.06467**, b2020.
- LIU, A. T. et al. **Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders**. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). **Anais...IEEE**, a2020.
- LIU, A. T.; LI, S.-W.; LEE, H. Tera: Self-supervised learning of transformer encoder representation for speech. **arXiv preprint arXiv:2007.06028**, b2020.
- LIU, B. Sentiment Analysis and Opinion Mining. **Synthesis Lectures on Human Language Technologies**, 2012.
- LIU, T.; YAO, J.-G.; LIN, C.-Y. **Towards improving neural named entity recognition with gazetteers**. Proceedings of the 57th annual meeting of the association for computational linguistics. **Anais...**a2019.



- LIU, X. et al. **Evaluating the Factuality of Large Language Models using Large-Scale Knowledge Graphs.**, 2024. Disponível em: <<https://arxiv.org/abs/2404.00942>>
- LIU, Y. et al. **RoBERTa: A Robustly Optimized BERT Pretraining Approach.**, 2019. Disponível em: <<https://arxiv.org/abs/1907.11692>>
- LIU, Y. et al. Roberta: A robustly optimized bert pretraining approach. **arXiv preprint arXiv:1907.11692**, b2019.
- LIU, Y.; HEARNE, J.; CONRAD, B. **Recognizing proper names in ur iii texts through supervised learning.** The Twenty-Ninth International Flairs Conference. **Anais...**2016.
- LIU, Y.; LAPATA, M. Text summarization with pretrained encoders. **arXiv preprint arXiv:1908.08345**, 2019.
- LIU, Z. et al. De-identification of clinical notes via recurrent neural network and conditional random field. **J Biomed Inform**, v. 75S, p. S34–S42, jun. 2017.
- LIU, Z. et al. **A Robustly Optimized BERT Pre-Training Approach with Post-Training.** Chinese Computational Linguistics: 20th China National Conference, CCL 2021, Hohhot, China, August 13–15, 2021, Proceedings. **Anais...**Berlin, Heidelberg: Springer-Verlag, 2021. Disponível em: <https://doi.org/10.1007/978-3-030-84186-7_31>
- LIVESO. **O que é BERT? - O mais recente algoritmo da Google.** Disponível em: <<https://liveseo.com.br/seo/o-que-e-bert-o-mais-recente-algoritmo-da-google/#:~:text=Bem%2C%20o%20BERT%2C%20de%20maneira,respostas%20possíveis%20para%20seus%20usuários>>.
- LO, C. **YiSi - a Unified Semantic MT Quality Evaluation and Estimation Metric for Languages with Different Levels of Available Resources.** Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 2: Shared Task Papers, Day 1. **Anais...**2019. Disponível em: <<https://doi.org/10.18653/v1/w19-5358>>
- LO, C.; WU, D. **MEANT: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles.** The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA. **Anais...**2011. Disponível em: <<https://aclanthology.org/P11-1023/>>
- LO, S. L. et al. Multilingual Sentiment Analysis: From Formal to Informal and Scarce Resource Languages. **Artificial Intelligence Review**, 2017.
- LOCATELLI, M. S. et al. **Examining the Behavior of LLM Architectures within the Framework of Standardized National Exams in Brazil.** Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society. **Anais...**AAAI Press, 2025.



- LOMMEL, A.; MELBY, A. **Tutorial: MQM-DQF: A Good Marriage (Translation Quality for the 21st Century)**. Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track). **Anais...**Boston, MA: Association for Machine Translation in the Americas, mar. 2018. Disponível em: <<https://aclanthology.org/W18-1925>>
- LOOI, Z. Q. et al. **MeLoDicA AI-Machine Learning Based Detection of Asthma via Vocal Audio Analysis**. 2024 IEEE Conference on Artificial Intelligence (CAI). **Anais...**IEEE, 2024.
- LOPES, C. DE S. et al. **Trend in the prevalence of depressive symptoms in Brazil: results from the Brazilian National Health Survey 2013 and 2019**. **Cad Saude Publica**, 6 maio 2022.
- LOPES, F.; TEIXEIRA, C.; GONÇALO OLIVEIRA, H. **Contributions to Clinical Named Entity Recognition in Portuguese**. Proceedings of the 18th BioNLP Workshop and Shared Task. **Anais...**Florence, Italy: Association for Computational Linguistics, ago. 2019. Disponível em: <<https://www.aclweb.org/anthology/W19-5024>>
- LÓPEZ, R. et al. **A qualitative analysis of a corpus of opinion summaries based on aspects**. Proceedings of the 9th Linguistic Annotation Workshop. **Anais...**2015.
- LOPEZ-GAZPIO, I. et al. Interpretable semantic textual similarity: Finding and explaining differences between sentences. **Knowledge-Based Systems**, v. 119, p. 186–199, 2017.
- LORÈ, F. et al. **An AI framework to support decisions on GDPR compliance**. **Journal of Intelligent Information Systems**, p. 1–28, 2023.
- LOSADA, D. E.; CRESTANI, F. **A Test Collection for Research on Depression and Language Use**. Experimental IR Meets Multilinguality, Multimodality, and Interaction. **Anais...**Cham: Springer, 2016.
- LOSADA, D. E.; CRESTANI, F.; PARAPAR, J. **eRISK 2017: CLEF lab on early risk prediction on the internet: experimental foundations**. Lecture Notes in Computer Science vol 10456. **Anais...**Cham: Springer, 2017.
- LOSADA, D. E.; CRESTANI, F.; PARAPAR, J. **Overview of eRisk: Early Risk Prediction on the Internet**. Lecture Notes in Computer Science vol 11018. **Anais...**Cham: Springer, 2018.
- LOSADA, D. E.; CRESTANI, F.; PARAPAR, J. **Overview of eRisk 2019 Early Risk Prediction on the Internet**. Lecture Notes in Computer Science vol 11696. **Anais...**2019.
- LOUIS, A.; HIGGINS, D. **Off-topic essay detection using short prompt texts**. Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications. **Anais...**Los Angeles, California: Association for Computatio-



nal Linguistics, jun. 2010.

LOUIS, A.; NENKOVA, A. Automatically assessing machine summary content without a gold standard. **Computational Linguistics**, v. 39, n. 2, p. 267–300, 2013.

LOVEYS, K. et al. **Small but Mighty: Affective Micropatterns for Quantifying Mental Health from Social Media Language**. Fourth Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. **Anais...**Vancouver, Canada: Association for Computational Linguistics, 2017.

LOVINS, J. B. Development of a stemming algorithm. **Mech. Transl. Comput. Linguistics**, v. 11, n. 1-2, p. 22–31, 1968.

LUCAS, J. et al. **Fighting Fire with Fire: The Dual Role of LLMs in Crafting and Detecting Elusive Disinformation**. (H. Bouamor, J. Pino, K. Bali, Eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. **Anais...**Singapore: Association for Computational Linguistics, dez. 2023. Disponível em: <<https://aclanthology.org/2023.emnlp-main.883>>

LUHN, H. P. The automatic creation of literature abstracts. **IBM Journal of research and development**, v. 2, n. 2, p. 159–165, 1958.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. **Advances in neural information processing systems**, v. 30, 2017.

MA, Q. et al. **Blend: a Novel Combined MT Metric Based on Direct Assessment - CASICT-DCU submission to WMT17 Metrics Task**. Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017. **Anais...**2017. Disponível em: <<https://doi.org/10.18653/v1/w17-4768>>

MACDONALD, C.; TONELLOTO, N. **Declarative Experimentation in Information Retrieval using PyTerrier**. Proceedings of ICTIR 2020. **Anais...**2020.

MACHADO, A. A. A. et al. **Personalitatem Lexicon: um léxico em português brasileiro para mineração de traços de personalidade em textos**. Proceedings of the Brazilian Symposium on Computers in Education. **Anais...**2015.

MACHADO, M. T.; PARDO, T. A. S. **NILC at ABSAPT 2022: Aspect Extraction for Portuguese**. Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2022), A Coruña, Spain, September 20, 2022. **Anais...**2022.

MACHADO, M. T.; PARDO, T. A. S.; RUIZ, E. E. S. **Creating a portuguese context sensitive lexicon for sentiment analysis**. Proceedings of the 13th international conference on computational processing of the Portuguese Language (PROPOR). **Anais...**2018.

MACIEL, A. M. B. **Para o reconhecimento da especificidade do termo jurídico**. mathesis—[s.l.] Universidade Federal do Rio Grande do Sul, RS, 2001.



MACOHIN, A.; CARNEIRO, J. V. V. Web Crawling e Web Scraping em sites de tribunais: publicidade processual e proteção de dados pessoais nas experiências europeia e brasileira. Em: WACHOWICZ, M. (Ed.). **Proteção de Dados Pessoais em Perspectiva: LGPD e RGPD na Ótica do Direito Comparado**. Curitiba: Gedai, UFPR, 2020.

MAIA, D. F. et al. **UlyssesSD-Br: Stance Detection in Brazilian Political Polls**. (G. Marreiros et al., Eds.) Progress in Artificial Intelligence. **Anais...** Cham: Springer International Publishing, 2022. Disponível em: <<https://github.com/Dyonnaton/UlyssesSD-Br>>

MAIA, M.; LEMLE, M.; FRANÇA, A. I. Efeito stroop e rastreamento ocular no processamento de palavras. **Ciências e Cognição** 2007, v. 12, p. 02–17, 2007.

MALENCHINI, F. M. et al. **Um Benchmark para Sistemas de Extração de Informação Aberta em Português**. Proceedings of the Symposium in Information and Human Language Technology (STIL 2019). **Anais...** Salvador, Bahia: SBC, out. 2019.

MALIK, V. et al. **ILDC for CJPE: Indian Legal Documents Corpus for Court Judgment Prediction and Explanation**. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). **Anais...** Association for Computational Linguistics, 2021.

MALOUF, R. **A Comparison of Algorithms for Maximum Entropy Parameter Estimation**. COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002). **Anais...** 2002. Disponível em: <<https://aclanthology.org/W02-2018>>

MALTA, D. et al. **Prevalence of sexual violence in school children in Brazil: National School Health Survey 2019**. **Population Medicine**, v. 5, abr. 2023.

MANDAL, A. et al. **Unsupervised approaches for measuring textual similarity between legal court case reports**. **Artificial Intelligence and Law**, v. 29, n. 3, p. 417–451, 2021.

MANI, I. Automatic Summarization. **John Benjamins Publishing Company**, v. 2, p. 399–408, 2001.

MANI, I.; MAYBURY, M. T. **Advances in automatic text summarization**. [s.l.] MIT press, 1999.

MANN, P.; MATSUSHIMA, E. H.; PAES, A. **Detecting Depression from Social Media Data as a Multiple-Instance Learning Task**. 10th International Conference on Affective Computing and Intelligent Interaction (ACII). **Anais...** 2022.

MANN, P.; PAES, A.; MATSUSHIMA, E. H. **See and Read: Detecting Depression Symptoms in Higher Education Students Using Multimodal Social Media Data**. Proceedings of the International AAAI Conference on Web and Social Media. **Anais...** 2020.



- MANN, W.; THOMPSON, S. **Rethorical Structure Theory: Toward a functional theory of text organization**. *Text*, v. 8, p. 243–281, jan. 1988.
- MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. **Introduction to Information Retrieval**. New York, NY, USA: Cambridge University Press, 2008a.
- MANNING, C. D.; SCHÜTZE, H. *Foundations of Statistical Natural Language Processing*. 1999.
- MANNING, C. D.; SCHÜTZE, H.; RAGHAVAN, P. **Introduction to information retrieval**. [s.l.] Cambridge University Press Cambridge, 2008b.
- MARCUSCHI, L. A. **Produção textual, análise de gêneros e compreensão**. [s.l.] Parábola Ed., 2008.
- MARGARIDO, P. R. A. et al. **Automatic Summarization for Text Simplification: Evaluating Text Understanding by Poor Readers**. Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web. *Anais...: WebMedia '08*. New York, NY, USA: ACM, 2008. Disponível em: <<http://doi.acm.org/10.1145/1809980.1810057>>
- MARIE, B.; FUJITA, A.; RUBINO, R. **Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers**. *arXiv:2106.15195 [cs]*, jun. 2021.
- MARINHO, J. et al. **Automated Essay Scoring: An approach based on ENEM competencies**. *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional. Anais...SBC*, a2022.
- MARINHO, J.; ANCHIÊTA, R.; MOURA, R. **Essay-BR: a Brazilian Corpus to Automatic Essay Scoring Task**. *Journal of Information and Data Management*, v. 13, n. 1, p. 65–76, b2022.
- MARON, M. E. **Automatic indexing: an experimental inquiry**. *Journal of the ACM (JACM)*, v. 8, n. 3, p. 404–417, 1961.
- MARTINS, D. B. DE J. **Pós-edição automática de textos traduzidos automaticamente de inglês para português do Brasil**. Mestrado—São Carlos: Universidade Federal de São Carlos, 2014.
- MARTINS, D. B. DE J.; CASELI, H. DE M. **Automatic machine translation error identification**. *Machine Translation*, v. 29, n. 1, p. 1–24, 2015.
- MARTINS, R. T. et al. **An interlingua aiming at communication on the Web: How language-independent can it be?** NAACL-ANLP 2000 Workshop: Applied Interlinguas: Practical Applications of Interlingual Approaches to NLP. *Anais...2000*. Disponível em: <<https://aclanthology.org/W00-0204>>
- MARTINS, T. B. F. et al. **Readability Formulas Applied to Textbooks in Brazilian**



Portuguese. [s.l.] ICMSC-USP, 1996.

MASCARENHAS, M. D. M. et al. *Análise das notificações de violência por parceiro íntimo contra mulheres, Brasil, 2011-2017*. *Revista Brasileira de Epidemiologia*, v. 23, p. e200007.SUPL.1, 2020.

MATHEUS GAUY, M. et al. *Contrasting deep learning audio models for direct respiratory insufficiency detection versus blood oxygen saturation estimation*. *Intelligence-Based Medicine*, v. 13, p. 100331, 2026.

MATOS, A. et al. *Accent classification is challenging but pre-training helps: a case study with novel brazilian portuguese datasets*. Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 1. *Anais...*2024.

MATTEI, L. D. et al. ATE ABSITA@ EVALITA2020: Overview of the Aspect Term Extraction and Aspect-based Sentiment Analysis Task. *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, 2020.

MATTHIESSEN, M. C. M. I. *Applying systemic functional linguistics in healthcare contexts*. *Text and Talk*, v. 33, n. 4-5, p. 437-447, 19 ago. 2013.

MATTHIESSEN, M. C. M. I.; TERUYA, K.; WU, C. Multilingual studies as a multi-dimensional space of interconnected language studies. Em: *Meaning in context : strategies for implementing intelligent applications of language studies*. [s.l.] Continuum, 2008. p. 146-221.

MAX, A. *Writing for Language-Impaired Readers*. In: Gelbukh A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2006. Lecture Notes in Computer Science, vol 3878. *Anais...*Springer, Berlin, Heidelberg, 2006.

MAXWELL, K. T.; SCHAFER, B. *Concept and context in legal information retrieval*. Em: *Legal Knowledge and Information Systems*. [s.l.] IOS Press, 2008. p. 63-72.

MAYER, R. E. Elaboration techniques that increase the meaningfulness of technical text: An experimental test of the learning strategy hypothesis. *Journal of Educational Psychology*, v. 72, n. 6, p. 770-784, 1980.

MAYFIELD, E.; BLACK, A. W. *Should You Fine-Tune BERT for Automated Essay Scoring?* Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications. *Anais...*Association for Computational Linguistics, jul. 2020.

MAZIERO, E. G.; PARDO, T. A. S.; ALUÍSIO, S. M. Ferramenta de Análise Automática de Inteligibilidade de Corpus (AIC). *NILC - ICMC-USP*, 2008.

MCCALLUM, A.; LI, W. *Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons*. Proceedings



of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4. **Anais...**2003.

MCNAMARA, D. S. et al. **Coh-Metrix Common Core T.E.R.A. version 1.0.**, 2013. Disponível em: <<http://www.commoncoretera.com/>>

MCNAMARA, D. S. et al. **Automated Evaluation of Text and Discourse with Coh-Metrix**. 1a. ed. [s.l.] Cambridge University Press, 2014.

MCWHORTER, J. H. **The Worlds Simplest Grammars Are Creole Grammars**. **Linguistic Typology**, v. 5, n. 2-3, jan. 2001.

MENDES, A. R.; CASELI, H. M. **Identifying Fine-grained Depression Signs in Social Media Posts**. Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC 2024). **Anais...**2024.

MEYER, C. F. et al. The world wide web as linguistic corpus. Em: **Corpus Analysis**. [s.l.] Brill Rodopi, 2003. p. 241–254.

MIKOLOV, T. et al. **Distributed Representations of Words and Phrases and their Compositionality**. (C. J. Burges et al., Eds.)Advances in Neural Information Processing Systems. **Anais...**Curran Associates, Inc., b2013. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf>

MIKOLOV, T. et al. **Efficient Estimation of Word Representations in Vector Space**. Proceedings of the International Conference on Learning Representations (ICLR 2013) - Workshop Track. **Anais...**Scottsdale, Arizona, USA: a2013.

MILLER, G. A. WordNet: A Lexical Database for English. **Communications of the ACM**, v. Vol. 38, No. 11, p. 39–41, 1995.

MINAEE, S. et al. Deep learning–based text classification: a comprehensive review. **ACM computing surveys (CSUR)**, v. 54, n. 3, p. 1–40, 2021.

MIWA, M.; BANSAL, M. **End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures**. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais...**Association for Computational Linguistics, 2016.

MOHAN, S. et al. **The Impact of Toxic Language on the Health of Reddit Communities**. Proceedings of the Canadian Conference on AI. **Anais...**2017.

MOLLAS, I. et al. ETHOS: a multi-label hate speech detection dataset. **Complex & Intelligent Systems**, 2022.

MONTEIRO, R. A. et al. **Contributions to the Study of Fake News in Portuguese: New Corpus and Automatic Detection Results**. Proceedings of the 13th international conference on computational processing of the Portuguese Language. **Anais...**Canela,



Rio Grande do Sul, Brazil: Springer International Publishing, set. 2018.

MONTORO, A. F. **Curso de Teoria Geral do Direito - Aula 2: A linguagem do direito: semântica, sintática e pragmática**. Disponível em: <<http://www.dialdata.com.br/ilam/aula2>>.

MOORKENS, J. et al. **Correlations of perceived post-editing effort with measurements of actual effort**. *Machine Translation*, v. 29, n. 3/4, p. 267–284, 2015.

MOORKENS, J. **Under pressure: translation in times of austerity**. *Perspectives*, v. 25, n. 3, p. 464–477, fev. 2017.

MORENO, G. C. DE L. et al. **ALT: um software para análise de legibilidade de textos em Língua Portuguesa**. *preprint*, 2022.

MORENO, J.; BRESSAN, G. **FACTCK.BR: a new dataset to study fake news**. : WebMedia '19. New York, NY, USA: Association for Computing Machinery, 2019. Disponível em: <<https://doi.org/10.1145/3323503.3361698>>

MORENO SCHNEIDER, J. et al. **Lynx: A knowledge-based AI service platform for content processing, enrichment and analysis for the legal domain**. *Information Systems*, v. 106, p. 101966, 2022.

MOTA, C. R3M, uma participação minimalista no Segundo HAREM. **quot; In Cristina Mota; Diana Santos (ed) Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM** Linguateca 2008, 2008.

MOTA, C. C. et al. **Reconhecimento de entidades nomeadas em documentos jurídicos em português utilizando redes neurais**. Encontro Nacional de Inteligência Artificial e Computacional (ENIAC). *Anais...SBC*, 2021.

MOTA, C.; SANTOS, D. (EDS.). **Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM**. [s.l.] Linguateca, 2008.

MOTA, C.; SANTOS, D.; RANCHHOD, E. Avaliação de reconhecimento de entidades mencionadas: princípio de HAREM. **Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa**, p. 161–175, 2007.

MOTTA, E. **Sentenças Judiciais e Acessibilidade Textual e Terminológica**. *Domínios de Linguagem*, v. 15, n. 3, p. 761–813, 2021.

MOTTA, E. **SENTENÇAS JUDICIAIS E LINGUAGEM SIMPLES: um encontro possível e necessário**. mathesis—[s.l.] Universidade Federal do Rio Grande do Sul, RS, 2022.

MOUAWAD, P.; DUBNOV, T.; DUBNOV, S. Robust detection of COVID-19 in cough sounds: using recurrence dynamics and variable Markov model. *SN Computer Science*, v. 2, n. 1, p. 34, 2021.



