Programa de Pós-Graduação em

LINGUÍSTICA

ANOTAÇÃO DE CORPUS: CARACTERIZAÇÃO DE ENTIDADES NOMEADAS EM TWEETS DO MERCADO FINANCEIRO

São Carlos
2025

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE EDUCAÇÃO E CIÊNCIAS HUMANAS
PROGRAMA DE PÓS-GRADUAÇÃO EM LINGUÍSTICA

ANOTAÇÃO DE *CORPUS*: CARACTERIZAÇÃO DE ENTIDADES NOMEADAS EM *TWEETS* DO MERCADO FINANCEIRO

Laís Piai
BOLSISTA SOFTEX/InovaUSP/MCTI

Dissertação apresentada ao Programa de Pós-Graduação em Linguística da Universidade Federal de São Carlos como parte dos requisitos para a obtenção do título de Mestra em Linguística, na linha de Descrição, Análise e Processamento Automático de Línguas Naturais.

Orientadora: Profa. Dra. Ariani Di Felippo

São Carlos– São Paulo– Brasil
Laís Piai, 2025

Piai, Laís

Anotação de corpus: caracterização de Entidades Nomeadas em tweets do mercado financeiro / Laís Piai -- 2025.
122f.

Dissertação (Mestrado) - Universidade Federal de São Carlos, campus São Carlos, São Carlos
Orientador (a): Ariani Di Felippo
Banca Examinadora: Profa. Dra. Maria Cláudia de Freitas (PUC/RJ), Profa. Dra. Paula Christina Figueira Cardoso (UFPA)
Bibliografia

1. Anotação de corpus. 2. Entidade nomeada. 3. Conteúdo gerado por usuário. I. Piai, Laís. II. Título.

Ficha catalográfica desenvolvida pela Secretaria Geral de Informática (SIn)

DADOS FORNECIDOS PELO AUTOR

Bibliotecário responsável: Arildo Martins - CRB/8 7180

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Educação e Ciências Humanas
Programa de Pós-Graduação em Linguística

Folha de Aprovação

Defesa de Dissertação de Mestrado da candidata Laís Piai, realizada em 30/07/2025.

Comissão Julgadora:

Profa. Dra. Ariani Di Felippo (UFSCar)

Profa. Dra. Maria Cláudia de Freitas (PUC/RJ)

Profa. Dra. Paula Christina Figueira Cardoso (UFPA)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa de Pós-Graduação em Linguística.

Learning how to recompose the words (...)
Recalling, retreating,
Returning, retrieving,
A small talk you're missing,
More clever, but older now.

(Rebirth – Angra)

Agradecimentos

Esta conquista não seria possível sem o alicerce fundamental da minha família e sem a Fé que me sustentou nos momentos de dúvida. Agradeço aos meus pais, Orjana e Valdison, por cada sacrifício, por todo o incentivo e, principalmente, por acreditarem em mim antes mesmo que eu acreditasse. À minha irmã, Luísa, pela parceria e pelo apoio que sempre me impulsionaram. Estendo meu carinho e gratidão aos meus padrinhos, Tio Zé e Tia Zi (*in memoriam*), e à minha prima, Nádia, pelo carinho e pela certeza de que eu sempre tive com quem contar. Ter a torcida de vocês é um privilégio, e sei que não chegaria tão longe sem ela. Desejo que este trabalho, de alguma forma, faça jus a tanto amor.

Ao meu companheiro, Matheus, dedico um agradecimento especial por ter sido meu porto seguro e meu principal interlocutor ao longo de todo este processo. Ninguém mais poderia ter compreendido tão bem esta jornada do mestrado, pois a compartilhamos em momentos similares. Agradeço por ter sido meu leitor mais atencioso, por oferecer um olhar crítico e uma segunda perspectiva sempre que a minha se esgotava, e por ser meu especialista para as perguntas mais absurdas sobre o Mercado Financeiro. Sua paciência, seu incentivo e, acima de tudo, nossa sintonia foram essenciais para que eu chegasse até aqui.

Expresso minha gratidão à minha orientadora, Profa. Dra. Ariani Di-Felippo, pela orientação segura, pelo apoio e pela oportunidade de integrar este projeto, que me apresentou às possibilidades do Processamento de Língua Natural. Estendo meus agradecimentos ao Prof. Dr. Norton Trevisan Roman, por seus valiosos direcionamentos ao longo da pesquisa. Agradeço a ambos por me ensinarem, através do exemplo, que a colaboração e o diálogo interdisciplinar são o caminho para se construir soluções.