MOUJAHID, H. et al. Combining CNN and Grad-Cam for COVID-19 Disease Prediction and Visual Explanation. **Intelligent Automation & Soft Computing**, v. 32, n. 2, 2022.

MOUSAVI, E. **Complex Network Series: Part 1 — An Overview.**, 2024. Disponível em: <<https://medium.com/@ebimsv/complex-network-series-part-1-an-overview-ddd9e0b6ec3a>>

MURPHY, K. P. **Machine learning: a probabilistic perspective.** [s.l.] MIT press, 2012.

NADEAU, D. **Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision.** tese de doutorado—[s.l.] University of Ottawa, 2007.

NAGAO, M. **A Framework of a Mechanical Translation between Japanese and English by Analogy Principle.** Em: NIRENBURG, S.; SOMERS, H. L.; WILKS, Y. A. (Eds.). **Readings in Machine Translation.** [s.l.] The MIT Press, 1984.

NAIR, S. S.; JEEVEN, V. A brief overview of metadata formats. **DESIDOC Journal of Library & Information Technology**, v. 24, n. 4, 2004.

NAIR, V.; HINTON, G. E. **Rectified linear units improve restricted boltzmann machines.** Icml. **Anais...**2010.

NARDE, W. **Análise de notícias falsas em rede social: uma abordagem utilizando transferência de aprendizagem e Transformers.** https://www.monografias.ufop.br/bitstream/35400000/3122/6/MONOGRAFIA_AnaliseNoticiasFalsas.pdf, 2021.

NASAR, Z.; JAFFRY, S. W.; MALIK, M. K. Named entity recognition and relation extraction: State-of-the-art. **ACM Computing Surveys (CSUR)**, v. 54, n. 1, p. 1–39, 2021.

NASCIMENTO, D. N. C. R. DO. **Sumarização de artigos científicos em português no domínio da saúde.** mathesis—[s.l.] (Mestrado em Informática) - Programa de Pós-Graduação em Informática da PUC-Rio, Rio de Janeiro, 2023.

NASCIMENTO, G. et al. **Hate speech detection using brazilian imageboards.** Proceedings of the 25th Brazillian Symposium on Multimedia and the Web. **Anais...**2019.

NASCIMENTO, R. DA S. et al. **Identificando Sinais de Comportamento Depressivo em Redes Sociais.** Anais do VII Brazilian Workshop on Social Network Analysis and Mining. **Anais...**Porto Alegre, Brazil: SBC, 2018.

NATH, N.; LEE, S.-H.; LEE, I. **NEAR: Named Entity and Attribute Recognition of Clinical Concepts.** **J. of Biomedical Informatics**, v. 130, n. C, jun. 2022.

NECO, R. P.; FORCADA, M. L. **Asynchronous translations with recurrent**



neural nets. Proceedings of International Conference on Neural Networks (ICNN'97). **Anais...**1997.

NETO, J. R. C. S. A. V. S.; FALEIROS, T. DE P. **Deep Active-Self Learning Applied to Named Entity Recognition**. (A. Britto, K. Valdivia Delgado, Eds.) Intelligent Systems. **Anais...**Cham: Springer International Publishing, 2021. Disponível em: <<https://avio11.github.io/resources/aposentadoria/aposentadoria.html>>

NEURALMIND. **NeuralMind disponibiliza modelo BERT do Google em português**. **Neuralmind blog**. Disponível em: <<https://neuralmind.ai/2020/01/26/neuralmind-disponibiliza-modelo-bert-inteligencia-artificial-do-google-em-portugues/>>.

NEWELL, E. et al. **Assessing the Verifiability of Attributions in News Text**. (G. Kondrak, T. Watanabe, Eds.) Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers). **Anais...**Taipei, Taiwan: Asian Federation of Natural Language Processing, nov. 2017. Disponível em: <<https://aclanthology.org/I17-1076>>

NEWMAN, M. E. J. **Networks: An Introduction**. Oxford: Oxford University Press, 2010.

NEWMAN, M. E. J.; GIRVAN, M. **Finding and evaluating community structure in networks**. **Physical Review E**, v. 69, n. 2, p. 026113, 2004.

NEWMAN, N. et al. **Reuters institute digital news report 2020**. [s.l.] Report of the Reuters Institute for the Study of Journalism, 2020.

NGUYEN, D. B.; THEOBALD, M.; WEIKUM, G. J-NERD: joint named entity recognition and disambiguation with rich linguistic features. **Transactions of the Association for Computational Linguistics**, v. 4, p. 215–229, 2016.

NICHOLS, J. **Linguistic Diversity in Space and Time**. [s.l.] University of Chicago Press, 1998.

NIIZUMI, D. et al. **Masked spectrogram modeling using masked autoencoders for learning general-purpose audio representation**. HEAR: Holistic Evaluation of Audio Representations. **Anais...**PMLR, 2022.

NIIZUMI, D. et al. **Masked modeling duo: Learning representations by encouraging both networks to model the input**. ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). **Anais...**IEEE, 2023.

NIIZUMI, D. et al. **Towards Pre-training an Effective Respiratory Audio Foundation Model**. **arXiv preprint arXiv:2505.15307**, 2025.

NOGUEIRA, R. et al. **Document expansion by query prediction**. **arXiv preprint arXiv:1904.08375**, 2019.



- NUNES, M. DAS G. V. et al. O uso de interlíngua para comunicação via Internet: a decodificação UNL-português. **Revista Tecnologia da Informação**, v. 3, n. 1, p. 49–55, 2003.
- NUNES, R. **LegiSubject-Br-Keywords**. <https://huggingface.co/datasets/ronunes/LegiSubject-Br-Keywords>, 2024.
- NUNES, R. O. et al. **Out of Sesame Street: A Study of Portuguese Legal Named Entity Recognition Through In-Context Learning**. INSTICC; SciTePress, a2024.
- NUNES, R. O. et al. **A Named Entity Recognition Approach for Portuguese Legislative Texts Using Self-Learning**. (P. Gamallo et al., Eds.) Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1. **Anais...**Santiago de Compostela, Galicia/Spain: Association for Computational Linguistics, mar. b2024. Disponível em: <<https://aclanthology.org/2024.propor-1.30>>
- NUNES, R. O. **Data contamination in specialized named entity recognition corpora**. Dissertação de Mestrado—Porto Alegre, Brasil: Universidade Federal do Rio Grande do Sul, 2025.
- O'BRIEN, S. Towards predicting post-editing productivity. **Machine translation**, v. 25, p. 197–215, 2011.
- O'BRIEN, S. et al. **Dynamic Quality Evaluation Framework**. [s.l.] TAUS Labs Report. The Translation Automation User Society-TAUS, 2011.
- OAEPUBLISH. **CES 2025 Article.**, 2025. Disponível em: <<https://www.oaepublish.com/articles/ces.2025.34>>
- OCH, F. J.; NEY, H. **The Alignment Template Approach to Statistical Machine Translation**. **Computational Linguistics**, v. 30, n. 4, p. 417–449, dez. 2004.
- ÖHMAN, A. et al. The public health turn on violence against women: analysing Swedish healthcare law, public health and gender-equality policies. **BMC Public Health**, v. 20, p. 1–12, 2020.
- OKANO, E. Y. et al. **Fake News Detection on Fake.Br Using Hierarchical Attention Networks**. (P. Quaresma et al., Eds.) Computational Processing of the Portuguese Language. **Anais...**Cham: Springer International Publishing, 2020.
- OKANO, E. Y.; RUIZ, E. E. S. **Using linguistic cues to detect fake news on the brazilian portuguese parallel corpus Fake.br**. Proceedings of the Symposium in Information and Human Language Technology. **Anais...**Brazilian Computer Society, 2019.
- OKSANEN, A. et al. **Semantic Finlex: Transforming, Publishing, and Using Finnish Legislation and Case Law As Linked Open Data on the Web**. Em: PERUGINELLI, G.; FARO, S. (Eds.). **Knowledge of the Law in the Big Data Age**. Frontiers em Artificial Intelligence e Applications. [s.l.] IOS Press, 2019b. v. 317p. 212–228.



OKSANEN, A. et al. **ANOPPI: A Pseudonymization Service for Finnish Court Documents**. JURIX. *Anais...*a2019.

OKSANEN, A. et al. **An Anonymization Tool for Open Data Publication of Legal Documents**. Joint Proceedings of ISWC2022 Workshops. *Anais...CEUR-WS. org*, 2022. Disponível em: <<https://ceur-ws.org/Vol-3257/>>

OLIVAL, F.; CAMERON, H.; VIEIRA, R. **As Memórias Paroquiais: do manuscrito ao digital**. Actas da Jornada de Humanidades Digitais do CIDEHUS (to appear). *Anais...*2022.

OLIVEIRA, A. et al. **Toxic Text Classification in Portuguese: Is LLaMA 3.1 8B All You Need?** Proceedings of the 15th Brazilian Symposium in Information and Human Language Technology. *Anais...*a2024.

OLIVEIRA, I. L. Uma mentira repetida mil vezes se transforma em verdade? Reflexões sobre as dinâmicas discursivas e seus efeitos na saúde. Em: **Desinformação o mal do século: Distorções, inverdades, fake news: a democracia ameaçada**. [s.l.: s.n.]. p. 299–315.

OLIVEIRA, L. E. S. et al. **SemClinBr - a multi-institutional and multi-specialty semantically annotated corpus for Portuguese clinical NLP tasks**. *Journal of Biomedical Semantics*, v. 13, n. 1, a2022.

OLIVEIRA, L. F. A. DE et al. **Challenges In Annotating A Treebank Of Clinical Narratives In Brazilian Portuguese**. Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings. *Anais...*Berlin, Heidelberg: Springer-Verlag, b2022. Disponível em: <https://doi.org/10.1007/978-3-030-98305-5_9>

OLIVEIRA, L. F. DE; OLIVEIRA, O. N., Jr. **The Landscape of Wearable Sensors and Automated Literature Analysis with Large-Language Models**. *ACS Omega*, v. 10, n. 37, p. 42127–42134, 2025.

OLIVEIRA, L. M. DE; MOURA, M. F. DE; OLIVEIRA, O. N., Jr. **Using network analysis and large-language models to obtain a landscape of the literature on dressing materials for wound healing: The predominance of chitosan and other biomacromolecules**. *International Journal of Biological Macromolecules*, v. 306.3, p. 141565, 2025.

OLIVEIRA, L.; CLARO, D.; SOUZA, M. **DptOIE: a Portuguese open information extraction based on dependency analysis**. *Artificial Intelligence Review*, v. 56, p. 1–32, dez. c2022.

OLIVEIRA, N. et al. **Processamento de Linguagem Natural para Identificação de Notícias Falsas em Redes Sociais: Ferramentas, Tendências e Desafios**. Em: [s.l.] SBC, 2020.

OLIVEIRA, R. L. DE; MARTINS, J. T.; PARABONI, I. **Mental health prediction from social media connections**. *New Review of Hypermedia and Multimedia*, b2024.



OLIVEIRA, R. L. DE; PARABONI, I. **A Bag-of-Users approach to mental health prediction from social media data**. 16th International Conference on Computational Processing of Portuguese (PROPOR 2024). **Anais...**Santiago de Compostela, Spain: 2024.

OLIVEIRA, V. et al. **Combining prompt-based language models and weak supervision for labeling named entity recognition on legal documents**. **Artificial Intelligence and Law**, p. 1–21, fev. c2024.

OMOHUNDRO, S. M. **Five balltree construction algorithms**. Berkeley, CA, USA: International Computer Science Institute, 1989.

ONAGA, T.; FUJITA, M.; YOSHINOBU, K. **Japanese Legal Bar Problem Solver Focusing on Person Names**. Proceedings of the Tenth International Competition on Legal Information Extraction/Entailment (COLIEE 2023). **Anais...**2023. Disponível em: <<https://sites.ualberta.ca/~rabelo/COLIEE2023>>

ORENGO, V. M.; BURIOL, L. S.; COELHO, A. R. **A study on the use of stemming for monolingual ad-hoc Portuguese information retrieval**. Workshop of the Cross-Language Evaluation Forum for European Languages. **Anais...**Springer, 2006.

ORENGO, V. M.; HUYCK, C. **A Stemming Algorithm for the Portuguese Language**. Proceedings Eighth Symposium on String Processing and Information Retrieval. **Anais...**IEEE Computer Society, 2001.

ORLANDIC, L.; TELJEIRO, T.; ATIENZA, D. The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. **Scientific Data**, v. 8, n. 1, p. 156, 2021.

OSTENDORFF, M. et al. **Evaluating document representations for content-based legal literature recommendations**. Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law. **Anais...**2021. Disponível em: <<https://doi.org/10.1145/3462757.3466073>>

OTT, M. et al. **Finding Deceptive Opinion Spam by Any Stretch of the Imagination**. (D. Lin, Y. Matsumoto, R. Mihalcea, Eds.) Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. **Anais...**Portland, Oregon, USA: Association for Computational Linguistics, jun. 2011. Disponível em: <<https://aclanthology.org/P11-1032>>

OUYANG, L. et al. Training language models to follow instructions with human feedback. **Advances in Neural Information Processing Systems (NeurIPS)**, v. 35, p. 27730–27744, 2022.

PACK, A.; BARRETT, A.; ESCALANTE, J. **Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability**. **Computers and Education: Artificial Intelligence**, v. 6, p. 100234, 2024.

PAETZOLD, G. H.; SPECIA, L. **Unsupervised Lexical Simplification for Non-**



native Speakers. Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. **Anais...** AAAI'16.Phoenix, Arizona: AAAI Press, a2016. Disponível em: <<http://dl.acm.org/citation.cfm?id=3016387.3016433>>

PAETZOLD, G.; SPECIA, L. **Inferring Psycholinguistic Properties of Words.** NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016. **Anais...**b2016. Disponível em: <<http://aclweb.org/anthology/N/N16/N16-1050.pdf>>

PAETZOLD, G.; SPECIA, L. **Understanding the Lexical Simplification Needs of Non-Native Speakers of English.** (Y. Matsumoto, R. Prasad, Eds.)Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. **Anais...**Osaka, Japan: The COLING 2016 Organizing Committee, dez. c2016. Disponível em: <<https://aclanthology.org/C16-1069>>

PAETZOLD, G.; SPECIA, L. **Lexical Simplification with Neural Ranking.** Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. **Anais...**Valencia, Spain: Association for Computational Linguistics, abr. 2017. Disponível em: <<http://www.aclweb.org/anthology/E17-2006>>

PAIOLA, P. H. **Sumarização abstrativa de textos em português utilizando aprendizado de máquina.** mathesis—[s.l.] (Mestrado em Ciências da Computação) - Programa de Pós-Graduação em Ciência da Computação, Universidade Estadual Paulista "Júlio de Mesquita Filho", 2022.

PAIOLA, P. H. et al. **RecognaSumm: A Novel Brazilian Summarization Dataset.** (P. Gamallo et al., Eds.)Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1. **Anais...**Santiago de Compostela, Galicia/Spain: Association for Computational Linguistics, mar. 2024. Disponível em: <<https://aclanthology.org/2024.propor-1.63>>

PAIOLA, P. H.; ROSA, G. H. DE; PAPA, J. P. **Deep Learning-Based Abstractive Summarization for Brazilian Portuguese Texts.** Intelligent Systems: 11th Brazilian Conference, BRACIS 2022, Campinas, Brazil, November 28 – December 1, 2022, Proceedings, Part II. **Anais...**Berlin, Heidelberg: Springer-Verlag, 2022. Disponível em: <https://doi.org/10.1007/978-3-031-21689-3_34>

PAIS, V. et al. **Named Entity Recognition in the Romanian Legal Domain.** Proceedings of the Natural Legal Language Processing Workshop 2021. **Anais...**Punta Cana, Dominican Republic: Association for Computational Linguistics, nov. 2021.

PAIS, V. et al. **LegalNERo: A linked corpus for named entity recognition in the Romanian legal domain.** *Semantic Web journal*, 2024.

PALACIO-NIÑO, J.-O.; BERZAL, F. **Enhancing Community Detection in Networks: A Comparative Analysis of Local Metrics and Hierarchical Algo-**



- rithms**. arXiv preprint, 2024. Disponível em: <<https://arxiv.org/html/2408.09072v2>>
- PANWAR, H. et al. A deep learning and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-Scan images. **Chaos, Solitons & Fractals**, v. 140, p. 110190, 2020.
- PAPA, J. P.; FALCÃO, A. X.; SUZUKI, C. T. N. **Supervised pattern classification based on optimum-path forest**. **International Journal of Imaging Systems and Technology**, v. 19, n. 2, p. 120–131, 2009.
- PAPINENI, K. et al. **BLEU: A Method for Automatic Evaluation of Machine Translation**. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. **Anais...**: ACL '02.USA: Association for Computational Linguistics, 2002. Disponível em: <<https://doi.org/10.3115/1073083.1073135>>
- PARAGUASSU, L. et al. **MedSimples: An Automated Simplification Tool for Promoting Health Literacy in Brazil**. DHandNLP@PROPOR. **Anais...**2020. Disponível em: <<https://api.semanticscholar.org/CorpusID:218910691>>
- PARDO, T. A. S. **Gistsumm: Um sumariador automático baseado na ideia principal de textos**. [s.l.] Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional, Universidade de São Paulo, 2002.
- PARDO, T. A. S.; RINO, L. H. M. **TeMário: Um corpus para sumarização automática de textos**. [s.l.] Série de Relatórios Técnicos da Universidade de São Carlos, 2003.
- PARIDA, S.; MOTLICEK, P. **Abstract text summarization: A low resource challenge**. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). **Anais...**2019.
- PARK, S. et al. Activities on Facebook Reveal the Depressive State of Users. **J Med Internet Res**, v. 15, n. 10, p. e217, 2013.
- PAROUBEK, P.; CHAUDIRON, S.; HIRSCHMAN, L. **Principles of Evaluation in Natural Language Processing**. Traitement Automatique des Langues, Volume 48, Numéro 1 : Principes de l'évaluation en Traitement Automatique des Langues [Principles of Evaluation in Natural Language Processing]. **Anais...**France: ATALA (Association pour le Traitement Automatique des Langues), 2007. Disponível em: <<https://aclanthology.org/2007.tal-1.1>>
- PASQUALINI, B. **Corpop : um corpus de referência do português popular escrito do Brasil**. UFRGS - Porto Alegre - RS: Instituto de Letras - UFRGS, 2018.
- PASQUALOTTI, P. R. **WordNet Affect BR – uma base de expressões de emoção em Português**. [s.l.] Novas Edições Acadêmicas, 2015.



- PELLE, R. P. DE; MOREIRA, V. **Offensive Comments in the Brazilian Web: a dataset and baseline results**. Anais do VI Brazilian Workshop on Social Network Analysis and Mining. **Anais...**2017.
- PENNEBAKER, J. W. et al. The development and psychometric properties of LIWC2015. **The University of Texas at Austin**, 2015.
- PENNEBAKER, J. W.; FRANCIS, M. E.; BOOTH, R. J. **Linguistic Inquiry and Word Count**. [s.l.] Lawrence Erlbaum Associates, 2001.
- PENNINGTON, J.; SOCHER, R.; MANNING, C. **GloVe: Global Vectors for Word Representation**. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). **Anais...**Doha, Qatar: Association for Computational Linguistics, out. 2014. Disponível em: <<https://aclanthology.org/D14-1162>>
- PEREIRA, D. A. A Survey of Sentiment Analysis in the Portuguese Language. **Artificial Intelligence Review**, 2021.
- PEREIRA, V.; PINHEIRO, V. **Report - um sistema de extração de informações aberta para língua portuguesa**. Anais do X Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. **Anais...**SBC, 2015.
- PÉREZ-ROSAS, V.; MIHALCEA, R. **Cross-cultural Deception Detection**. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. **Anais...**Baltimore, MD, USA: Association for Computational Linguistics, 2014.
- PERSING, I.; NG, V. **Modeling Prompt Adherence in Student Essays**. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. **Anais...**Baltimore, Maryland: Association for Computational Linguistics, jun. 2014.
- PETERS, M. E. et al. **Semi-supervised sequence tagging with bidirectional language models**. Proc. of ACL-2017. **Anais...**Vancouver, Canada: Association for Computational Linguistics, 2017.
- PETERS, M. E. et al. **Deep Contextualized Word Representations**. (M. A. Walker, H. Ji, A. Stent, Eds.)Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers). **Anais...**Association for Computational Linguistics, 2018. Disponível em: <<https://doi.org/10.18653/v1/n18-1202>>
- PETRI, M. J. C. **Manual de Linguagem Jurídica**. 3rd. ed. São Paulo: Saraiva, 2017.
- PILÁN, I. et al. The text anonymization benchmark (tab): A dedicated corpus and evaluation framework for text anonymization. **Computational Linguistics**, v. 48, n. 4, p. 1053–1101, 2022.
- PINKAS, G. et al. **SARS-CoV-2 Detection From Voice**. **IEEE Open Journal of Engi-**



neering in Medicine and Biology, v. 1, p. 268–274, 2020.

PINTO, I. V. et al. **Fatores associados ao óbito de mulheres com notificação de violência por parceiro íntimo no Brasil**. *Ciência & Saúde Coletiva*, v. 26, n. 3, p. 975–985, mar. 2021.

PINTO, L.; MELGAR, A. **A classification model for Portuguese documents in the juridical domain**. 2016 11th Iberian Conference on Information Systems and Technologies (CISTI). *Anais...IEEE*, 2016.

PIRES, R. et al. **Sabiá: Portuguese Large Language Models**. (M. C. Naldi, R. A. C. Bianchi, Eds.) *Intelligent Systems*. *Anais...Cham*: Springer Nature Switzerland, 2023.

PIRES, V. B.; GUERREIRO, D.; et al. **Portuguese Fake News Classification with BERT models**. Encontro Nacional de Inteligência Artificial e Computacional (ENIAC). *Anais...SBC*, 2024.

PIRES, V.; SILVA, D. G. E. **Portuguese Fake News Classification with BERT models**. *Anais do XXI Encontro Nacional de Inteligência Artificial e Computacional*. *Anais...Porto Alegre, RS, Brasil*: SBC, 2024. Disponível em: <<https://sol.sbc.org.br/index.php/eniac/article/view/33848>>

PIRINA, I.; ÇÖLTEKIN, ÇAĞRI. **Identifying Depression on Reddit: The Effect of Training Data**. Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task. *Anais...2018*.

PIROVANI, J. P. C. **CRF+ LG: uma abordagem híbrida para o reconhecimento de entidades nomeadas em português**. tese de doutorado—[s.l.] Universidade Federal do Espírito Santo, Vitória (Brasil), 2019.

PITLER, E.; LOUIS, A.; NENKOVA, A. **Automatic evaluation of linguistic quality in multi-document summarization**. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. *Anais...2010*.

POLO, F. M. et al. **LegalNLP – Natural Language Processing methods for the Brazilian Legal Language.**, 2021. Disponível em: <<https://arxiv.org/abs/2110.15709>>

PONTES, L. B. L.; OLIVEIRA, H. T. A. DE; ASSIS BOLDT, F. DE. **Avaliação de Modelos Neurais para Sumarização de Código-fonte**. *Anais do XLIX Seminário Integrado de Software e Hardware*. *Anais...SBC*, 2022.

PONTIKI, M. et al. **SemEval-2014 Task 4: Aspect Based Sentiment Analysis**. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). *Anais...Association for Computational Linguistics*, 2014. Disponível em: <<https://aclanthology.org/S14-2004/>>

PONTIKI, M. et al. **SemEval-2015 Task 12: Aspect Based Sentiment Analysis**. Proceedings of the 9th International Workshop on Semantic Evaluation. *Anais...2015*.



PONTIKI, M. et al. **SemEval-2016 Task 5: Aspect Based Sentiment Analysis**. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). **Anais...**2016.

POPOVIC, M.; BURCHARDT, A. **From Human to Automatic Error Classification for Machine Translation Output**. Proceedings of the 15th Conference of the European Association for Machine Translation. **Anais...**Leuven, Belgium: 2011. Disponível em: <<https://aclanthology.org/2011.eamt-1.36.pdf>>

POPOVIĆ, M. **chrF: character n-gram F-score for automatic MT evaluation**. Proceedings of the Tenth Workshop on Statistical Machine Translation. **Anais...**Lisbon, Portugal: Association for Computational Linguistics, set. 2015. Disponível em: <<https://aclanthology.org/W15-3049>>

PORTER, M. F. **An algorithm for suffix stripping**. **Program**, v. 14, n. 3, p. 130–137, 1980.

PRADEEP, R. et al. **H2oloo at trec 2020: When all you got is a hammer... deep learning, health misinformation, and precision medicine**. **Corpus**, v. 5, n. d3, p. d2, 2020.

PUSTEJOVSKY, J. **The Generative Lexicon**. Cambridge, USA: MIT Press, 1995.

QIU, H. et al. **AMRFact: Enhancing Summarization Factuality Evaluation with AMR-Driven Negative Samples Generation**. (K. Duh, H. Gomez, S. Bethard, Eds.) Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). **Anais...**Mexico City, Mexico: Association for Computational Linguistics, jun. 2024. Disponível em: <<https://aclanthology.org/2024.naacl-long.33/>>

QIU, Q. et al. BiLSTM-CRF for geological named entity recognition from the geoscience literature. **Earth Science Informatics**, v. 12, n. 4, p. 565–579, 2019.

QUARESMA, P.; FINATTO, M. J. B. **Information Extraction from Historical Texts: a Case Study**. DHandNLP@ PROPOR. **Anais...**2020.

QUARESMA, P.; GONÇALVES, T. **Using Linguistic Information and Machine Learning Techniques to Identify Entities from Juridical Documents**. Em: FRANCESCONI, E. et al. (Eds.). **Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. p. 44–59.

QUINLAN, J. R. **C4.5: Programs for Machine Learning**. San Mateo, CA: Morgan Kaufmann Publishers, 1993.

RADEV, D. R. **A common theory of information fusion from multiple text sources step one: cross-document structure**. 1st SIGdial workshop on Discourse and Dialogue. **Anais...**2000.



- RAFFEL, C. et al. **Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer**. *Journal of Machine Learning Research*, v. 21, n. 140, p. 1–67, 2020.
- RAGHAVAN, U. N.; ALBERT, R.; KUMARA, S. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, v. 76, n. 3, p. 036106, 2007.
- RAMISCH, R. **Caracterização de desvios sintáticos em redações de estudantes do ensino médio: subsídios para o processamento automático das línguas naturais**. mathesis—[s.l.] Universidade Federal de São Carlos, 2020.
- RANA, M. S. et al. **Deepfake Detection: A Systematic Literature Review**. *IEEE Access*, v. 10, p. 25494–25513, 2022.
- RAU, L. F. **Extracting company names from text**. Proceedings the Seventh IEEE Conference on Artificial Intelligence Application. *Anais...IEEE Computer Society*, 1991.
- RAYNER, K. Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin - APA*, vol. 124 n. 3, p. 372–422, 1998.
- REAL, L.; FONSECA, E.; GONÇALO OLIVEIRA, H. **The ASSIN 2 Shared Task: A Quick Overview**. Computational Processing of the Portuguese Language: 14th International Conference, PROPOR 2020, Evora, Portugal, March 2–4, 2020, Proceedings. *Anais...Berlin, Heidelberg: Springer-Verlag*, 2020. Disponível em: <https://doi.org/10.1007/978-3-030-41505-1_39>
- REAL, L.; OSHIRO, M.; MAFRA, A. **B2W-Reviews01-An open product reviews corpus**. the Proceedings of the XII Symposium in Information and Human Language Technology. *Anais...2019*.
- RECUERO, R. **Redes Sociais na Internet**. [s.l.] Ciber Cultura, 2009.
- REI, R. et al. **COMET: A Neural Framework for MT Evaluation**. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). *Anais...Online: Association for Computational Linguistics*, nov. 2020. Disponível em: <<https://aclanthology.org/2020.emnlp-main.213>>
- REIS, G. B. Predição da Complexidade Textual de Notícias Jornalísticas usando uma Plataforma Crowdsourcing. *Monografia Conclusão Curso - USP*, 2017.
- RESENDE, G. et al. **(Mis)Information Dissemination in WhatsApp: Gathering, Analyzing and Countermeasures**. Proceedings of the World Wide Web Conference. *Anais...2019*.
- REYES, A.; ROSSO, P.; BUSCALDI, D. From Humor Recognition to Irony Detection: The Figurative Language of Social Media. *Data & Knowledge Engineering*, 2012.
- RIBEIRO, A. S. **O projecto MONSOON: perspectivas digitais da Índia por-**



tuguesa. Actas da Jornada de Humanidades Digitais do CIDEHUS (to appear). **Anais...**2022.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. ” **Why should i trust you?” Explaining the predictions of any classifier**. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. **Anais...**2016.

RICARTE NETO, F. A. et al. **Team PiLN at ABSAPT 2022: Lexical and BERT Strategies for Aspect-Based Sentiment Analysis in Portuguese**. Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2022), A Coruña, Spain, September 20, 2022. **Anais...**2022.

RICKARD, E. M. S. A. M. L. K. A. B. D. F. A. N. S. **Predicting Depression From Language-Based Emotion Dynamics: Longitudinal Analysis of Facebook and Twitter Status Updates**. **Journal of Medical Internet Research**, v. 20, n. 5, p. e168, 2018.

RIJSBERGEN, C. JOOST. VAN. **Information Retrieval**. [s.l.] Butterworths, 1979.

RILOFF, E. et al. **Automatically constructing a dictionary for information extraction tasks**. AAAI. **Anais...**Citeseer, 1993.

RILOFF, E.; JONES, R.; et al. **Learning dictionaries for information extraction by multi-level bootstrapping**. AAAI/IAAI. **Anais...**1999.

RINO, L. H. M.; PARDO, T. A. S. **A Sumarização Automática de textos: principais características e metodologias**. Anais do XXIII Congresso da Sociedade Brasileira de Computação. **Anais...**2003.

RO, Y.; LEE, Y.; KANG, P. **Multi²OIE: Multilingual Open Information Extraction Based on Multi-Head Attention with BERT**. Findings of the Association for Computational Linguistics: EMNLP 2020. **Anais...**Online: Association for Computational Linguistics, nov. 2020. Disponível em: <<https://aclanthology.org/2020.findings-emnlp.99>>

ROARK, B.; CHARNAIAK, E. Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction. **arXiv preprint cs/0008026**, 2000.

ROBERTSON, S. E.; SPÄRCK JONES, K. Relevance weighting of search terms. **Journal of the American Society for Information science**, v. 27, n. 3, p. 129–146, 1976.

ROBERTSON, S. E.; WALKER, S. **Okapi/keenbow at trec-8**. TREC. **Anais...**Citeseer, 1999. Disponível em: <<https://trec.nist.gov/pubs/trec8/papers/okapi.pdf>>

ROBOTTI, C. et al. Machine Learning-based Voice Assessment for the Detection of Positive and Recovered COVID-19 Patients. **Journal of Voice**, 2021.

ROCCHIO-JR, J. J. Relevance feedback in information retrieval. **The SMART retrieval**



system: experiments in automatic document processing, 1971.

RODRÍGUEZ, M. M.; BEZERRA, B. L. D. *Processamento de linguagem natural para reconhecimento de entidades nomeadas em textos jurídicos de atos administrativos (portarias)*. *Revista de Engenharia e Pesquisa Aplicada*, v. 5, n. 1, p. 67–77, 2020.

ROGERS, A.; KOVALEVA, O.; RUMSHISKY, A. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, v. 8, p. 842–866, 2021.

ROSA, G. M. et al. *Yes, bm25 is a strong baseline for legal case retrieval*. *arXiv preprint arXiv:2105.05686*, b2021.

ROSA, G. M. et al. *A cost-benefit analysis of cross-lingual transfer methods*. *arXiv preprint arXiv:2105.06813*, a2021.

ROSVALL, M.; BERGSTROM, C. T. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, v. 105, n. 4, p. 1118–1123, 2008.

ROTH, D.; YIH, W. Global inference for entity and relation identification via a linear programming formulation. *Introduction to statistical relational learning*, p. 553–580, 2007.

ROY, K.; GARG, T.; PALIT, V. *Knowledge Graph Guided Semantic Evaluation of Language Models For User Trust*. 2023 IEEE Conference on Artificial Intelligence (CAI). *Anais...*2023.

RUBIN, V. L. TALIP Perspectives, Guest Editorial Commentary: Pragmatic and Cultural Considerations for Deception Detection in Asian Languages. v. 13, n. 2, p. 10:1–10:8, 2014.

RUBIN, V. L.; CHEN, Y.; CONROY, N. J. Deception detection for news: Three types of fakes. *Proceedings of the Association for Information Science and Technology*, v. 52, n. 1, p. 1–4, 2015.

RUBIN, V. L.; CONROY, N. J. Challenges in automated deception detection in computer-mediated communication. *Proceedings of the American Society for Information Science and Technology*, v. 48, n. 1, p. 1–4, 2011.

RUPPENHOFER, J. et al. *FrameNet II: Extended Theory and Practice*. [s.l: s.n.].

RUSSELL, M. A. *Mineração de Dados da Web Social*. Primeira edição ed. São Paulo: O'Reilly Novatec, 2011.

RUSSELL-ROSE, T.; CHAMBERLAIN, J.; AZZOPARDI, L. *Information retrieval in the workplace: A comparison of professional search practices*. *Information Processing & Management*, v. 54, n. 6, p. 1042–1057, 2018.



SACRAMENTO, A. DA S. B.; SOUZA, M. **Joint Event Extraction with Contextualized Word Embeddings for the Portuguese Language**. Brazilian Conference on Intelligent Systems. **Anais...**Springer, 2021.

SAGER, N. Natural language information formatting: the automatic conversion of texts to a structured data base. Em: **Advances in computers**. [s.l.] Elsevier, 1978. v. 17p. 89–162.

SAGER, N.; FRIEDMAN, C.; LYMAN, M. S. **Medical language processing: computer management of narrative data**. [s.l.] Addison-Wesley Longman Publishing Co., Inc., 1987.

SAINZ, O. et al. NLP Evaluation in trouble: On the Need to Measure LLM Data Contamination for each Benchmark. **Findings of the Association for Computational Linguistics: EMNLP 2023**, 2023.

SAKIYAMA, K. M. **Geração Automática de Verbatações para Recuperação de Informações no Domínio Jurídico Brasileiro**. mathesis—[s.l.] Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo, 2023.

SALES ALMEIDA, T. et al. Sabiá-2: A New Generation of Portuguese Large Language Models. **arXiv e-prints**, p. arXiv–2403, 2024.

SALMINEN, J. et al. **Creating and detecting fake reviews of online products**. **Journal of Retailing and Consumer Services**, v. 64, p. 102771, 2022.

SALTON, G.; MCGILL, M. J. **Introduction to Modern Information Retrieval**. [s.l.] McGraw-Hill, 1983.

SALVI, C. et al. **Going Viral: How Fear, Socio-Cognitive Polarization and Problem-Solving Influence Fake News Detection and Proliferation During COVID-19 Pandemic**. **Frontiers in Communication**, v. 5, p. 127, 2021.

SAMY, D. **Reconocimiento y clasificación de entidades nombradas en textos legales en español**. **Procesamiento del lenguaje natural**, v. 67, p. 103–114, 2021.

SANDERSON, M. et al. Test collection based evaluation of information retrieval systems. **Foundations and Trends® in Information Retrieval**, v. 4, n. 4, p. 247–375, 2010.

SANG, E. F. T. K.; DE MEULDER, F. **Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition**. Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003. **Anais...**2003. Disponível em: <<https://aclanthology.org/W03-0419>>

SANSONE, C.; SPERLÍ, G. **Legal Information Retrieval systems: State-of-the-art and open issues**. **Information Systems**, v. 106, p. 101967, 2022.

SANTANA, B. S. **A computational-linguistic-based approach to support the**



analysis of the discursive configuration of violence on social media. tese de doutorado—[s.l.] Universidade Federal do Rio Grande do Sul, 2023.

SANTOS, C. N. DOS; GUIMARÃES, V. **Boosting Named Entity Recognition with Neural Character Embeddings.** (X. Duan et al., Eds.) Proceedings of the 5th Named Entity Workshop. **Anais...** Association for Computational Linguistics, 2015.

SANTOS, D.; CARDOSO, N. **Breve introdução ao HAREM.** (D. Santos, N. Cardoso, Eds.) Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área. **Anais...** Linguatca, a2007. Disponível em: <<http://www.linguatca.pt/LivroHAREM/>>

SANTOS, D.; CARDOSO, N. **A golden resource for named entity recognition in portuguese.** Proceeding of the 7th International conference on the computational processing of portuguese. **Anais...** Springer, b2007.

SANTOS, D.; CARDOSO, N.; SECO, N. **Avaliação no HAREM: Métodos e medidas.** (D. Santos, N. Cardoso, Eds.) Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área. **Anais...** Linguatca, 2007.

SANTOS, D.; ROCHA, P. **The key to the first CLEF with Portuguese: Topics, questions and answers in CHAVE.** Workshop of the Cross-Language Evaluation Forum for European Languages. **Anais...** b2004.

SANTOS, D.; ROCHA, P. CHAVE: topics and questions on the Portuguese participation in CLEF. a2004.

SANTOS, H. D. P. DOS et al. **Fall Detection in EHR using Word Embeddings and Deep Learning.** 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE). **Anais...** a2019.

SANTOS, H. D. P. D.; ULBRICH, A. H. D. P. S.; VIEIRA, R. **Evaluation of a Prescription Outlier Detection System in Hospital's Pharmacy Services.** 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). **Anais...** IEEE, a2021.

SANTOS, J. et al. **Assessing the Impact of Contextual Embeddings for Portuguese Named Entity Recognition.** Proceedings of the 8th Brazilian Conference on Intelligent Systems. **Anais...** b2019.

SANTOS, J. et al. **De-identification of clinical notes using contextualized language models and a token classifier.** Brazilian Conference on Intelligent Systems. **Anais...** Springer, b2021.

SANTOS, J. et al. Named entity recognition specialised for Portuguese 18th-century History research. **Proceedings of the International Conference on the Computational treatment of Portuguese, PROPOR,** a2024.



SANTOS, J.; SANTOS, H. D. P. DOS; VIEIRA, R. **Fall Detection in Clinical Notes using Language Models and Token Classifier**. (A. G. S. de Herrera et al., Eds.) Proceedings of the 33rd IEEE International Symposium on Computer-Based Medical Systems. **Anais...**a2020.

SANTOS, L. B. DOS et al. A Lightweight Regression Method to Infer Psycholinguistic Properties for Brazilian Portuguese. **International Conference on Text, Speech, and Dialogue**, p. 281–289, a2017.

SANTOS, L. B. DOS et al. **Enriching Complex Networks with Word Embeddings for Detecting Mild Cognitive Impairment from Speech Transcripts**. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais...**Vancouver, Canada: Association for Computational Linguistics, jul. b2017. Disponível em: <<https://aclanthology.org/P17-1118/>>

SANTOS, R. et al. **Measuring the Impact of Readability Features in Fake News Detection**. (N. Calzolari et al., Eds.) Proceedings of the Twelfth Language Resources and Evaluation Conference. **Anais...**Marseille, France: European Language Resources Association, b2020. Disponível em: <<https://aclanthology.org/2020.lrec-1.176>>

SANTOS, R. L. DE S. **Detecção Automática de Notícias Falsas em Português**. Ph.D. Thesis—São Carlos, Brazil: Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2022.

SANTOS, R. L. DE S.; PARDO, T. A. S. **Fact-Checking for Portuguese: Knowledge Graph and Google Search-Based Methods**. Proceedings of the 14th International Conference on the Computational Processing of Portuguese (PROPOR). **Anais...**: Lecture Notes em Artificial Intelligence (LNAI).Évora, Portugal: Springer, 2020.

SANTOS, R. L. DE S.; PARDO, T. A. S. **Structural Characterization and Graph-based Detection of Fake News in Portuguese**. Proceedings of the XIV Symposium in Information and Human Language (STIL). **Anais...**2021.

SANTOS, W. R. DOS; FUNABASHI, A. M. M.; PARABONI, I. **Searching Brazilian Twitter for signs of mental health issues**. 12th International Conference on Language Resources and Evaluation (LREC-2020). **Anais...**Marseille, France: ELRA, c2020.

SANTOS, W. R. DOS; OLIVEIRA, R. L. DE; PARABONI, I. **SetembroBR: a social media corpus for depression and anxiety disorder prediction**. **Language Resources and Evaluation**, v. 58, n. 1, p. 273–300, b2024.