Agradeço aos meus colegas do Núcleo Interinstitucional de Linguística Computacional (NILC), que me mostraram na prática que a jornada da pesquisa, embora muitas vezes descrita como solitária, não precisa ser. Sou imensamente grata por ter feito parte de um laboratório com tanta tradição, onde aprendi que a Linguística e a Computação são, de fato, boas amigas. Agradeço a cada um que, com bolo, café e boas conversas, provou que o Processamento de Língua Natural fica muito melhor quando bem acompanhado, e que tornaram a caminhada do mestrado infinitamente mais leve e significativa. Obrigada Ana, Júlio, Xiana, Elvis, Ariadne, Gabriel e Maju!

Este trabalho foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei N. 8.248, de 23 de outubro de 1991, no âmbito do PPI-Softex, coordenado pela Softex e publicado como Residência em TIC 13, DOU 01245.010222/2022-44. A autora deste trabalho também agradece ao Centro de Inteligência Artificial (C4AI-USP) e o apoio da Fundação de Apoio à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM Corporation.

Resumo

A anotação de *corpus* desempenha um papel central no Processamento de Línguas Naturais (PLN), servindo tanto como base para a construção e avaliação de sistemas de Aprendizado de Máquina quanto como recurso essencial para a investigação do comportamento linguístico em diferentes domínios. A anotação de Entidades Nomeadas (ENs), em particular, configura-se como uma tarefa especialmente desafiadora em conteúdo gerado por usuários (CGU), como os *tweets*, uma vez que a linguagem informal e os fenômenos de gênero e domínio demandam metodologias adaptadas. Diante desse cenário, este trabalho realizou a anotação de ENs no DANTEStocks, um *corpus* em língua portuguesa composto por 4.048 *tweets* (84.396 *tokens*) sobre o mercado financeiro. Embora esse recurso já contasse com uma primeira versão anotada, essa considerava apenas as 10 categorias genéricas do Segundo HAREM. Diante disso, este trabalho conduziu uma reanotação independente. A metodologia adotada partiu da taxonomia das 10 categorias do Segundo HAREM, utilizadas na anotação anterior, e a expandiu para um conjunto de 47 tipos, com a proposição de quatro novos (certificado, indicador, *ticker* e usuário), de modo a aumentar a granularidade. Essa reavaliação foi fundamentada em decisões linguisticamente motivadas e implementada por um único anotador, por meio de uma abordagem semiautomática. Esse método combinou a aplicação de regras baseadas em pistas estruturais e morfossintáticas com a curadoria humana, o que permitiu não só gerar uma anotação de referência, mas também um novo conjunto de diretrizes. A partir dessa nova anotação, que resultou em 20.092 entidades, correspondentes a 24.825 *tokens*, a caracterização do *corpus* revelou um perfil linguístico dominado por entidades unitárias, isto é, compostas por um único *token*, e pelos tipos *ticker*, moeda e virtual, confirmando a forte influência do domínio. Em suma, as contribuições desta dissertação são: o enriquecimento do *corpus* DANTEStocks com uma anotação de ENs de granularidade fina, um conjunto de diretrizes para anotação de CGU/*tweets* e uma série de discussões sobre os desafios enfrentados e as estratégias adotadas para superá-los.

Palavras-chave: PLN; Conteúdo Gerado por Usuário; Tweet; Entidade Nomeada; Corpus; Mercado Financeiro.

Abstract

Corpus annotation is a cornerstone of Natural Language Processing (NLP), providing the foundation for training and evaluating Machine Learning systems, as well as for investigating linguistic behavior in various domains. A key challenge in this area is the annotation of Named Entities (NEs) in user-generated content (UGC). The informal language and a wide range of platform and domain-specific phenomena found in tweets demand highly adapted annotation methodologies. This dissertation addresses this challenge by conducting an independent re-annotation of the DANTEStocks, a Portuguese corpus of 4,048 financial market tweets (84,396 tokens). While a previous annotation based on the 10 general categories of the Second HAREM existed, our work expands this taxonomy to a more granular set of 47 types. This was achieved by refining the original guidelines based on linguistically motivated decisions and by introducing four new domain-specific types: certificate, indicator, ticker, and user. The annotation was carried out by a single annotator using a semi-automatic approach that combined rule-based methods with manual curation, resulting in a new reference annotation and a comprehensive set of guidelines. The resulting annotation comprises 20,092 entities (24,825 tokens). A detailed characterization of the corpus revealed a linguistic profile dominated by single-token (i.e. those consisting of a single word) entities and by specific stock market types such as ticker, money, and virtual. The main contributions of this work are therefore: the enrichment of the DANTEStocks corpus with a fine-grained NE annotation, a set of guidelines for annotating UGC, and a discussion of the strategies developed to overcome the inherent challenges of this task.