SANTOS, W. R. DOS; PARABONI, I. **Predição de transtorno depressivo em redes sociais: BERT supervisionado ou ChatGPT zero-shot?** XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL-2023). **Anais...**Porto Alegre, Brasil: SBC, 2023. Disponível em: <<https://sol.sbc.org.br/index.php/stil/article/view/25433>>

SANTOS, W. R. DOS; PARABONI, I. **Prompt-based mental health screening**



- from social media text. Brazilian Workshop on Social Network Analysis and Mining (BraSNAM-2024). **Anais...**Brasilia, DF: 2024.
- SANTOS, W. R. DOS; YOON, S.; PARABONI, I. **Mental Health Prediction from Social Media Text Using Mixture of Experts**. **IEEE Latin America Transactions**, v. 21, n. 6, p. 723–729, 2023.
- SAQUETE, E. et al. **Fighting Post-truth using Natural Language Processing: A Review and Open Challenges**. **Expert Systems with Applications**, v. 141, p. 112943, 2019.
- SARACEVIC, T. Ciência da informação: origem, evolução e relações. **Perspectivas em ciência da informação**, v. 1, n. 1, 1996.
- SARAH HICKEY. **Nimdzi 100 - Language Services Industry Market Report 2020.pdf**. [s.l: s.n.].
- SARDINHA, T. B. **Linguística de Corpus: histórico e problemática**. **DELTA: Documentação de Estudos em Linguística Teórica e Aplicada**, v. 16, n. 2, p. 323–367, 2000.
- SARMENTO, M. A.; OLIVEIRA, H. T. A. DE. **Sumarização Automática de Artigos de Notícias em Português: Da Extração à Abstração com Abordagens Clássicas e Modelos de Neurais**. Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL). **Anais...**SBC, 2024.
- SCAO, T. L. et al. **BLOOM: A 176B-Parameter Open-Access Multilingual Language Model**. **CoRR**, v. abs/2211.05100, 2022.
- SCARTON, C. et al. Simplifica: a tool for authoring simplified texts in Brazilian Portuguese guided by readability assessments. **Proceedings of the 2010 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies**, p. 41–44, 2010.
- SCARTON, C. E.; ALUÍSIO, S. M. **Análise da Inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-Metrix para o Português**. **Linguamática**, v. 2, n. 1, p. 45–61, 2010.
- SCARTON, C.; SPECIA, L. Learning Simplifications for Specific Target Audiences. **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers)**, p. 712–718, 2018.
- SCHANK, R. C. et al. **MARGIE: Memory Analysis Response Generation, and Inference on English**. **IJCAI. Anais...**1973.
- SCHMITZ, M. et al. **Open language learning for information extraction**. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. **Anais...**: EMNLP-CoNLL '12. Stroudsburg, PA, USA: Association for Computational Linguistics; Association for Computational



Linguistics, 2012. Disponível em: <<http://dl.acm.org/citation.cfm?id=2390948.2391009>>

SCHNEIDER, E. T. R. et al. **BioBERTpt - A Portuguese Neural Language Model for Clinical Named Entity Recognition**. (A. Rumshisky et al., Eds.) Proceedings of the 3rd Clinical Natural Language Processing Workshop. **Anais...**Online: Association for Computational Linguistics, nov. 2020. Disponível em: <<https://aclanthology.org/2020.clinicalnlp-1.7>>

SCHNEIDER, E. T. R. et al. **A GPT-2 Language Model for Biomedical Texts in Portuguese**. 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS). **Anais...**2021.

SCHNEIDER, E. T. R. et al. **CardioBERTpt: Transformer-based Models for Cardiology Language Representation in Portuguese**. 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS). **Anais...**2023.

SCHRAAGEN, M. et al. **Evaluation of Named Entity Recognition in Dutch online criminal complaints**. **Computational Linguistics in the Netherlands Journal**, v. 7, p. 3–16, 2017.

SCHRIML, L. M. et al. Disease Ontology: a backbone for disease semantic integration. **Nucleic acids research**, v. 40, n. D1, p. D940–D946, 2012.

SCHUBERT, G.; FREITAS, L. A. DE. **A Construção de um Corpus para Detecção de Ironia e Sarcasmo em Português**. Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional. **Anais...**2020.

SCHULLER, B. W. et al. **The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primitives**. Interspeech 2021. **Anais...**2021.

SCHUSTER, M.; PALIWAL, K. K. **Bidirectional recurrent neural networks**. **IEEE transactions on Signal Processing**, v. 45, n. 11, p. 2673–2681, 1997.

SCHWARTZ, R. et al. Green ai. **Communications of the ACM**, v. 63, n. 12, p. 54–63, 2020.

SCIENCEDIRECT. **Complex Networks - an overview.**, b2025. Disponível em: <<https://www.sciencedirect.com/topics/computer-science/complex-networks>>

SCIENCEDIRECT. **Complex Networks Article.**, a2025. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0966692325001619>>

SEBASTIANI, F. Machine learning in automated text categorization. **ACM computing surveys (CSUR)**, v. 34, n. 1, p. 1–47, 2002.

SEKINE, S. **Description of the Japanese NE system used for MET-2**. Seventh



Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998. **Anais...**1998.

SELLAM, T.; DAS, D.; PARIKH, A. P. **BLEURT: Learning Robust Metrics for Text Generation**. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020. **Anais...**2020. Disponível em: <<https://doi.org/10.18653/v1/2020.acl-main.704>>

SELVARAJU, R. R. et al. **Grad-cam: Visual explanations from deep networks via gradient-based localization**. Proceedings of the IEEE international conference on computer vision. **Anais...**2017.

SEMENOV, A. et al. **Discerning Depression Propensity Among Participants of Suicide and Depression-Related Groups of Vk.com**. Analysis of Images, Social Networks and Texts. **Anais...**Cham: Springer International Publishing, 2015.

SENA, C. F. L.; CLARO, D. B. InferPortOIE: A Portuguese Open Information Extraction system with inferences. **Natural Language Engineering**, v. 25, n. 2, p. 287–306, 2019.

SENA, C. F. L.; CLARO, D. B. **PragmaticOIE: a pragmatic open information extraction for Portuguese language**. **Knowl. Inf. Syst.**, v. 62, n. 9, p. 3811–3836, 2020.

SENA, C. F. L.; GLAUBER, R.; CLARO, D. B. **Inference Approach to Enhance a Portuguese Open Information Extraction**. Proceedings of the 19th International Conference on Enterprise Information Systems - Volume 3: ICEIS. **Anais...**INSTICC; SciTePress, 2017.

SENNRICH, R.; HADDOW, B.; BIRCH, A. **Improving Neural Machine Translation Models with Monolingual Data**. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016). **Anais...**2016. Disponível em: <<https://arxiv.org/abs/1511.06709>>

SERRAS, F. et al. **Analysing and Validating Language Complexity Metrics Across South American Indigenous Languages**. (T. Kuribayashi et al., Eds.) Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics. **Anais...**Bangkok, Thailand: Association for Computational Linguistics, ago. 2024. Disponível em: <<https://aclanthology.org/2024.cmcl-1.13/>>

SHAMMA, S. A. et al. **Information Extraction from Arabic Law Documents**. 2020 IEEE 14th International Conference on Application of Information and Communication Technologies (AICT). **Anais...**2020.

SHAOWEI, Z. et al. Survey of Supervised Joint Entity Relation Extraction Methods. **Journal of Frontiers of Computer Science & Technology**, v. 16, n. 4, 2022.

SHARDLOW, M. **A Survey of Automated Text Simplification**. **International Journal of Advanced Computer Science and Applications(IJACSA)**, Special Issue on Natural Language Processing 2014, v. 4, n. 1, 2014.



- SHEIKHALISHAHI, S. et al. Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. **JMIR Med Inform**, v. 7, n. 2, p. e12239, abr. 2019.
- SHEN, G. et al. **Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution**. 26th International Joint Conference on Artificial Intelligence, IJCAI-17. **Anais...**2017.
- SHEN, J. H.; RUDZICZ, F. **Detecting Anxiety on Reddit**. Fourth Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. **Anais...**Vancouver, Canada: Association for Computational Linguistics, 2017.
- SHEN, T. et al. **Cross-Domain Depression Detection via Harvesting Social Media**. Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18. **Anais...**International Joint Conferences on Artificial Intelligence Organization, 2018.
- SHERMIS, M. D.; BURSTEIN, J. **Handbook of Automated Essay Evaluation: Current Applications and New Directions**. [s.l.] Routledge/Taylor & Francis Group, 2013.
- SHIBATA, T.; MIYAMURA, Y. **LCES: Zero-shot Automated Essay Scoring via Pairwise Comparisons Using Large Language Models**, 2025. Disponível em: <<https://arxiv.org/abs/2505.08498>>
- SHICKEL, B. et al. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. **IEEE J Biomed Health Inform**, v. 22, n. 5, p. 1589–1604, out. 2017.
- SHIMANAKA, H.; KAJIWARA, T.; KOMACHI, M. **Machine Translation Evaluation with BERT Regressor**. **arXiv**, v. abs/1907.12679, 2019.
- SHRESTHA, A.; SPEZZANO, F. **Detecting Depressed Users in Online Forums**. 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). **Anais...**2019.
- SHTERIONOV, D. et al. Human versus Automatic Quality Evaluation of NMT and PBSMT. **Machine Translation**, v. 32, n. 3, p. 217–235, 2018.
- SHUJA, J. et al. COVID-19 open source data sets: a comprehensive survey. **Applied Intelligence**, v. 51, n. 3, p. 1296–1325, 2021.
- SIDDHARTHAN, A. Syntactic Simplification and Text Cohesion. **Research on Language and Computation - Springer**, 2006.
- SILLA, C. N.; FREITAS, A. A. A survey of hierarchical classification across different application domains. **Data Mining and Knowledge Discovery**, v. 22, n. 1-2, p. 31–72, 2011.
- SILVA, A. P. DA et al. Risco de queda relacionado a medicamentos em hospitais: abor-



- dagem de aprendizado de máquina. **Acta Paulista de Enfermagem**, v. 36, a2023.
- SILVA, D. P. P. DA et al. Interpretability analysis of deep models for COVID-19 detection. **Artificial Intelligence in Health**, v. 1, n. 3, p. 114–126, 2024.
- SILVA, F. DA et al. **Towards analysis on textual inference at ASSIN-2 dataset**. Proceedings of the 14th Brazilian Symposium in Information and Human Language Technology. **Anais...**2023.
- SILVA, F. L. V. DA et al. **ABSAPT 2022 at IberLEF: Overview of the Task on Aspect-Based Sentiment Analysis in Portuguese**. **Procesamiento del Lenguaje Natural**, v. 69, p. 199–205, a2022.
- SILVA, F. N. et al. **Using network science and text analytics to produce surveys in a scientific topic**. **Journal of Informetrics**, v. 10, n. 2, p. 487–502, 2016.
- SILVA, F. R. A. DA. **Deteccção de Ironia e Sarcasmo em Língua Portuguesa: uma abordagem utilizando Deep Learning**. <https://github.com/fabio-ricardo/deteccaoo-ironia>, 2018.
- SILVA, I. A. L. DA et al. Translation, post-editing and directionality. **Translation in transition: Between cognition, computing and technology**, p. 107–134, a2017.
- SILVA, J. F. F. **Estratégias para sumarização de documentos**. mathesis—[s.l.] (Mestrado em Engenharia informática) - Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa, 2022.
- SILVA, M. J.; CARVALHO, P.; SARMENTO, L. **Building a Sentiment Lexicon for Social Judgement Mining**. Proceedings of the 10th International Conference on Computational Processing of the Portuguese Language. **Anais...**2012.
- SILVA, N. F. F. DA et al. **Evaluating Topic Models in Portuguese Political Comments About Bills from Brazil's Chamber of Deputies**. (A. Britto, K. Valdivia Delgado, Eds.)Intelligent Systems. **Anais...**Cham: Springer International Publishing, 2021.
- SILVA, N. L. DA; DI FELIPPO, A. **Descrição e Análise do Fenômeno da Contradição para a Sumarização Automática Multidocumento**. [s.l.] Série de Relatórios Técnicos do Núcleo Interinstitucional de Linguística Computacional, 2014.
- SILVA, R. M. et al. **Towards Automatically Filtering Fake News in Portuguese**. **Expert Systems with Applications**, v. 146, p. 1–48, b2020.
- SILVA, R. M. et al. **Towards Automatically Filtering Fake News in Portuguese**. **Expert Systems With Applications**, v. 146, p. 1–14, a2020.
- SILVA, R. M. et al. Fake News Detection in Portuguese Under Large Language Model-Generated Content. **Journal of the Brazilian Computer Society**, p. 1–18, 2025.



- SILVA, R. M.; ALMEIDA, T. A. **How concept drift can impair the classification of fake news**. Proceedings of the 9th Symposium on Knowledge Discovery, Mining and Learning (KDMiLe'21). **Anais...**2021.
- SILVA, R. M.; ALMEIDA, T. A.; YAMAKAMI, A. MDLText: An Efficient and Lightweight Text Classifier. **Knowledge-Based Systems**, v. 118, p. 152–164, b2017.
- SILVA, R. M.; PIRES, P. R.; ALMEIDA, T. A. Incremental learning for fake news detection. **Journal of Information and Data Management – JIDM**, v. 13, p. 566–579, b2022.
- SINGH, D.; GARG, R. **A survey of Community Detection algorithms and its comparative performance analysis**. **Computer Science Review**, v. 58, 2025.
- SIQUEIRA, F. et al. **Ulysses Tesemão: a new large corpus for Brazilian legal and governmental domain**. **Language Resources and Evaluation**, p. 1–20, jul. 2024.
- SJÖHOLM, J. **Probability as readability: A new machine learning approach to readability assessment for written Swedish**. [s.l.] LiU Electronic Press, 2012.
- Slator 2019 Language Industry Market Report. p. 33, 2019.
- SLEIMI, A. et al. **An automated framework for the extraction of semantic legal metadata from legal texts**. **Empirical Software Engineering**, v. 26, p. 1–50, 2021.
- SMIRNOVA, A.; CUDRÉ-MAUROUX, P. Relation extraction using distant supervision: A survey. **ACM Computing Surveys (CSUR)**, v. 51, n. 5, p. 1–35, 2018.
- SMITH, K. S. **On Integrating Discourse in Machine Translation**. Proceedings of the Third Workshop on Discourse in Machine Translation. **Anais...**2017.
- SMYWIŃSKI-POHL, A. et al. **Automatic Extraction of Amendments from Polish Statutory Law**. Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law. **Anais...**: ICAIL '21. New York, NY, USA: Association for Computing Machinery, 2021.
- SNOVER, M. G. et al. **A Study of Translation Edit Rate with Targeted Human Annotation**. Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, AMTA 2006, Cambridge, Massachusetts, USA, August 8-12, 2006. **Anais...**2006. Disponível em: <<https://aclanthology.org/2006.amta-papers.25/>>
- SOARES, M. O que é letramento? **Presença Pedagógica Volume 2, n. 10**, p. 15–25, 1996.
- SOBAHI, N. et al. Explainable COVID-19 detection using fractal dimension and vision transformer with Grad-CAM on cough sounds. **Biocybernetics and Biomedical Engineering**, v. 42, n. 3, p. 1066–1080, 2022.



- SOCHER, R. et al. **Semantic compositionality through recursive matrix-vector spaces**. Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. **Anais...**2012.
- SODERLAND, S. et al. **CRYSTAL inducing a conceptual dictionary**. Proceedings of the 14th international joint conference on Artificial intelligence-Volume 2. **Anais...**1995.
- SOLORIO, T. MALINCHE: A NER system for Portuguese that reuses knowledge from Spanish. **Reconhecimento de entidades mencionadas em português: Documentação e atas do HAREM, a primeira avaliação conjunta na área, Capítulo**, v. 10, p. 123–136, 2007.
- SONG, H. et al. **Feature Attention Network: Interpretable Depression Detection from Social Media**. 32nd Pacific Asia Conference on Language, Information and Computation. **Anais...**Hong Kong: Association for Computational Linguistics, 2018.
- SOUSA, A. et al. **Cross-Lingual Annotation Projection for Argument Mining in Portuguese**. (G. Marreiros et al., Eds.)Progress in Artificial Intelligence. **Anais...**Springer International Publishing, 2021.
- SOUSA, M. C. P. DE. O Corpus Tycho Brahe: contribuições para as humanidades digitais no Brasil. **Filologia e linguística portuguesa**, v. 16, n. esp., p. 53–93, 2014.
- SOUSA, R. F. DE; BRUM, H. B.; NUNES, M. DAS G. V. **A bunch of helpfulness and sentiment corpora in brazilian portuguese**. Proceedings of Symposium in Information and Human Language Technology. **Anais...**2019.
- SOUZA, E. et al. **An Information Retrieval Pipeline for Legislative Documents from the Brazilian Chamber of Deputies**. Em: **Legal Knowledge and Information Systems**. [s.l.] IOS Press, 2021a. p. 119–126.
- SOUZA, E. N. P. DE; CLARO, D. B.; GLAUBER, R. A Similarity Grammatical Structures Based Method for Improving Open Information Systems. **j-jucs**, v. 24, n. 1, p. 43–69, 28 jan. 2018.
- SOUZA, E. N. P.; CLARO, D. B. **Extração de Relações utilizando Features Diferenciadas para Português**. **Linguamática**, v. 6, n. 2, p. 57–65, 2014.
- SOUZA, F. D.; FILHO, J. B. DE O. E. S. **BERT for sentiment analysis: Pre-trained and fine-tuned alternatives**. International Conference on Computational Processing of the Portuguese Language. **Anais...**Springer, 2022.
- SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. **BERTimbau: pretrained BERT models for Brazilian Portuguese**. (R. Cerri, R. C. Prati, Eds.)Proceedings of the 2020 Brazilian Conference on Intelligent Systems. **Anais...**Springer International Publishing, a2020.
- SOUZA FREIRE, P. M.; MATIAS DA SILVA, F. R.; GOLDSCHMIDT, R. R. **Fake news detection based on explicit and implicit signals of a hybrid crowd: An approach inspired**



in *meta-learning*. **Expert Systems with Applications**, v. 183, p. 115414, 2021.

SOUZA, J. W. DA C. **Aprofundamento da caracterização linguístico-computacional da complementaridade em um corpus jornalístico multidocumento**. tese de doutorado—[s.l.] (Doutorado em Linguística) - Programa de Pós-Graduação em Linguística, Universidade Federal de São Carlos, 2019.

SOUZA, J. W. DA C.; FELIPPO, A. D. Caracterização da complementaridade temporal: subsídios para sumarização automática multidocumento. **Alfa: Revista de Linguística (São José do Rio Preto)**, v. 62, p. 125–150, 2018.

SOUZA, J. W. DA C.; FELIPPO, A. D.; PARDO, T. A. S. **Investigação da Identificação da Redundância na Sumarização Multidocumento**. Anais do III Student Workshop on Information and Human Language Technology. **Anais...**a2011.

SOUZA, M. et al. **Construction of a Portuguese Opinion Lexicon from multiple resources**. Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology. **Anais...**b2011.

SOUZA, V.; NOBRE, J.; BECKER, K. **Characterization of Anxiety, Depression, and their Comorbidity from Texts of Social Networks**. Anais do XXXV Simpósio Brasileiro de Bancos de Dados. **Anais...**Porto Alegre, Brazil: SBC, b2020.

SOUZA, V.; NOBRE, J.; BECKER, K. A Deep Learning Ensemble to Classify Anxiety, Depression, and their Comorbidity from Texts of Social Networks. **Journal of Information and Data Management**, v. 12, n. 3, p. 306–325, b2021.

SPÄRCK JONES, K. Report on the need for and provision of an 'ideal' information retrieval test collection. **Computer Laboratory**, 1975.

SPÄRCK JONES, K.; WALKER, S.; ROBERTSON, S. E. A probabilistic model of information retrieval: development and comparative experiments. **Information processing & management**, v. 36, n. 6, p. 809–840, 2000.

SPARCK-JONES, K. **Automatic Summarizing: Factors and Directions**. In Mani, I. And Maybury, M., editors, **Advances in Automatic Text Summarization**. MIT Press, 1998.

SPEER, R.; CHIN, J.; HAVASI, C. **ConceptNet 5.5: an open multilingual graph of general knowledge**. Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. **Anais...**: AAAI'17.San Francisco, California, USA: AAAI Press, 2017.

SPEICHER, T. et al. **Potential for discrimination in online targeted advertising**. Proceedings of the Conference on Fairness, Accountability and Transparency. **Anais...**ACM, 2018.

SPRINGEROPEN APPLIED NETWORK SCIENCE. **Small-world and scale-free properties in networks.**, 2017. Disponível em: <<https://appliednetsci.springeropen.c>



om/articles/10.1007/s41109-017-0054-z>

SRIVASTAVA, N. et al. Dropout: a simple way to prevent neural networks from overfitting. **The journal of machine learning research**, v. 15, n. 1, p. 1929–1958, 2014.

SRIVASTAVA, S.; SHARMA, G. **OmniVec2 - A Novel Transformer Based Network for Large Scale Multimodal and Multitask Learning**. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). **Anais...**2024.

STANOJEVIC, M.; SIMA'AN, K. **BEER: BEtter Evaluation as Ranking**. Proceedings of the Ninth Workshop on Statistical Machine Translation, WMT@ACL 2014, June 26-27, 2014, Baltimore, Maryland, USA. **Anais...**2014. Disponível em: <<https://doi.org/10.3115/v1/w14-3354>>

STANOVSKY, G. et al. **Supervised open information extraction**. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). **Anais...**2018.

STEYVERS, M.; TENENBAUM, J. B. **The large-scale structure of semantic networks: statistical analyses and a model of semantic growth**. **Cognitive science**, v. 29, n. 1, p. 41–78, 2005.

STJ. **Supremo Tribunal de Justiça — Composição**. <https://www.stj.jus.br/sites/porta/p/Institucional/Composicao>, maio 24DC.

STRIEN, D. VAN et al. **Assessing the impact of OCR quality on downstream NLP tasks**. ICAART 2020 - Proceedings of the 12th International Conference on Agents and Artificial Intelligence. **Anais...**2020.

SU, J.; CARDIE, C.; NAKOV, P. **Adapting Fake News Detection to the Era of Large Language Models**. Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2024. **Anais...**2024.

SU, K.-Y.; WU, M.-W.; CHANG, J.-S. **A new quantitative quality measure for machine translation systems**. Proceedings of the 14th conference on Computational linguistics -. **Anais...**Association for Computational Linguistics, 1992. Disponível em: <<http://dx.doi.org/10.3115/992133.992137>>

SUN, C.; EMONET, V.; DUMONTIER, M. **A Comprehensive Comparison of Automated FAIRness Evaluation Tools**. (K. Wolstencroft et al., Eds.)13th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences, SWAT4HCLS 2022, Virtual Event, Leiden, The Netherlands, January 10th to 14th, 2022. **Anais...**: CEUR Workshop Proceedings.CEUR-WS.org, 2022. Disponível em: <<http://ceur-ws.org/Vol-3127/paper-6.pdf>>

SUTSKEVER, I.; VINYALS, O.; LE, Q. V. **Sequence to Sequence Learning with Neural Networks**. (Z. Ghahramani et al., Eds.)Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014,



December 8-13 2014, Montreal, Quebec, Canada. **Anais...**2014. Disponível em: <<https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html>>

SWEET, P. L. Every bone of my body: Domestic violence and the diagnostic body. **Social Science & Medicine**, v. 122, p. 44–52, 2014.

TAKAMATSU, S.; SATO, I.; NAKAGAWA, H. **Reducing wrong labels in distant supervision for relation extraction**. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais...**2012.

TAN, K. L.; LEE, C. P.; LIM, K. M. A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research. **Applied Sciences**, 2023.

TANDOC JR., E. C.; LIM, Z. W.; LING, R. **Defining “Fake News”**. **Digital Journalism**, v. 6, n. 2, p. 137–153, 2018.

TAUS. **TAUS - The Translation Industry in 2022 Report.**, 2020. Disponível em: <<https://info.taus.net/translation-industry-2022-report-download>>. Acesso em: 19 ago. 2020

TAVARES, L. L.; SILVA, R. M.; ALMEIDA, T. A. **Towards Automatically Creating Large Labeled Datasets for Training Question Domain Classifiers**. 2018 International Joint Conference on Neural Networks (IJCNN). **Anais...**2018. Disponível em: <<https://doi.org/10.1109/ijcnn.2018.8489124>>

TEAM, G. et al. **Gemma: Open Models Based on Gemini Research and Technology.**, 2024. Disponível em: <<https://arxiv.org/abs/2403.08295>>

TEIXEIRA, E. N.; FONSECA, M. C. M.; SOARES, M. E. Resolução do pronome nulo em Português Brasileiro: Evidência de movimentação ocular. **VEREDAS: Sintaxe das Línguas Brasileiras**, v. 18, 2014.

TEIXEIRA, F. et al. Classification of control/pathologic subjects with support vector machines. **Procedia computer science**, v. 138, p. 272–279, 2018.

TEIXEIRA, S. C. S. B.; MARENGO, S. M. D. A.; FINATTO, M. J. B. **Construindo fichas terminológicas para estudos sócio-históricos**. **Revista Diálogos**, v. 10, n. 3, p. 261–279, 2022.

TEIXEIRA, S. H.; ZAMORA, M. H. Pensando a interseccionalidade a partir da vida e morte de Marielle Franco. **Dignidade Re-Vista**, 2019.

THAKAR, H.; BHATT, B. **Fake News Detection: Recent Trends and Challenges**. **Social Network Analysis and Mining**, v. 14, 2024.

THOMAS, C. et al. Automatic Detection and Rating of Dementia of Alzheimer Type through Lexical Analysis of Spontaneous Speech. **Proceedings of the IEEE Interna-**



tional Conference on Mechatronics and Automation, p. 1569–1574, 2005.

THORNE, J.; VLACHOS, A. **Automated Fact Checking: Task Formulations, Methods and Future Directions.** (E. M. Bender, L. Derczynski, P. Isabelle, Eds.) Proceedings of the 27th International Conference on Computational Linguistics. **Anais...**Santa Fe, New Mexico, USA: Association for Computational Linguistics, ago. 2018.

TIRRELL, L. Toxic Speech: Inoculations and Antidotes. **The Southern Journal of Philosophy**, 2018.

TOHALINO, J. V.; AMANCIO, D. R. **Extractive multi-document summarization using multilayer networks.** **Physica A: Statistical Mechanics and its Applications**, v. 503, p. 526–539, 2018.

TOLLES, J.; MEURER, W. J. **Logistic Regression: Relating Patient Characteristics to Outcomes.** **JAMA**, v. 316, n. 5, p. 533–534, ago. 2016.

TORAL, A. et al. **Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation.** Proceedings of WMT. **Anais...**Brussels, Belgium: 2018.

TORRENT, T. T. et al. **Representing Context in FrameNet: A Multidimensional, Multimodal Approach.** **Frontiers in Psychology**, v. Volume 13 - 2022, 2022.

TORRENT, T. T. et al. **A flexible tool for a qualia-enriched FrameNet: the FrameNet Brasil WebTool.** **Language Resources and Evaluation**, jan. 2024.

TORRENT, T. T.; CONEGLIAN, A. V. L. **Caligrama: Revista de Estudos Românicos**, v. 30, n. 1, p. 8–22, 2025.

TRAAG, V. A.; WALTMAN, L.; ECK, N. J. VAN. **From Louvain to Leiden: guaranteeing well-connected communities.** **Scientific Reports**, v. 9, n. 1, p. 5233, 2019.

TRAJANO, D.; BORDINI, R. H.; VIEIRA, R. OLID-BR: offensive language identification dataset for Brazilian Portuguese. **Language Resources and Evaluation**, 2023.

TRIFU, R. et al. **Linguistic indicators of language in major depressive disorder (MDD). An evidence based research.** **Journal of Evidence-Based Psychotherapies**, v. 17, p. 105–128, mar. 2017.

TROTZEK, M.; KOITKA, S.; FRIEDRICH, C. M. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. **IEEE Transactions on Knowledge and Data Engineering**, 2018.

TSOUMAKAS, G.; KATAKIS, I. Multi-label classification: An overview. **International Journal of Data Warehousing and Mining**, v. 3, n. 3, p. 1–13, 2007.



- TSUGAWA, S. et al. **Recognizing Depression from Twitter Activity**. 33rd Annual ACM Conference on Human Factors in Computing Systems. **Anais...**New York, USA: Association for Computing Machinery, 2015.
- TURCHIOE, M. R. et al. **Systematic review of current natural language processing methods and applications in cardiology**. **Heart**, v. 108, n. 12, p. 909–916, 2022.
- UCHIDA, H.; ZHU, M.; DELLA SENTA, T. A gift for a millennium. **IAS/UNU, Tokyo**, 1999.
- USZKOREIT, H.; LOMMEL, A. **Multidimensional Quality Metrics: A New Unified Paradigm for Human and Machine Translation Quality Assessment**. [s.l: s.n.].
- VAJJALA, S.; MEURERS, D. Readability-based Sentence Ranking for Evaluating Text Simplification. **CoRR**, v. abs/1603.06009, 2016.
- VAN DER LAAN, R. H. **Tesouro e terminologia: uma inter-relação lógica**. Tese (Doutorado em Letras)—Porto Alegre: Universidade Federal do Rio Grande do Sul; Instituto de Letras, Programa de Pós-Graduação em Letras, 2002.
- VARELA, P. J.; JUSTINO, E. J.; OLIVEIRA, L. E. **Identificação de Autoria de Textos através do uso de Classes Linguísticas da Língua Portuguesa (Authorship Identification Using Linguistic Classes for Portuguese)**[in Portuguese]. Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology. **Anais...**2011.
- VARGAS, D. F.; VAN DER LANN, R. H. A contribuição da terminologia na construção de linguagens documentárias como os tesouros. **Biblos**, v. 25, n. 1, p. 21–34, 2011.
- VARGAS, F. et al. **HateBR: A Large Expert Annotated Corpus of Brazilian Instagram Comments for Offensive Language and Hate Speech Detection**. Proceedings of the Thirteenth Language Resources and Evaluation Conference. **Anais...**b2022.
- VARGAS, F. et al. **HateBR: A Large Expert Annotated Corpus of Brazilian Instagram Comments for Offensive Language and Hate Speech Detection**. Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022). **Anais...**Marseille, France: European Language Resources Association, a2022. Disponível em: <<https://aclanthology.org/2022.lrec-1.777>>
- VARGAS, F. et al. **Rhetorical Structure Approach for Online Deception Detection: A Survey**. (N. Calzolari et al., Eds.)Proceedings of the Thirteenth Language Resources and Evaluation Conference. **Anais...**Marseille, France: European Language Resources Association, jun. c2022. Disponível em: <<https://aclanthology.org/2022.lrec-1.635>>
- VARGAS, F. et al. **Predicting Sentence-Level Factuality of News and Bias of Media Outlets**. (R. Mitkov, G. Angelova, Eds.)Proceedings of the 14th International



Conference on Recent Advances in Natural Language Processing. **Anais...**Varna, Bulgaria: INCOMA Ltd., Shoumen, Bulgaria, set. 2023. Disponível em: <<https://aclanthology.org/2023.ranlp-1.127>>

VARGAS, F. et al. **Improving Explainable Fact-Checking via Sentence-Level Factual Reasoning**. (M. Schlichtkrull et al., Eds.) Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER). **Anais...**Miami, Florida, USA: Association for Computational Linguistics, nov. 2024. Disponível em: <<https://aclanthology.org/2024.fever-1.23/>>

VARGAS, F. A.; SANTOS, R. S. S. D.; ROCHA, P. R. **Identifying Fine-Grained Opinion and Classifying Polarity on Coronavirus Pandemic**. Proceedings of the Brazilian Conference on Intelligent Systems. **Anais...**2020.

VARGAS, F.; PARDO, T.; BENEVENUTO, F. **Socially Responsible and Explainable Automated Fact-Checking and Hate Speech Detection**. Anais do XXXVIII Concurso de Teses e Dissertações. **Anais...**Porto Alegre, RS, Brasil: SBC, 2025. Disponível em: <<https://sol.sbc.org.br/index.php/ctd/article/view/36355>>

VAROL, O. et al. **Online human-bot interactions: Detection, estimation, and characterization**. Proceedings of the International AAAI Conference on Web and Social Media. **Anais...**AAAI Press, 2017.

VASWANI, A. et al. **Attention is All you Need**. (I. Guyon et al., Eds.) Advances in Neural Information Processing Systems. **Anais...**Curran Associates, Inc., b2017. Disponível em: <<https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>>

VASWANI, A. et al. Attention is all you need. **Advances in neural information processing systems**, v. 30, a2017.

VIEIRA, C.; SOUZA, J.; CAVALCANTI, G. **Deteção de Fake News em Português: Análise Comparativa entre Métodos de Representação em Português, Inglês e Multilíngues**. Anais do XIV Brazilian Workshop on Social Network Analysis and Mining. **Anais...**Porto Alegre, RS, Brasil: SBC, 2025. Disponível em: <<https://sol.sbc.org.br/index.php/brasnam/article/view/36378>>

VIEIRA, J. M. M. **The Brazilian Portuguese eye tracking corpus with a predictability study focusing on lexical and partial prediction**. mathesis—Universidade Federal do Ceará, Biblioteca Universitária: Federal University of Ceará (UFC), 2020.

VIEIRA, R. et al. Enriching the 1758 Portuguese Parish Memories (Alentejo) with Named Entities. **Journal of Open Humanities Data**, v. 7, p. 20, 2021.

VILAR, D. et al. **Error Analysis of Statistical Machine Translation Output**. Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06). **Anais...**Genoa, Italy: European Language Resources Association (ELRA), 2006. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2006/pdf/413_pdf.pdf>



f>

VILLAR, G. S.; FINATTO, M. J. B. Acessibilidade textual e terminológica: novos glosários sobre oncologia para a ferramenta MedSimples. **Mandinga-Revista de Estudos Linguísticos (ISSN: 2526-3455)**, v. 7, n. 2, p. 23–42, 2023.

VIMAL, B. et al. **MFCC Based Audio Classification Using Machine Learning**. 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT). **Anais...**2021.

VIRIDIANO, M. et al. **Framed Multi30K: A Frame-Based Multimodal-Multilingual Dataset**. (N. Calzolari et al., Eds.)Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). **Anais...**Torino, Italia: ELRA; ICCL, 2024. Disponível em: <<https://aclanthology.org/2024.lrec-main.656/>>

VITÓRIO, D. et al. **Ulysses-RFSQ: A Novel Method to Improve Legal Information Retrieval Based on Relevance Feedback**. (J. C. Xavier-Junior, R. A. Rios, Eds.)Intelligent Systems. **Anais...**Cham: Springer International Publishing, 2022.

VITÓRIO, D. et al. **Building a relevance feedback corpus for legal information retrieval in the real-case scenario of the Brazilian Chamber of Deputies**. **Language Resources and Evaluation**, 2024.

VLACHOS, A.; RIEDEL, S. **Fact checking: Task definition and dataset construction**. Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science. **Anais...**Association for Computational Linguistics, 2014.

VOSOUGHI, S.; ROY, D.; ARAL, S. **The spread of true and false news online**. **Science**, v. 359, n. 6380, p. 1146–1151, 2018.

WAGNER FILHO, J. A. et al. **The brWaC Corpus: A New Open Resource for Brazilian Portuguese**. (N. Calzolari et al., Eds.)Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). **Anais...**Miyazaki, Japan: European Language Resources Association (ELRA), 2018. Disponível em: <<https://aclanthology.org/L18-1686>>

WANDERLEY, M. G. et al. **A Moving Target: Detecting Concept Drift in Brazilian Portuguese Fake News**. Proceedings of the 16th Symposium in Information and Human Language Technology (STIL'2025). **Anais...**Fortaleza, CE, Brazil: SBC, out. 2025.

WANG, L. et al. **Relation classification via multi-level attention cnns**. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais...**2016.

WANG, S. et al. **Want To Reduce Labeling Cost? GPT-3 Can Help**. (M.-F. Moens et al., Eds.)Findings of the Association for Computational Linguistics: EMNLP 2021.



Anais... Punta Cana, Dominican Republic: Association for Computational Linguistics, nov. 2021.

WANG, S. et al. **GPT-NER: Named Entity Recognition via Large Language Models.**, 2023. Disponível em: <<https://arxiv.org/abs/2304.10428>>

WANG, W. Y. **“Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection.** Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). **Anais...** Vancouver, Canada: Association for Computational Linguistics, jul. 2017.

WARDLE, C.; DERA KHSHAN, H. **Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making.** [s.l.] Council of Europe, 2017. Disponível em: <<https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-research/168076277c>>.

WASSERMAN, S.; FAUST, K. **Social network analysis: Methods and applications.** [s.l.] Cambridge university press, 1994.

WATANABE, W. M. et al. **Facilita: helping the reading of texts available on the web.** XV Brazilian Symposium on Multimedia and the Web, WebMedia '09, Fortaleza, Ceará, Brazil, October 5-7, 2009. **Anais...**a2009. Disponível em: <<http://doi.acm.org/10.1145/1858477.1858516>>

WATANABE, W. M. et al. **Facilita: reading assistance for low-literacy readers.** Proceedings of the 27th Annual International Conference on Design of Communication, SIGDOC 2009, Bloomington, Indiana, USA, October 5-7, 2009. **Anais...**b2009. Disponível em: <<http://doi.acm.org/10.1145/1621995.1622002>>

WATASE, T. et al. Severity Classification Using Dynamic Time Warping-Based Voice Biomarkers for Patients With COVID-19: Feasibility Cross-Sectional Study. **JMIR Biomedical Engineering**, v. 8, n. 1, p. e50924, 2023.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of “small-world” networks. **Nature**, v. 393, n. 6684, p. 440–442, 1998.

WAY, A. **Quality Expectations of Machine Translation.** Em: MOORKENS, J. et al. (Eds.). **Translation Quality Assessment: From Principles to Practice.** Cham: Springer International Publishing, 2018. p. 159–178.

WAY, A.; FORCADA, M. L. **Editors’ foreword to the invited issue on SMT and NMT. Machine Translation**, v. 32, n. 3, p. 191–194, set. 2018.

WEI, J. et al. Emergent abilities of large language models. **arXiv preprint arXiv:2206.07682**, 2022.

WEI, X. et al. **ChatIE: Zero-Shot Information Extraction via Chatting with ChatGPT.**, 2024. Disponível em: <<https://arxiv.org/abs/2302.10205>>



WHO. **Comprehensive mental health action plan 2013–2030**. [s.l.] World Health Organization; World Health Organization, 2021.

WHO. **World Health Organization**. World Health Organization, 2024. Disponível em: <https://www.who.int/health-topics/violence-against-women#tab=tab_2>. Acesso em: 6 abr. 2024

WILKINSON, M.; DUMONTIER, M.; AALBERSBERG, ET AL. The FAIR Guiding Principles for scientific data management and stewardship. **Scientific data**, v. 3, n. 1, p. 1–9, 2016.

WILKINSON, M.; DUMONTIER, M.; SANSONE, ET AL. **Evaluating FAIR maturity through a scalable, automated, community-governed framework**. **Sc. Data**, v. 6, n. 1, p. 1–12, 2019.

WIVES, L. K. **Técnicas de Recuperação de Informações Com Ênfase em Informações Textuais**. tese de doutorado—[s.l.] Universidade Federal do Rio Grande do Sul, 1997.

WOLFRAM, W. **Variation and Language: Overview**. Em: BROWN, K. (Ed.). **Encyclopedia of Language & Linguistics (Second Edition)**. Second Edition ed. Oxford: Elsevier, 2006. p. 333–341.

WOLINSKI, F.; VICHOT, F.; DILLET, B. **Automatic processing proper names in texts**. Proc. Conference on European Chapter of the Association for Computational Linguistics. **Anais...EACL**, 1995.

WU, H. et al. SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. **J Am Med Inform Assoc**, v. 25, n. 5, p. 530–537, 2018.

WU, Y. et al. Google’s neural machine translation system: Bridging the gap between human and machine translation. **arXiv preprint arXiv:1609.08144**, 2016.

XAVIER, C. C.; LIMA, V. L. S. DE; SOUZA, M. Open information extraction based on lexical semantics. **Journal of the Brazilian Computer Society**, v. 21, n. 1, p. 1–14, 2015.

XAVIER, R. C. **Português no Direito: Linguagem Forense**. Rio de Janeiro: Forense, 2002. p. 1

XIA, P. et al. **LOME: Large Ontology Multilingual Extraction**. (D. Gkatzia, D. Seddah, Eds.) Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. **Anais...Online: Association for Computational Linguistics**, abr. 2021. Disponível em: <<https://aclanthology.org/2021.eacl-demos.19>>

XIA, T.; HAN, J.; MASCOLO, C. Exploring machine learning for audio-based respiratory



condition screening: A concise review of databases, methods, and open issues. **Experimental Biology and Medicine**, v. 247, n. 22, p. 2053–2061, 2022.

XIAO, C. et al. **Human-AI Collaborative Essay Scoring: A Dual-Process Framework with LLMs**. Proceedings of the 15th International Learning Analytics and Knowledge Conference. **Anais...**: LAK '25. New York, NY, USA: Association for Computing Machinery, 2025.

XU, S. et al. **Automated Lightweight Model for Asthma Detection Using Respiratory and Cough Sound Signals**. **Diagnostics**, v. 15, n. 9, 2025.