Keywords: NLP; User-Generated Content; Tweet; Named Entity; Corpus; Stock Market.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

ACE	*Automatic Content Extraction*
AM	Aprendizado de Máquina
API	*Application Programming Interface*
B3	Brasil, Bolsa, Balcão
BERT	*Bidirectional Encoder Representations from Transformers*
CGU	Conteúdo Gerado por Usuários
CoNLL	*Conference on Computational Natural Language Learning*
CRF	Conditional Random Fields
EN	Entidades Nomeadas
GPT	*Generative Pre-trained Transformer*
HAREM	Avaliação de Reconhecimento de Entidades Mencionadas
IA	Inteligência Artificial
IAA	*Inter-Annotator Agreement*
IberLef	*Iberian Languages Evaluation Forum*
ICL	*In-Context Learning*
LLM	*Large Language Models*
LSTM	*Long Short-Term Memory*
MUC	*Message Understanding Conference*
NER	*Named Entity Recognition*
PLN	Processamento de Língua Natural
REN	Reconhecimento de Entidades Nomeadas
RNN	*Recurrent Neural Network*
RS	Revisão Sistemática
WNUT	*Workshop on Noisy User-generated Text*

Sumário

Capítulo 1

Introdução

1.1 Contexto

No Processamento de Língua Natural (PLN), *corpus* é uma coleção de textos produzidos naturalmente, coletada com um propósito específico e armazenada eletronicamente (Sinclair, 2005). Quando adicionada alguma informação linguística (explícita) a essa coleção, diz-se que o *corpus* foi "anotado". A depender do segmento de texto sob anotação, as etiquetas podem indicar diferentes análises linguísticas, sejam elas morfológicas, sintáticas, semânticas, pragmáticas ou discursivas (Leech, 2005; Freitas, 2022; Jurafsky; Martin, 2025).

Além disso, a execução da anotação de *corpus* pode ser feita de três maneiras distintas, que variam em função do volume e complexidade do trabalho humano envolvido. Tais maneiras são: (i) manual, feita por humanos, (ii) semiautomática, realizada por máquina com posterior revisão manual ou (iii) automática, totalmente realizada pela máquina. Se a anotação possui algum grau de supervisão humana, como na manual e na semiautomática, os resultados são os chamados *corpora* padrão ouro.

Os *corpora* anotados são a base do Aprendizado de Máquina (AM) supervisionado em PLN. Nesse paradigma, um modelo é treinado com dados rotulados para aprender a mapear entradas e saídas, sendo os *corpora* o recurso utilizado tanto para esse treinamento, quanto para a posterior avaliação do modelo (Russell; Norvig, 2016).

Para que um *corpus* anotado propicie um bom aprendizado automático, dois fatores são fundamentais (Duran; Pardo, 2024). O primeiro deles é a consistência da anotação: as classes/etiquetas devem ser usadas de maneira uniforme, evitando que um mesmo fenômeno, conforme definido pelas diretrizes definidas para a tarefa de anotação, receba categorizações diferentes

ao longo do *corpus*. Aliás, a depender dos fenômenos e do volume de dados, garantir consistência (e, por conseguinte, qualidade) é uma tarefa bastante custosa. Garantir essa consistência, não se limita ao cálculo da concordância entre anotadores (em inglês, *Inter-Annotator Agreement* – IAA), mas depende também da criação de diretrizes metodológicas robustas e de seu refinamento contínuo. O segundo fator diz respeito à capacidade do modelo de aprendizado empregado, que precisa ser capaz de "capturar" a lógica expressa na anotação para reproduzi-la automaticamente.

O recente avanço do aprendizado de máquina não supervisionado, que utiliza dados não rotulados e é impulsionado pela vasta capacidade computacional, possibilitou o surgimento dos *Large Language Models* (LLM), como os da família *Generative Pre-trained Transformer* (GPT). Embora esses modelos, pré-treinados em enormes volumes de texto, tenham gerado um salto de eficácia no PLN, os *corpora* anotados continuam sendo recursos de fundamental importância por uma série de razões.