XU, W.; CALLISON-BURCH, C.; NAPOLES, C. **Problems in Current Text Simplification Research: New Data Can Help**. **Transactions of the Association for Computational Linguistics**, v. 3, p. 283–297, 2015.

YADAV, S. et al. **Identifying Depressive Symptoms from Tweets: Figurative Language Enabled Multitask Learning Framework**. 28th International Conference on Computational Linguistics. **Anais...** Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020.

YAN, M. Y.; GUSTAD, L. T.; NYTRØ, Ø. Sepsis prediction, early detection, and identification using clinical text for machine learning: a systematic review. **J Am Med Inform Assoc**, v. 29, n. 3, p. 559–575, jan. 2022.

YANG, H. et al. Clinical Trial Classification of SNS24 Calls with Neural Networks. **Future Internet**, v. 14, n. 5, p. 130, 2022.

YANG, K.-C. et al. Scalable and generalizable social bot detection through data selection. **Proceedings of the AAAI Conference on Artificial Intelligence**, v. 34, n. 01, p. 1096–1103, a2020.

YANG, P.; FANG, H.; LIN, J. **Anserini: Enabling the use of lucene for information retrieval research**. Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval. **Anais...** 2017.

YANG, R. et al. **Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking**. Association for Computational Linguistics (ACL), b2020.

YASUNAGA, M. et al. **Deep Bidirectional Language-Knowledge Graph Pretraining**, 2022. Disponível em: <<https://arxiv.org/abs/2210.09338>>

YATES, A.; COHAN, A.; GOHARIAN, N. **Depression and Self-Harm Risk Assessment in Online Forums**. Conference on Empirical Methods in Natural Language Processing. **Anais...** Copenhagen, Denmark: Association for Computational Linguistics, 2017.

YAZDAVAR, A. H. et al. **Semi-Supervised Approach to Monitoring Clinical De-**



pressive Symptoms in Social Media. IEEE/ACM International Conference on Advances in Social Network Analysis and Mining. **Anais...**2017.

YU, X.; LAM, W. **Jointly identifying entities and extracting relations in encyclopedia text via a graphical model approach.** Coling 2010: Posters. **Anais...**2010.

YUAN, W.; NEUBIG, G.; LIU, P. **BARTScore: Evaluating Generated Text as Text Generation.** Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual. **Anais...**a2021. Disponível em: <<https://proceedings.neurips.cc/paper/2021/hash/e4d2b6e6fdeca3e60e0f1a62fee3d9dd-Abstract.html>>

YUAN, Y. et al. **A relation-specific attention network for joint entity and relation extraction.** International joint conference on artificial intelligence. **Anais...**International Joint Conference on Artificial Intelligence, b2021.

YUE, Z. et al. **Evidence-Driven Retrieval Augmented Response Generation for Online Misinformation.** (K. Duh, H. Gomez, S. Bethard, Eds.)Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). **Anais...**Mexico City, Mexico: Association for Computational Linguistics, jun. 2024. Disponível em: <<https://aclanthology.org/2024.naacl-long.313>>

ZANUZ, L.; RIGO, S. J. **Fostering Judiciary Applications with New Fine-Tuned Models for Legal Named Entity Recognition in Portuguese.** (V. Pinheiro et al., Eds.)Computational Processing of the Portuguese Language. **Anais...**Cham: Springer International Publishing, 2022.

ZAROCOSTAS, J. How to fight an infodemic. **The lancet**, v. 395, n. 10225, p. 676, 2020.

ZELENINA, M. **Eye Tracking for NLP.** SlideShare, 2015. Disponível em: <<https://www.slideshare.net/mariezelenina/presentation-2-47610828>>

ZELENKO, D.; AONE, C.; RICHARDELLA, A. Kernel methods for relation extraction. **Journal of machine learning research**, v. 3, n. Feb, p. 1083–1106, 2003.

ZENG, D. et al. **Relation classification via convolutional deep neural network.** Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers. **Anais...**2014.

ZHANG, A. et al. **Dive into Deep Learning.** [s.l.] Cambridge University Press, 2023.

ZHANG, H. **The Optimality of Naive Bayes.** Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference. **Anais...**2004.

ZHANG, S. et al. **Opt: Open pre-trained transformer language models.** **arXiv preprint arXiv:2205.01068**, 2022.



ZHANG, S. X. et al. **Predictors of Depression and Anxiety Symptoms in Brazil during COVID-19.** *Int J Environ Res Public Health*, v. 18, n. 13, 30 jun. 2021.

ZHANG, S.; DUH, K.; VAN DURME, B. **Cross-lingual open information extraction with neural sequence-to-sequence models.** Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. *Anais...*2017.

ZHANG, T. et al. **BERTScore: Evaluating Text Generation with BERT.** 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. *Anais...*OpenReview.net, 2020. Disponível em: <<https://openreview.net/forum?id=SkeHuCVFDr>>

ZHANG, Y. et al. **A Comprehensive Survey on Process-Oriented Automatic Text Summarization with Exploration of LLM-Based Methods.**, 2025. Disponível em: <<https://arxiv.org/abs/2403.02901>>

ZHAO, S.; GRISHMAN, R. **Extracting relations with integrated information using kernel methods.** Proceedings of the 43rd annual meeting of the association for computational linguistics (acl'05). *Anais...*2005.

ZHONG, H. et al. **How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence.** Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. *Anais...*Association for Computational Linguistics, 2020.

ZHOU, L. et al. A comparison of classification methods for predicting deception in computer-mediated communication. *Journal of Management Information Systems*, v. 20, n. 4, p. 139–165, 2004.

ZHOU, L. An empirical investigation of deception behavior in instant messaging. *IEEE Transactions on Professional Communication*, v. 48, n. 2, p. 147–160, 2005.

ZHUANG, F. et al. **A comprehensive survey on transfer learning.** *Proceedings of the IEEE*, v. 109, n. 1, p. 43–76, 2020.

ZILIO, L.; FINATTO, M. J.; VIEIRA, R. **Named Entity Recognition Applied to Portuguese Texts from the XVIII Century.** (C. Trojahn et al., Eds.)Proceedings of the Second Workshop on Digital Humanities and Natural Language Processing (2nd DHandNLP 2022) co-located with International Conference on the Computational Processing of Portuguese (PROPOR 2022), Virtual Event, Fortaleza, Brazil, 21st March, 2022. *Anais...*: CEUR Workshop Proceedings.CEUR-WS.org, 2022. Disponível em: <<http://ceur-ws.org/Vol-3128/paper10.pdf>>

ZILIO, L.; LAZZARI, R. R.; FINATTO, M. J. B. NLP for historical Portuguese: Analysing 18th-century medical texts. **Proceedings of the International Conference on the Computational treatment of Portuguese, PROPOR**, 2024.

ZIPF, G. K. **Human behavior and the principle of least effort: an introduction**



to human ecology. [s.l.] Addison-Wesley Press, 1949.

ZOBEL, J. **How reliable are the results of large-scale information retrieval experiments?** Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. **Anais...ACM**, 1998.



Apêndice A

Considerações prévias à sumarização automática

Apêndice do Capítulo 5

Jackson Wilke da Cruz Souza
Paula Christina Figueira Cardoso
Crysttian Arantes Paixão

Publicado em: 13/03/2024

Ao analisarmos os métodos e sistemas de sumarização, torna-se necessário estabelecer e detalhar alguns procedimentos importantes para trabalhar com sumarização de textos. Muitas vezes, o usuário de sistemas de SA tem em mente o que deseja realizar com o seu texto, porém nem sempre se atenta aos detalhes que podem influenciar no processamento, análise e resultados.

A) Formalidades e variações da língua

O primeiro detalhe é conhecer a qualidade do texto que se deseja fazer a sumarização. A origem do texto determina a qualidade e o padrão da grafia textual. Um texto pode vir de diferentes fontes para compor um *corpus* ou corpora; Por exemplo, em textos de redes sociais, nos quais o processo de escrita não segue, necessariamente a variedade, é formal da língua e possui limitação de caracteres, dependendo do público que gerou as mensagens, demandará uma atenção especial. Uma base de texto que siga uma padronização, de preferência normalizado, reduzirá em muito o tratamento das inconsistências, otimizando a análise (cf. Capítulo 2).

Problemas envolvendo acentuação são os mais comuns, desconsiderando o problema de grafia. Vamos ao exemplo do uso da palavra “não”. Em mensagens do Twitter, um usuário pode escrevê-la de várias formas para transmitir a mensagem, como “n”, “não” e “nao”, sendo que todas essas variações correspondem ao “não”. Para resolver esse problema, o uso de dicionário seria uma opção para corrigir as palavras (corretor ortográfico), como existem nos editores de textos tradicionais.

B) Obtenção dos textos-fonte

Dependendo da origem dos textos, alguns problemas podem ser encontrados. No caso das redes sociais, quando se possui uma interface de programação de aplicações - API, o texto é coletado de forma automatizada. Porém, os problemas com a falta de padronização ainda podem persistir.

Alguns pesquisadores realizam estudos envolvendo apenas textos que possuem uma normalização, exemplo os que analisam trabalhos de conclusão de curso - TCC, dissertações e teses. Existem ainda os que utilizam artigos, de revistas e jornais, para realizarem suas pesquisas.

Nesses textos, por conta da formalização textual, muitos problemas são minimizados. Entretanto, eles ainda podem ocorrer. Um exemplo é quando a base é composta por



textos que seguem a formação dos programas Word da Microsoft, Write do OpenOffice, documentos do Google e em PDF (*Portable Document Format*). Esses formatos utilizam uma linguagem para a formatação dos textos que não é visível ao usuário, como *Extensible Markup Language* - XML do docx do Word, que tem que ser tratada no momento da importação para análise, pois apenas o texto deve compor o *corpus*. Para visualizar a fonte de um arquivo docx, sugerimos que o arquivo docx analisado seja aberto em um editor de texto simples, como o bloco de notas do Windows. No caso de arquivos no formato PDF, a conversão também é necessária. Existem ainda textos que são oriundos de páginas da internet que são capturadas, nos quais a linguagem que estrutura o site na maioria das vezes, é a Linguagem de Marcação de HiperTexto - HTML.

Esse exemplo é fácil de se constatar. Abra uma página da internet, selecione o texto e cole no Word. Você notará que a formatação das palavras também foi mantida. A pergunta é: como? A resposta é simples: junto com o texto desejado, você copiou um texto de marcação que irá compor o seu documento; logo, este deve ser tratado antes de realizar a análise.

Uma forma de contornar esse problema é copiar o texto da internet, colar em um editor de textos simples e, em seguida, copiar novamente o texto do editor para o Word. Dessa forma, a linguagem invisível ao olhar humano é removida. Porém, trata-se de um serviço manual e dependente da quantidade de textos, o que o torna inviável. Assim, em algum momento nesse processo, será necessário utilizar recursos computacionais para o processamento textual. Em suma, independente da fonte, deve-se realizar uma limpeza dos arquivos, o que usualmente recebe o nome de pré-processamento, deixando apenas o texto a ser analisado.

Uma sugestão é que os textos a serem analisados, na medida do possível, estejam em formato de texto puro, como os criados em blocos de notas. Recomenda-se sempre que, após a importação ou carregamento de uma base textual, uma inspeção visual seja realizada em uma amostra do texto para a verificação de possíveis problemas.

C) Fator cultural

Dependendo da região do país, existem fatores culturais que impactam diretamente a forma de expressão, como o uso de gírias. Portanto, em alguns casos, como desejamos monitorar o que está repercutindo, devemos padronizar essa informação. Para isso, torna-se necessário o uso de léxicos de gírias. O léxico serve como uma orientação na troca das palavras de mesmo significado para se estabelecer um padrão. Basicamente, o léxico funcionaria como um “corretor ortográfico”, estabelecendo o padrão textual para análise. Como exemplo, temos o uso dos termos “pão de sal”, “pão”, “pãozinho”, “cassquinho”, entre outros, que referem-se a mesma coisa.

O desafio imposto é a obtenção desses léxicos. Na maioria dos casos, o usuário deverá criá-lo, do zero ou a partir de algum outro que contemple os termos que deseje, pois dificilmente encontrará, salvo exceções, um léxico específico para sua aplicação.

D) Emojis

Outro fato relevante em textos de redes sociais é o uso de emojis, ideogramas e smileys. O usuário se vale de tais recursos em suas mensagens para transmitir uma informação relevante, o que em determinada parte do texto pode dificultar o processo de sumarização, demandando configurações especiais. No processo de SA, os emojis serão considerados como palavras e, dependendo da análise, deverá ser utilizado um dicionário de sinônimos para descrevê-los.

Além disso, consideremos outros sistemas, nos quais o usuário de certas redes sociais pode se expressar diretamente com outros elementos multimodais (como vídeos e fotos).



A interação dos usuários também leva em conta o regionalismo, que torna necessário a adaptação de um léxico para tentar minimizar o efeito de idiossincrasias em função da informalidade textual. As temáticas abordadas devem ser consideradas para determinar se o texto pode compor um sumário.

E) Expressões regulares

Aos leitores interessados em análise de textos, recomendamos um estudo, pelo menos introdutório, sobre expressões regulares. Como referência, indicamos o livro de Expressões Regulares (Jargas, 2016). O livro apresenta de forma simples os fundamentos de expressões e suas aplicações. Basicamente, uma expressão regular é utilizada para representar um padrão, o qual é muito utilizado dentro das linguagens de programação que podem ser executadas no processamento textual.

Um exemplo é o caso da palavra “não” com diversas ocorrências que podem ser representadas pela expressão regular “[Nn][aã]?o?”, referindo-se às variações “Não”, “Nao”, “N”, “não”, “nao” e “n”. Grande parte desse pré-processamento depende do uso de funções da linguagem de programação, que recebem como parâmetro uma expressão regular para realizar as operações. O uso dessas funções otimiza o processamento textual, permitindo padronizar o texto, reduzindo problemas e viabilizando as análises.

F) Pré-processamento

Até o momento, relatamos o problema envolvendo a criação da base de dados. Porém, antes de analisar o texto, devem ser feitas, como capitalização, acentuação, símbolos de pontuação e *stopwords*, por exemplo

Algumas perguntas podem surgir ao se pensar que uma palavra escrita com letras maiúsculas e minúsculas é diferente. Como exemplo, temos: o computador diferencia “Carro” e “carro”? ou existe um tipo de caixa para otimizar o processamento? Você já sabe a resposta! A forma com a qual são armazenadas é o que as diferencia. Lembre-se sempre que o computador não possui discernimento entre o significado da palavra, mas sim de como ela é representada. Nesse caso, letras maiúsculas e minúsculas possuem codificações diferentes. O mesmo ocorre com a acentuação.

Já para os símbolos de pontuação, em algum determinado momento, o usuário deve escolher qual o caractere que servirá de delimitador de palavras (*tokens*, em PLN). Em alguns idiomas, esse delimitador é o espaço em branco, como na frase “O dia estava lindo, mas pode chover no período da tarde”. Facilmente, você consegue identificar as palavras que compõem essa oração. Porém o computador, dependendo da forma como for programado, pode considerar que a palavra “lindo” não esteja presente na frase, por conta da pontuação. Nesse caso “lindo,” com destaque para a presença da vírgula, seria a palavra considerada. A vírgula também é processada como um caractere, logo pode influenciar a análise em decorrência que “lindo” é diferente de “lindo,”. Isso ocorre muito em nuvens de palavras quando não se atenta pela remoção desses caracteres de pontuação. Logo, uma palavra com determinada frequência poderá ocorrer um número maior de vezes no texto, enviesando a análise.

Mais um detalhe a ser analisado é a forma da escrita que constitui o texto. Apesar de ocorrer em menor frequência, temos alguns itens a serem sempre analisados. O primeiro é o uso de separação silábica, pois a palavra “casa” é diferente de “ca-sa” quando inserida em um texto e processada pelo computador, pois o “-” é um símbolo.

Dependendo da situação, o texto ainda tem de ser pré-processado para eliminar palavras sem valor informacional, comumente denominadas *stopwords*. No caso de SA, nem sempre existe a eliminação dessa “classe” de palavras, mas vale chamar a atenção. Muitas vezes, essa lista de palavras a serem removidas consta de um léxico já padronizado.



Porém, em determinados domínios e públicos torna-se necessário atualizá-lo ou adaptá-lo. Como exemplo, considere um estudo que envolva a palavra-chave “COVID” a ser coletada em mensagens no Twitter. As mensagens que serão coletadas terão a palavra “COVID” em seu texto, sendo que não há adição de informação no seu uso, podendo ser removida do corpus. Todos os textos, de forma geral, terão essa palavra e dependendo da análise, ela é dispensável, ou seja, é interessante removê-la.



Apêndice B

Exemplo de sumarização abstrativa

Apêndice do Capítulo 5

Pedro Henrique Paiola

Gabriel Lino Garcia

João Vitor Mariano Correia

João Renato Ribeiro Manesco

Danilo Samuel Jodas

Douglas Rodrigues

João Paulo Papa

Publicado em: 16/04/2026

Este apêndice apresenta os exemplos de sumarização omitidos do texto do capítulo de sumarização automática.

Quadro B.1: Sumários gerados por cada modelo



Modelo	Texto
Documento1 Problema em sonda japonesa	Leia mais: Segundo a empresa, quando a sonda atingiu uma altitude de cerca de 100 km, ela ativou corretamente a sequência de descida. No entanto, o software estimou erroneamente que a sua altitude era zero quando ela ainda pairava a cerca de 5 km do chão. Ou seja, a Hakuto-R pensava que já havia pousado, mas seguiu descendo lentamente a uma velocidade inferior a 1 m/s. Por fim, seu sistema de propulsão ficou sem combustível e, apesar de nesse momento a equipe não conseguir mais contato com a espaçonave, acredita-se que ela caiu em queda livre em solo lunar. Confusão é a razão mais provável para esse erro do software. Enquanto voava para o local de pouso, a espaçonave passou pela borda de uma cratera. Lá, o sensor obteve uma leitura de altitude de 3 km e, aparentemente, isso foi maior do que o valor de altitude estimado anteriormente pela equipe que projetou a missão. Por isso, o software da sonda pensou erroneamente que o sensor relatou uma medida anormal e rejeitou-a. Isso ocorreu porque o programa tinha a capacidade de desconsiderar valores anormais de altitude no módulo de pouso como medida de segurança para caso houvesse um problema com o sensor. Com isso, infelizmente, o software falhou ao incorporar o verdadeiro terreno lunar na rota de pouso da espaçonave. Se a aterrissagem da sonda japonesa tivesse sido bem sucedida, essa seria a primeira missão privada e o primeiro veículo construído no Japão a pousar na Lua. Apesar do fracasso, a ispace usará os dados dessa missão para melhorar suas próximas espaçonaves com destino ao satélite natural, com lançamentos previstos em 2024 e 2025. Layse Ventura é jornalista formada pela Uerj e mestre em Engenharia e Gestão do Conhecimento pela Ufsc. Acumula mais de 10 anos de experiência como repórter, copywriter e analista de SEO.
Sumário de referência	Cratera confundiu sonda japonesa que caiu na Lua, revela ispace. Essa seria a primeira missão privada e o primeiro veículo construído no Japão a pousar na Lua.
PTT5-Summ:	Ispace tenta pousar na Lua com erro de software. O software da ispace tentou incorporar o terreno lunar na rota de pouso da espaçonave.
Qwen3-8B:	A sonda lunar Hakuto-R, desenvolvida pela empresa ispace, falhou em sua missão de pouso na Lua devido a um erro no software. Quando atingiu cerca de 100 km de altitude, o sistema ativou a sequência de descida, mas estimou incorretamente a altitude como zero, mesmo que a sonda estivesse a 5 km do solo. Como resultado, a sonda continuou descendo lentamente, sem combustível, e caiu em queda livre. O erro provavelmente ocorreu porque o sensor de altitude capturou uma leitura de 3 km em uma cratera, que foi rejeitada pelo software como anormal, levando a uma falha na navegação. Apesar do fracasso, os dados da missão serão usados para melhorar futuras missões, previstas para 2024 e 2025.
InternLM-Chatbode-7B:	A sonda japonesa Hakuto-R, que foi lançada para a Lua em 2021, caiu em queda livre após um erro de software que estimou erroneamente que a sonda havia pousado. O software estimou que a sonda tinha pousado quando ela ainda estava voando a uma distância de 5 km do chão. O sistema de propulsão da sonda ficou sem combustível e ela caiu em queda livre. A ispace, a empresa que operou a missão, usará os dados dessa missão para melhorar suas próximas espaçonaves com destino ao satélite natural, com lançamentos previstos em 2024 e 2025.

Sabiá-7B:

A missão Hakuto-R, da empresa ispace, foi a primeira missão privada a pousar na Lua. A sonda japonesa, que foi lançada em 2019, tinha como objetivo realizar um pouso suave na superfície lunar. No entanto, o pouso não foi bem sucedido, e a sonda acabou caindo em uma cratera. Apesar do fracasso, a ispace usará os dados dessa missão para melhorar suas próximas espaçonaves com destino ao satélite natural, com lançamentos previstos em 2024 e 2025.

Fonte:

<https://www.tecmundo.com.br/ciencia/229069-missao-hakuto-r-falha-na-lua-e-ispac-vai-usar-dados-para-melhorar-missao-de-2024.htm>

Texto 2:

A missão Hakuto-R, da empresa ispace, foi a primeira missão privada



<https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/>



Sobre as/os autoras/es

Helena de Medeiros Caseli

Helena Caseli é formada em Ciência da Computação pela UFU, com pós-graduação no ICMC/USP. É professora titular na UFSCar, onde atua desde 2008 e coordena o **LALIC** (Laboratório de Linguística e Inteligência Computacional). Suas áreas de interesse incluem aprendizado de máquina, tradução automática e PLN aplicado a redes sociais em domínios como política e saúde mental. Foi a idealizadora, é uma das organizadoras deste livro e colaborou com os Capítulos **O que é PLN?**, **Sequência de caracteres e palavras**, **E agora, PLN?**, 4 e 10.

Email: helenacaseli@ufscar.br

Maria das Graças Volpe Nunes

Maria das Graças Volpe Nunes é formada em Ciência da Computação pela UFSCar, mestre pela USP e doutora pela PUC-Rio. Foi docente no ICMC-USP, São Carlos, de 1981 a 2013, onde agora é professora sênior. Atualmente é membro do C4AI, da USP, IBM e FAPESP. É uma das fundadoras do **NILC**, onde atuou em várias áreas do PLN para português: revisão ortográfica e gramatical, tradução automática, sumarização, análise de sentimentos, entre outras. É uma das organizadoras deste livro e colaborou com os Capítulos **O que é PLN?**, **Questões éticas em IA e PLN** e **E agora, PLN?**.

Email: gracan@icmc.usp.br

Adriana Silvina Pagano

Adriana Pagano é formada em Letras pela Universidad Nacional de La Plata. Fez seu mestrado no Programa de Inglês e Literatura Correspondente da UFSC e seu doutorado no Programa de Estudos Linguísticos da UFMG. É professora titular da Faculdade de Letras da UFMG desde 1995. Suas áreas de interesse incluem tradução, produção textual multilíngue e modelagem sistêmico-funcional da linguagem. Neste livro, colaborou com os Capítulos **O que é PLN?**, **A ordem e a função das palavras em uma sentença**, **Ferramentas e recursos para o processamento sintático** e 9.

Email: apagano@ufmg.br

Adriano L. I. Oliveira

Adriano Oliveira é engenheiro eletrônico (1993), mestre e doutor em Computação pela UFPE (1997 e 2004). Foi Professor Visitante da École de Technologie Supérieure (ÉTS, Université du Québec, Canadá), 2018-2019. É Professor da UFPE e bolsista de produtividade em pesquisa do CNPq, nível 1D. Neste livro, colaborou com o Capítulo **13**.

Email: alio@cin.ufpe.br



Aline Macohin

Aline Macohin é formada em Direito e Computação, com Mestrado em Computação Aplicada pela UTFPR e Doutorado em Direito pela UFPR. É advogada, analista de sistemas no poder público há 13 anos, membro da Comissão de Direito Digital e Proteção de dados da OAB/PR e professora de pós-graduação na PUCPR. Autora do site <https://www.iaresponsavel.com.br>. Suas áreas de interesse incluem aprendizado de máquina, PLN, governança de inteligência artificial e proteção de dados. Neste livro, colaborou com o Capítulo 12.

Email: aline.macohin@gmail.com

Amanda Pontes Rassi

Amanda Rassi é bacharel e mestre em Linguística pela Universidade Federal de Goiás (UFG) e doutora na mesma área pela Universidade Federal de São Carlos (UFSCar). Tem experiência com correção automática de redações, tendo atuado na Redação Nota 1000 e atualmente na Somos Educação. Suas áreas de interesse incluem sintaxe, semântica, lexicologia, lexicografia e PLN. Neste livro, colaborou com os Capítulos *Sequência de caracteres e palavras*, *A ordem e a função das palavras em uma sentença* e 7.

Email: amanda.rassi@somoseducacao.com.br

Ana Paula Banza

Ana Paula Banza é Licenciada em Línguas e Literaturas Modernas pela Universidade de Lisboa (Portugal); Doutora em Linguística pela Universidade de Évora (Portugal), onde lecciona, no Departamento de Linguística e Literaturas, desde 1991, sendo, atualmente, professora associada com agregação. Desenvolve investigação nas áreas da Filologia, Crítica Textual, Linguística Histórica/História da Língua e Variação linguística, centrando-se os seus interesses na edição crítica de textos literários e não literários, com destaque para a obra do Padre António Vieira, e no estudo dos fenómenos de mudança e de variação na língua portuguesa. Neste livro, colaborou com o Capítulo 15.

Email: anabanza@uevora.pt

Ana Sofia Ribeiro

Ana Sofia Ribeiro é formada em História e doutorada em História pela Universidade do Porto. É investigadora auxiliar do CIDEHUS, Universidade de Évora, Portugal. As suas áreas de interesse estão ligadas aos métodos de utilização de PLN aos textos históricos e como os resultados de técnicas como o reconhecimento de entidades nomeadas (NER) ou grafos de conhecimento (*knowledge graphs*) podem ser usados como variáveis da investigação em história. Neste livro, colaborou com o Capítulo 15.

Email: asvribeiro@uevora.pt

André Carlos Ponce de Leon Ferreira de Carvalho

André Carlos de Carvalho é professor titular do ICMC-USP e Vice-Diretor do Centro de Inteligência Artificial e Aprendizado de Máquina (CIAAM) da USP. Seus principais



interesses de pesquisa são Aprendizado de Máquina, Mineração de Dados e Ciência de Dados, com aplicações em várias áreas. Escreveu vários livros, entre eles *Inteligência Artificial: Uma abordagem de Aprendizado de Máquina*, publicado pelo GrupoGen em 2011 e prêmio Jabuti 2012, e *A General Introduction to Data Analytics*, publicado pela Wiley, em 2018. Neste livro, colaborou com o Capítulo 13.

Email: andre@icmc.usp.br

Beatriz Raposo de Medeiros

Beatriz Raposo de Medeiros é professora associada de Fonética da Universidade de São Paulo (USP), onde coordena o Laboratório de Fonética e Linguagem Lafalin. É doutora em Linguística pela Universidade Estadual de Campinas (Unicamp). Realizou pós-doutorado no Laboratoire Langue Parole, na Université Aix-Marseille em 2007. Em 2023 recebeu a BPE, bolsa de pesquisa da FAPESP, em período como professora visitante da University of Southern California, para estudo da laringe e sua relação mais estreita com capacidades entoacionais humanas. Seu trabalho de pesquisa abrange comparações entre fala e canto, tratando aspectos como a inteligibilidade, a organização temporal, a sincronização e a estabilidade da frequência fundamental. Neste livro, colaborou com o Capítulo 11.

Email: biarm@usp.br

Brenda Salenave Santana

Brenda Salenave Santana é formada em Ciência da Computação pela Universidade Federal de Santa Maria (UFSM), com mestrado e doutorado em Computação pela Universidade Federal do Rio Grande do Sul (UFRGS). É docente na Universidade Federal de Pelotas (UFPel) desde 2023. Suas áreas de interesse incluem aprendizado de máquina e detecção de discursos de ódio. Neste livro, colaborou com o Capítulo 16.

Email: bssalenave@inf.ufpel.edu.br

Cassia Trojahn dos Santos

Cassia Trojahn dos Santos é formada em Ciência da Computação pela Unisc, mestre pela Unisinos e doutora pela Universidade de Evora (Portugal), com pós-doutorado pelo INRIA (França). Foi docente do Departamento de Informática e Matemática na Universidade de Grenoble (França) entre 2012 e 2024. Atualmente, é docente na Universidade de Grenoble Alpes, associada ao centro de pesquisa INRIA Alpes. Suas áreas de interesse incluem representação de conhecimento, alinhamento de ontologias, e dados FAIR. Neste livro, colaborou com o Capítulo 15.

Email: cassia.ts@gmail.com

Celso Ricardo Fernandes de Carvalho

Celso Ricardo Fernandes de Carvalho é formado em Fisioterapia pela UFSCar. Atualmente é Professor Associado da Faculdade de Medicina da Universidade de São Paulo (FMUSP). Tem interesse pelas áreas Processamento de Linguagem Natural e Inteligência Artificial em pessoas com doenças respiratórias. Neste livro, colaborou com o Capítulo 11.

Email: cscarval@usp.br



Claudia Maria Cabral Moro Barra

Claudia Moro é engenheira de computação pela PUCPR, com mestrado (UNICAMP) e doutorado (USP) em engenharia elétrica/biomédica. É professora titular da PUCPR onde atua desde 1995, vinculada ao Programa de Pós-graduação em Tecnologia em Saúde (desde 2003), atuando nas áreas de PLN em textos clínicos, inteligência artificial em saúde e, interoperabilidade entre sistemas de informações em saúde. Neste livro, colaborou com o Capítulo 9.

Email: claudia.moro@gmail.com

Crysttian Arantes Paixão

Crysttian Arantes Paixão é formado em Ciência da Computação pela Universidade Federal de Lavras (UFLA) e Estatística pela Anhembí Morumbi, com pós-graduação em Administração de Redes Linux, Modelagem Computacional e Estatística na UFLA e Pós-Doutorado em Modelagem Computacional pela Fundação Getúlio Vargas. É docente na Universidade Federal da Bahia (UFBA) desde 2022. Suas áreas de interesse incluem Estatística, Modelagem Computacional e Inteligência Artificial. Neste livro, colaborou com o Capítulo 5.

Email: crysttian@gmail.com

Daniela Barreiro Claro

Daniela Barreiro Claro é professora titular da Universidade Federal da Bahia (UFBA), fez seu Mestrado em Ciências da Computação pela Universidade Federal de Santa Catarina (UFSC) e o seu Doutorado na Université d'Angers (França). Em 2009, ela fundou o Grupo de Pesquisa FORMAS na área de Semântica e PLN. Suas principais áreas de pesquisa incluem extração de informação, aprendizado de máquina e aprendizado multimodal. Neste livro, colaborou com os Capítulos [Representação vetorial e semântica distribucional](#) e [3](#).

Email: dclaro@ufba.br

Danilo Samuel Jodas

Danilo Jodas é formado em Ciência da Computação pelo Centro Universitário do Norte Paulista (UNORP), com pós-graduação em processamento de imagens e aprendizado de máquina pela UNESP e FEUP (Universidade do Porto). É pesquisador de pós-doutorado na UNESP de Bauru desde 2019, atuando em pesquisas de aprendizado de máquina e PLN. Neste livro, colaborou com os Capítulos [5](#) e [Treinamento de Grandes Modelos de Linguagem na prática](#).

Email: danilo.jodas@unesp.br

Diego Raphael Amancio

Diego Raphael Amancio é doutor em Física Computacional pela USP (2013), tendo realizado estágios de pós-doutorado na própria USP (2014) e na Indiana University, nos Estados Unidos (2018). Atualmente, é professor associado (livre-docente) e pesquisador do ICMC/USP. Possui experiência na área de Ciência da Computação, com ênfase em



Redes Complexas, Cienciometria, Aprendizado de Máquina e PLN. Seus interesses de pesquisa incluem a modelagem de sistemas reais por meio de grafos, com destaque para a representação e análise da linguagem humana. Neste livro, colaborou com o Capítulo 18.
Email: diego@icmc.usp.br

Douglas Rodrigues

Douglas Rodrigues é formado em Informática para Gestão de Negócios pela FATEC/Botucatu, Mestrado (UNESP/Bauru) e Doutorado (UFSCar/São Carlos) em Ciências da Computação. Atualmente, atua como pós-doutorando e pesquisador no Recogna/UNESP. Suas áreas de interesse incluem aprendizado de máquina e otimização matemática. Neste livro, colaborou com os Capítulos 5 e [Treinamento de Grandes Modelos de Linguagem na prática](#).

Email: d.rodrigues@unesp.br

Elisa Terumi Rubel Schneider

Elisa Terumi é formada em Informática pela Universidade Tecnológica Federal do Paraná (UTFPR), com especialização em Desenvolvimento Web (PUCPR), mestrado em Bioinformática (UFPR) e doutorado em Informática atuando na linha de pesquisa de Inteligência Artificial (PUCPR). Tem interesse em PLN aplicado na área de saúde. Neste livro, colaborou com os Capítulos [Ferramentas e recursos para o processamento sintático](#) e 9.

Email: lisa.terumi@gmail.com

Ellen Souza

Ellen Souza é formada em Ciência da Computação com doutorado em inteligência artificial pelo CIn/UFPE e pós-doutoranda na área de processamento de textos legislativos no ICMC-USP desde 2020. É docente do curso de Bacharelado em Sistemas de Informação da UFRPE desde 2009. Suas áreas de interesse incluem aprendizado de máquina, recuperação de informação e reconhecimento de entidades nomeadas. Neste livro, colaborou com o Capítulo 13.

Email: ellen.ramos@ufrpe.br

Evelin Amorim

Evelin Amorim é formada em Ciência da Computação pela Universidade Federal do Espírito Santo (UFES), com doutorado em Ciência da Computação pela Universidade Federal de Minas Gerais (UFMG). É pesquisadora no INESC TEC, em Portugal, desde 2020. Suas áreas de interesse incluem correção automática de redações e extração de informação. Neste livro, colaborou com o Capítulo 1.

Email: evelin.amorim@gmail.com



Fátima Farrica

Fátima Farrica é doutorada em História pela Universidade de Évora (Portugal), com pós-graduação em Ciências Documentais: Arquivologia. É bolsista de investigação do CIDEHUS-Universidade de Évora (Portugal). As suas áreas de interesse incluem a investigação na área da História, projetos de arquivística, paleografia e tratamento documental. Neste livro, colaborou com o Capítulo 15.

Email: ffarrica@uevora.pt

Felipe Ribas Serras

Felipe Ribas Serras é Bacharel em Física pelo Instituto de Física da Universidade de São Paulo (USP) e Mestre em Ciência da Computação pelo Instituto de Matemática e Estatística da USP. Suas áreas de interesse incluem complexidade de linguagem e abordagens computacionais à variação linguística e tipologia. Neste livro, colaborou com os capítulos 6, *O papel dos dados no pré-treinamento de Grandes Modelos de Linguagem*.

Email: frserras@ime.usp.br

Fernanda Olival

Fernanda Olival é doutorada em História pela Universidade de Évora (Portugal), onde leciona desde 1991. É também diretora do CIDEHUS (Portugal). Trabalha sobre História Social e Política do período Moderno. Interessa-se por Processamento de Linguagem Natural (PLN) como ferramenta para anotar, gerir e recuperar informação, a partir de fontes históricas. Neste livro, colaborou com o Capítulo 15.

Email: mfo@uevora.pt

Flaviane Romani Fernandes Svartman

Flaviane Svartman possui graduação e doutorado em Linguística pela Unicamp, com período de estágio de doutorado na Faculdade de Letras da Universidade de Lisboa e pós-doutorado também pela Unicamp. Atua como docente na área de Filologia e Língua Portuguesa na Faculdade de Filosofia, Letras e Ciências Humanas da USP desde 2010. Suas investigações concernem ao estudo da fonologia e da fonética da língua portuguesa, com especial interesse na prosódia, na interface sintaxe-fonologia, na comparação entre variedades africanas, brasileiras e europeias do português e no PLN. Neste livro, colaborou com os Capítulos *Recursos para o processamento de fala* e 11.

Email: flaviane@gmail.com

Gabriel Lino Garcia

Gabriel Lino Garcia é graduado em Análise e Desenvolvimento de Sistemas pela FATEC e mestre em Inteligência Artificial (IA) e Processamento de Linguagem Natural (PLN) pela UNESP, onde faz doutorado focado na aplicação de modelos de linguagem na área médica. Seus interesses incluem aprendizado de máquina, grandes modelos de linguagem, aprendizado multimodal e aplicações médicas. Neste livro, colaborou com os Capítulos 5 e *Treinamento de Grandes Modelos de Linguagem na prática*.



Email: linaogolino@gmail.com

Helena Freire Cameron

Helena Freire Cameron é doutora em Linguística pela Universidade de Aveiro (Portugal) e pós-doutora em Humanidades Digitais e PLN pela Universidade de Évora (Portugal). Integrou vários projetos de investigação em Linguística Computacional (EUROTRA, DIGRAMA, Corpus Lexicográfico do Português, entre outros). É investigadora do CIDEHUS-Universidade de Évora (Portugal), onde integra um grupo multidisciplinar de Humanidades Digitais, envolvendo História, Linguística e Computação-PLN. Nos anos mais recentes tem-se dedicado ao estudo da variação ortográfica no século XVIII, do ponto de vista linguístico e computacional-PLN. Neste livro, colaborou com o Capítulo 15.

Email: helenac@ipportalegre.pt

Hidelberg O. Albuquerque

Hidelberg Albuquerque é formado em Análise e Desenvolvimento de Sistemas pelo Instituto Federal da Paraíba (IFPB), com pós-graduação pelo PPGI/UFPB, e atualmente doutorando em Ciência da Computação no Cin/UFPE. É docente na Unidade Acadêmica de Serra Talhada (UAST)/UFRPE desde 2015. Suas áreas de interesse incluem aprendizado de máquina, reconhecimento de entidade nomeada (NER), sumarização e recuperação de informação. Neste livro, colaborou com o Capítulo 13.

Email: hidelberg.albuquerque@ufrpe.br

Ivandr  Paraboni

Ivandr  Paraboni   doutor em Ci ncia da Computa  o pela University of Brighton (Inglaterra, 2003), com est gio de p s-doutorado junto   University of Aberdeen (Esc cia, 2012).   professor associado (livre docente) e pesquisador da USP/EACH. Possui experi ncia na  rea de Ci ncia da Computa  o, atuando principalmente nas  reas de PLN e IA. Interesses de pesquisa abrangem o processamento da l ngua humana em um sentido amplo do termo, incluindo m todos computacionais derivados das Ci ncias Cognitivas e aplica  es pr ticas de classifica  o de documentos extra dos da WEB. Neste livro, colaborou com os Cap tulos *Gera  o de linguagem natural* e 10.

Email: ivandre@usp.br

Jackson Wilke da Cruz Souza

Jackson Souza   bacharel (2012), mestre (2015) e doutor (2019) em Lingu stica pela Universidade Federal de S o Carlos (UFSCar). Atualmente,   professor da Universidade Federal da Bahia (UFBA), atuando junto ao Departamento de Ci ncia, Tecnologia e Inova  o e ao Programa de P s-Gradua  o em L ngua e Cultura da UFBA. Tem interesse nas  reas de an lise de discurso, sumariza  o autom tica e an lise e descri  o lingu stica. Neste livro, colaborou com os Cap tulos *Modelos discursivos* e 5.

Email: jackcruzsouza@gmail.com



João Paulo Papa

João Paulo Papa é formado em Sistemas de Informação pela UNESP, com pós-graduação em processamento de imagens e aprendizado de máquina pela UFSCar e Unicamp, respectivamente. Possui pós-doutorados na Unicamp, Harvard e MIT. É professor titular da Unesp onde atua desde 2009 e suas áreas de interesse incluem visão computacional, processamento de imagens, aprendizado de máquina e aprendizado multimodal. Neste livro, colaborou com os Capítulos 5 e [Treinamento de Grandes Modelos de Linguagem na prática](#).

Email: joao.papa@unesp.br

João Renato Ribeiro Manesco

João Renato Ribeiro Manesco é doutorando em Ciência da Computação pela UNESP, onde também concluiu mestrado e graduação. Durante o mestrado, realizou estágio de pesquisa no Media Integration and Communication Center da Universidade de Florença, Itália. Seus interesses de pesquisa incluem adaptação de domínio, modelos de linguagem generativos e multimodais, e aplicações em contexto médico. Neste livro, colaborou com os Capítulos 5 e [Treinamento de Grandes Modelos de Linguagem na prática](#).

Email: joao.r.manesco@unesp.br

João Vitor Mariano Correia

João Vitor Mariano Correia é bacharel em Ciência da Computação pela UNESP, onde atualmente cursa o mestrado no Programa de Pós-Graduação em Ciência da Computação. Seus principais interesses de pesquisa estão na área de Processamento de Linguagem Natural (PLN), com foco em representações de conhecimento. Neste livro, colaborou com os Capítulos 5 e [Treinamento de Grandes Modelos de Linguagem na prática](#).

Email: mariano.correia@unesp.br

Joaquim Santos

Joaquim Santos fez licenciatura em Matemática na Universidade Regional do Cariri (URCA), é mestre em Ciência da Computação pela PUCRS e atualmente é estudante de doutorado no programa de Computação Aplicada da UNISINOS. Tem trabalhado intensamente com Processamento de Linguagem Natural (PLN) em problemas de extração de informação, modelos de linguagem e bases de conhecimento. Neste livro, colaborou com o Capítulo 3.

Email: netojoaquim@edu.unisininos.br

Kenzo Sakiyama

Kenzo Sakiyama é formado em Engenharia de Computação pela Universidade Federal do Mato Grosso do Sul (UFMS), com mestrado em Ciência da Computação pelo ICMC/USP. Atualmente possui um doutorado em andamento também no ICMC/USP. Suas áreas de interesse incluem redes neurais profundas, mineração de texto em redes sociais e grafos. Neste livro, colaborou com o Capítulo 14.



Email: kenzosakiyama@usp.br

Larissa Cristina Berti

Larissa Berti é formada em Fonoaudiologia pela UNESP de Marília, com doutorado em Linguística pela Unicamp. É docente do Departamento de Fonoaudiologia da UNESP desde 2009. Sua área de pesquisa é produção e percepção da fala, além do uso de métodos instrumentais na análise de fala. Neste livro, colaborou com o Capítulo 11.

Email: larissa.berti@unesp.br

Larissa A. de Freitas

Larissa Freitas é doutora em Ciência da Computação pela PUCRS (2015). Atualmente é professora da Universidade Federal de Pelotas (UFPeL). Suas áreas de interesse incluem análise de sentimento, ontologias, *chatbots*, detecção de sarcasmo/ironia, detecção de discurso de ódio e detecção de notícias falsas. Neste livro, colaborou com o Capítulo 16.

Email: larissa@inf.ufpel.edu.br

Lilian Mie Mukai Cintho

Lilian Mie Mukai Cintho é formada em Enfermagem e Obstetrícia pela Universidade Federal de São Carlos (UFSCar), com pós-graduação em Tecnologia em Saúde na PUCPR. É docente colaboradora na Universidade Estadual de Ponta Grossa (UEPG). Suas áreas de interesse incluem, gerenciamento de programas de promoção à saúde, enfermagem médico cirúrgica e informática em saúde. Neste livro, colaborou com o Capítulo 9.

Email: miemukai@hotmail.com

Lívia Vicente Dutra

Lívia Vicente Dutra é formada em Letras pela Universidade Federal de Juiz de Fora (UFJF) e mestre em Tecnologia Linguística pela Universidade de Gotemburgo, onde atualmente é doutoranda no programa de Multilinguismo. É também pesquisadora do Laboratório FrameNet Brasil. Seus interesses de pesquisa incluem Semântica de Frames e Processamento de Linguagem Natural (PLN). Neste livro, colaborou com o Capítulo 17.

Email: livia.vdutralet@gmail.com

Lucas Daniel Volpiano Figueiredo

Lucas Daniel Volpiano Figueiredo é graduando em Sistemas de Informação pelo ICMC-USP. Foi estagiário de Inteligência Artificial no IFSC-USP por oito meses, atuando com implementação e pesquisa. Suas áreas de interesse incluem: PLN, aprendizado de máquina e inteligência artificial. Neste livro, colaborou com o Capítulo 18.

Email: volpi@usp.br



Marcelo Finger

Marcelo Finger é professor titular de Ciência da Computação no IME-USP, Pesquisador Principal do Centro de Inteligência Artificial USP-IBM-FAPESP C4AI, com interesses em Geração de Recursos de Processamento de Linguagem Natural (PLN), uso de PLN para a área de Saúde, estudo e interpretabilidade de redes neurais. Neste livro, colaborou com os Capítulos 6, 11 e [O papel dos dados no pré-treinamento de Grandes Modelos de Linguagem](#).

Email: mfinger@ime.usp.br

Marcelo Matheus Gauy

Marcelo Gauy é Bacharel em Matemática pela Universidade de São Paulo (USP), com mestrado na área de Combinatória pela USP e doutorado em Inteligência Artificial pelo ETH Zuerich, na Suíça. É docente na UNESP desde 2025. Suas áreas de interesse incluem aprendizado profundo, computação bioinspirada, processamento de áudio, combinatória e processos estocásticos. Neste livro, colaborou com o Capítulo 11.

Email: marcelo.gauy@unesp.br

Marcelo Gomes de Queiroz

Marcelo Queiroz é doutor em Ciência da Computação pelo IME-USP e docente do mesmo departamento desde 2002. Tem interesse em processamento de fala e voz e aplicações em medicina e linguística. Neste livro, colaborou com o Capítulo 11.

Email: mqz@ime.usp.br

Maria Cristina Ferreira de Oliveira

Cristina Oliveira é formada em Ciência da Computação pelo ICMC/USP (1985), com pós-graduação em Electronic Engineering pela University of Wales, Bangor, no Reino Unido (1990). É docente do ICMC, no Departamento de Ciências da Computação, desde 1986, como Professora Titular desde 2008. Atua em Ciência da Computação, em projetos de pesquisa que atualmente se concentram nos seguintes temas: visualização de informação, mineração visual de dados, análise visual de dados e aplicações. Neste livro, colaborou com o Capítulo 18.

Email: cristina@icmc.usp.br

Maria José Bocorny Finatto

Maria José Bocorny Finatto é linguista especializada em Estudos do Léxico e Terminologia. Investiga a história das terminologias médicas e modos de aplicar técnicas da Linguagem Simples para maior acessibilidade das informações escritas para diferentes perfis de pessoas no Brasil. Análise textual com recursos PLN é seu foco. Orientadora de mestrado e doutorado junto à Universidade Federal do Rio Grande do Sul (UFRGS). Neste livro, colaborou com os Capítulos [Sequência de caracteres e palavras](#), 12 e 15.

Email: mariafinatto@gmail.com



Mariane de Carvalho Pinto

Mariane Carvalho é graduada em Letras pela Universidade Federal de Juiz de Fora (UFJF), onde cursa o mestrado em Linguística Cognitiva e Computacional. Seu trabalho de pesquisa envolve a geração de roteiros de audiodescrição por Inteligência Artificial (IA), com base na Semântica de Frames e na anotação multimodal. É vinculada ao projeto FrameNet Brasil. Neste livro, colaborou com o Capítulo 17.

Email: mariane.carvalho@estudante.ufjf.br

Marlo Souza

Marlo Souza é formado em Ciência da Computação pela Unifacs e em Matemática pela Universidade Federal da Bahia (UFBA), com pós-graduação em PLN e Lógica pela PUC-RS e UFRGS, respectivamente. Atualmente, é professor adjunto na UFBA e co-líder do grupo de pesquisa FORMAS. Tem interesse em pesquisas nas áreas de semântica computacional, extração de informação e modelos de representação de conhecimento em PLN. Neste livro, colaborou com o Capítulo 3.

Email: msouza1@ufba.br

Matheus Cerqueira

Matheus Cerqueira é Técnico em Informática pelo Instituto Federal do Paraná (IFPR) de Jacarezinho-PR e graduando em Ciências de Computação pelo ICMC/USP desde 2020. Atualmente, também é estagiário em Engenharia de Software na Google. Suas áreas de interesse incluem Aprendizado de Máquina e, mais especificamente, Processamento de Linguagem Natural (PLN) e Visão Computacional. Neste livro, colaborou com o Capítulo 13.

Email: matheus.crqra@gmail.com

Maucha Andrade Gamonal

Maucha Gamonal atua como residente de pós-doutorado na Universidade Federal de Juiz de Fora (UFJF). É pesquisadora vinculada ao Projeto FrameNet Brasil desde 2009. Suas áreas de interesse incluem Linguística Cognitiva, Semântica de Frames e Linguística Computacional. É curiosa sobre a produção de sentido na linguagem humana e gosta de plantas, crianças, ar fresco e de seu golden retriever Luke. Neste livro, colaborou com o Capítulo 17.

Email: maucha.andrade@visitante.ufjf.br

Nádia Félix Felipe da Silva

Nádia Félix é formanda em Ciência da Computação pela Universidade Federal de Goiás (UFG), com pós-graduação em Inteligência Computacional no ICMC/USP. É docente na UFG desde 2016. Suas áreas de interesse incluem Aprendizado de Máquina, Processamento de Linguagem Natural (PLN) e Aprendizado Multimodal. Neste livro, colaborou com o Capítulo 13.

Email: nadia.felix@ufg.br



Oswaldo Novais de Oliveira Jr.

Oswaldo Novais de Oliveira Jr. é físico, professor do Instituto de Física da Universidade de São Paulo (USP) de São Carlos, e um dos membros fundadores do Núcleo Interinstitucional de Linguística Computacional (NILC). Suas áreas de interesse são a ciência dos materiais, além de PLN. Neste livro, colaborou com o Capítulo 18.

Email: chu@ifsc.usp.br

Paula Christina Figueira Cardoso

Paula Figueira Cardoso tem graduação em Informática pelo ILES de Santarém, mestrado em Engenharia Elétrica pela Universidade Federal do Pará (UFPA) e doutorado pelo Instituto de Ciências Matemáticas e de Computação (ICMC) – USP/São Carlos. Atualmente é docente na UFPA. Suas áreas de interesse incluem geração de textos, sumarização automática, tecnologias educacionais e pesquisas aplicadas na área de interface homem-máquina. Neste livro, colaborou com os Capítulos Modelos discursivos e 5.

Email: paulastm@gmail.com

Pedro Henrique Paiola

Pedro Henrique Paiola é mestre e doutorando em Ciência da Computação pela UNESP, onde também atua como professor substituto desde 2022. É pesquisador no projeto RADAR, desenvolvido em parceria com a Petrobras. Suas áreas de interesse incluem sumarização automática, modelos de linguagem e avaliação automática de texto. Neste livro, colaborou com os Capítulos 5 e Treinamento de Grandes Modelos de Linguagem na prática.

Email: pedro.paiola@unesp.br

Priscilla de Abreu Lopes

Priscilla de Abreu Lopes é formada em Ciência da Computação pela Universidade Federal de São Carlos (UFSCar), com pós-graduação em inteligência artificial no PPGCC/UFSCar. Suas áreas de interesse incluem Aprendizado de Máquina (AM) e Processamento de Linguagem Natural (PLN). Neste livro, colaborou com o Capítulo 7.

Email: alopes.priscilla@gmail.com

Rafael Oleques Nunes

Rafael Oleques Nunes é mestre e doutorando em Ciência da Computação no Instituto de Informática da UFRGS, bacharel em Ciência da Computação pela mesma instituição e técnico em Informática pelo IFSUL Campus Charqueadas. Pesquisador em PLN, concentra suas investigações em Reconhecimento de Entidades Nomeadas, qualidade de dados e uso de LLMs para criação e validação automática de dados em domínios especializados e de baixos recursos, com aplicações em Direito, Geologia e Saúde. Possui grande interesse nas interseções entre tecnologia e humanidades, explorando o potencial do PLN na análise de textos literários e históricos. Neste livro, colaborou com o Capítulo 13.

Email: rafael.nunes@inf.ufrgs.br



Raphael Montanari

Raphael Montanari é formado em Ciências de Computação pela Universidade de São Paulo (USP), com pós-graduação em Inteligência Artificial (IA) e robótica no ICMC/USP. É servidor na USP desde 2013. Suas áreas de interesse incluem aprendizado de máquina, visão computacional e aplicações de Processamento de Linguagem Natural (PLN) para robótica. Neste livro, colaborou com o Capítulo 14.

Email: rmontanari@usp.br

Renata Vieira

Renata Vieira é PhD em Informática e Ciências Cognitivas pela Universidade de Edimburgo, Escócia. É pesquisadora da área de Processamento de Linguagem Natural (PLN) atuando em diversas áreas de aplicação. É investigadora principal convidada do Centro de História, Culturas e Sociedades da Universidade de Évora (Portugal) e conselheira científica do Instituto de Inteligência Artificial na Saúde. Neste livro, colaborou com os Capítulos [Resolução de correferência](#), [Perguntas e Respostas](#), 3 e 15.

Email: renata.vieira@gmail.com

Renato Moraes Silva

Renato Moraes Silva é graduado em Licenciatura Plena em Informática pela UFMT (2010) e Mestre (2013) e Doutor (2017) em Engenharia Elétrica pela Unicamp. Foi pesquisador de pós-doutorado pela UFSCar/Sorocaba (2017-2021) e professor e pesquisador no Centro Universitário Facens (2021-2024). Atualmente, é docente no ICMC/USP. Sua pesquisa é voltada para o PLN, com ênfase em aplicações como análise de sentimentos, detecção de spam, classificação de mensagens instantâneas e identificação de *fake news*. Além disso, desenvolveu pesquisas que aplicam aprendizado de máquina a problemas de visão computacional e estudos que utilizam análises de dados no contexto do setor financeiro. Neste livro, colaborou com o Capítulo 8.

Email: renatoms@icmc.usp.br

Rodrigo Nogueira

Rodrigo Nogueira é formado em Engenharia Elétrica pela UNICAMP e possui doutorado em Ciência da Computação pela Universidade de Nova York (NYU). Ao longo de sua carreira, Nogueira contribuiu para os campos de Recuperação de Informação e PLN através da criação de modelos como BERTimbau, doc2query, monoT5, e mais recentemente, os modelos Sabiá 1, 2 e 3, que são LLMs especializados no Brasil. Nogueira é o fundador e CEO da Maritaca AI, uma empresa especializada no desenvolvimento de LLMs especializados no Brasil. Neste livro, colaborou com o Capítulo 14.

Email: rfn@unicamp.br

Roney Lira de Sales Santos

Roney Santos é formado em Ciência da Computação pela Universidade Federal do Piauí (UFPI) em 2014, com mestrado em Processamento de Linguagem Natural (PLN) na UFPI



e doutorado em PLN no ICMC/USP. É docente no ICTI/UFBA desde 2023. Suas áreas de interesse incluem inteligência artificial, aprendizado de máquina, PLN e detecção de conteúdo enganoso. Neste livro, colaborou com o Capítulo 8.

Email: roneysantos@ufba.br

Roseli A. F. Romero

Roseli A. F. Romero fez graduação e Mestrado no ICMC-USP e Doutorado na FEEC-UNICAMP. É professora titular do ICMC-USP desde 2013 e trabalha na área de inteligência artificial há 30 anos. Suas áreas de interesse incluem aprendizado de máquina, aprendizado profundo, com aplicação nas áreas de visão computacional, educação, meio ambiente, robótica e direito. Neste livro, colaborou com o Capítulo 14.

Email: rafrance@icmc.usp.br

Roseval Donisete Malaquias Junior

Roseval Donisete Malaquias Junior é bacharel em Ciência da Computação pela Universidade Federal de São Carlos (UFSCar) e mestrando em Ciência da Computação e Matemática Computacional no ICMC/USP. É pesquisador na Maritaca AI desde 2024. Suas áreas de interesse incluem aprendizado de máquina, grandes modelos de linguagem e pré-treinamento contínuo. Neste livro, colaborou com o Capítulo 14.

Email: roseval@usp.br

Sandra Maria Aluísio

Sandra Maria Aluísio é formada em Ciência da Computação pela UFSCar, com Doutorado em IA e PLN, pela USP. Foi docente de 1988 a 2018, no ICMC/USP, onde hoje atua no Programa Professor Sênior. As áreas de interesse incluem linguística de *corpus*, simplificação textual, recursos semânticos, detecção automática da estrutura do discurso científico e ferramentas de suporte à escrita, análise automática de distúrbios de linguagem, e processamento de fala. Neste livro, colaborou com os Capítulos [Recursos para o processamento de fala](#) e 6.

Email: sandra@icmc.usp.br

Sheila Castilho

Sheila Castilho é licenciada em Letras pela UNIOESTE. Possui mestrado em PLN pela Universidade do Algarve (PT) e pela University of Wolverhampton (UK). Completou seu doutorado pela Dublin City University em 2016. Atualmente, ela é Assistant Professor na Dublin City University. Seus interesses incluem tradução automática, avaliação automática e humana, tecnologias da tradução, e PLN para tradução. Neste livro, colaborou com o Capítulo 4.

Email: sheila.castilho@dcu.ie



Sidney Evaldo Leal

Sidney Leal é pesquisador e cientista de dados no Venturus; pós-doc em Processamento de Fala (Tarsila/C4AI/USP); doutor em Ciências da Computação na área de Inteligência Artificial; MBA em Gestão da Tecnologia da Informação; bacharel em Ciência da Computação e técnico em Processamento de Dados. Atua na indústria desde 1995 em projetos de tecnologia de diversos portes para governos, bancos e multinacionais. Áreas de interesse: aprendizagem de máquina, processamento de linguagem natural, reconhecimento e síntese de fala, complexidade/simplificação de textos e internet das coisas. Neste livro, colaborou com o Capítulo 6.

Email: sidleal@gmail.com

Thiago Alexandre Salgueiro Pardo

Thiago Alexandre Salgueiro Pardo é formado em Ciências da Computação, com foco em pesquisa em PLN. Fez graduação e Mestrado na UFSCar e doutorado e pós-doutorado no ICMC/USP. É professor titular no ICMC, onde atua desde 2006. Seus interesses incluem análise de sentimentos, detecção de notícias falsas, sumarização de textos e modelos linguístico-computacionais de representação textual e métodos automáticos de análise. Neste livro, colaborou com o Capítulo 8.

Email: taspardo@icmc.usp.br

Tiago Agostinho de Almeida

Tiago Almeida é professor associado da UFSCar e especialista em ciência de dados, aprendizado de máquina e processamento de linguagem natural. Possui experiência em ensino e pesquisa em Inteligência Computacional, Aprendizado de Máquina e Reconhecimento de Padrões. É autor de um livro premiado de aprendizado de máquina e +100 artigos científicos. Neste livro, colaborou com o Capítulo 1.

Email: talmeida@ufscar.br

Tiago Timponi Torrent

Tiago Torrent é doutor em Linguística pela Universidade Federal do Rio de Janeiro (UFRJ), tendo feito mestrado e graduação em Letras na Universidade Federal de Juiz de Fora (UFJF). É professor do Programa de Pós-Graduação em Linguística e lidera o laboratório FrameNet Brasil da UFJF desde 2010. Seus interesses de pesquisa são modelos de língua aplicados à saúde digital e ao processamento computacional da multimodalidade. Neste livro, colaborou com o Capítulo 17.

Email: tiago.torrent@ufjf.br

Vitor Hugo Bridi Dei Santi

Vitor Hugo Bridi Dei Santi é formado em Física Computacional pela Universidade de São Paulo (USP). Atualmente, é mestrando no Programa de Pós-Graduação em Física Computacional do IFSC-USP. Suas áreas de interesse incluem redes complexas, simulações



computacionais e análise de literatura científica por técnicas de PLN e IA. Neste livro, colaborou com o Capítulo 18.

Email: vitor.santi@usp.br

Viviane Pereira Moreira

Viviane P. Moreira é doutora em computação na Middlesex University (2004), Inglaterra, mestre em ciência da computação na Universidade Federal do Rio Grande do Sul (UFRGS, 1999) e bacharel em análise de sistemas na UCPel (1996). É professora na UFRGS e orientadora do Programa de Pós-Graduação em Computação. Sua pesquisa concentra-se em recuperação de informação, Processamento de Linguagem Natural (PLN) e mineração de dados. Neste livro, colaborou com o Capítulo 2.

Email: viviane@inf.ufrgs.br

Yohan Bonescki Gumiel

Yohan Bonescki Gumiel é formado em Engenharia de Controle e Automação, com uma carreira dedicada à tecnologia e inovação. Ele fez doutorado na área de PLN voltada a textos clínicos, um campo de estudo que demonstra sua paixão por unir tecnologia e saúde. Atualmente, Yohan ocupa o cargo de cientista de dados na A3data e é pós-doutorando na UFMG, onde continua a aprofundar sua pesquisa e conhecimento em áreas críticas. Suas áreas de interesse incluem aprendizado de máquina, PLN e aplicações voltadas à saúde. Neste livro, colaborou com o Capítulo 9.

Email: yohan.gumiel@gmail.com



Uma iniciativa



brasileiraspln.com

Apoio