Uma delas é o refinamento dos próprios LLMs. Diz-se isso porque, embora os LLMs aprendam bem a partir de textos não anotados na fase de pré-treinamento, os modelos podem ser refinados para uma tarefa específica, o que é comumente feito por meio de *corpora* anotados menores e específicos.

Outra razão é a avaliação do desempenho dos LLMs e de outras ferramentas e sistemas de PLN construídos segundo outros paradigmas de Inteligência Artificial (IA), como o simbólico e probabilístico. Nesse caso, uma amostragem do conhecimento humano esperado, por meio de um *corpus* anotado, em determinada tarefa é sempre necessária. O alto custo financeiro e os impactos ambientais envolvidos no pré-treinamento dos LLMs também justificam a relevância desses recursos.

Além dessas, Freitas (2022) destaca outras duas razões. Uma delas é a de que nem toda pesquisa em PLN é voltada para o desenvolvimento de sistemas ou aplicações. Muitas pesquisas focam na produção de descrições linguísticas para diversos fins, sendo essas descrições pautadas em *corpora*. Por meio de uma anotação, aliás, é possível gerar uma caracterização linguística dos dados, pois, verificando a frequência das etiquetas anotadas, pode-se definir o perfil linguístico do gênero e domínio de um *corpus*. A outra razão é, como a própria autora classifica, filosófica.

No que diz respeito ao estudo científico da linguagem, o paradigma dos LLMs muitas vezes prioriza a **eficácia**, em detrimento da **explicabilidade** do processo gerativo. Essa abordagem

resulta em modelos que frequentemente funcionam como uma caixa-preta, embora produzam bons resultados, seu processo interno de tomada de decisão é opaco. Como destaca Freitas (2022), essa opacidade representa um desafio, pois o modelo pode gerar resultados que não refletem necessariamente a distribuição real dos fenômenos linguísticos nos dados. Nesse sentido, os *corpora* anotados tornam-se essenciais. Eles funcionam como uma base empírica verificável, uma âncora simbólica, que permite não apenas avaliar, mas também interpretar e compreender os resultados do modelo em relação a dados linguísticos concretos e analisados por humanos.

Diante disso, vê-se que a relevância dos *corpora* anotados é variada. O mencionado refinamento dos modelos pré-treinados, em particular, é comumente feito quando se trata de tarefas linguísticas que requerem categorização específica, sendo um desses casos o Reconhecimento de Entidades Nomeadas (REN), em inglês, *Named Entity Recognition* (NER).

O REN consiste, na verdade, em duas tarefas automáticas relacionadas: (i) identificação das Entidades Nomeadas (EN) em um texto e (ii) classificação dessas ENs conforme o contexto em categorias predefinidas (Jurafsky; Martin, 2025). De modo geral, ENs são elementos nominais relevantes para uma área de conhecimento ou tarefa (Freitas, 2022), como pessoas, locais e organizações. Entretanto, também podem ser elementos como expressões temporais e expressões numéricas que não são, aliás, da classe nominal.

A seguir, o exemplo em (1) extraído de um anotador de REN (Albuquerque; Costa et al., 2022), apresenta quatro menções a Entidades Nomeadas (ENs). As menções pertencem, respectivamente, às categorias PESSOA, FUNDAMENTO, ORGANIZAÇÃO (ORG) e LOCAL(LOC). Com exceção da categoria FUNDAMENTO, específica do domínio de origem do exemplo, o *corpus* jurídico Ulysses-NER, as outras três categorias são as mais típicas de ENs, bastante difundidas em aplicações de REN para textos de domínio geral.

(1) [$_{\text{PESSOA}}$ **Afonso Florence]** altera a **[**$_{\text{FUNDAMENTO}}$ **Lei nº 6.088, de 16 de julho de 1974]**, que dispõe sobre a criação da **[**$_{\text{ORG}}$ **Companhia de Desenvolvimento do Vale do São Francisco – Codevasf]**, incluindo a **[**$_{\text{LOC}}$ **Bacia do Rio Paraguaçu]** entre suas áreas de atuação, nos termos que especifica e dá outras providências.

Embora o Reconhecimento de Entidades Nomeadas tenha atingido um alto grau de maturidade para textos de linguagem formal (Nadeau; Sekine, 2007),com modelos baseados em

LLMs atingindo patamares de medida-F1[1] superiores a 90% para o inglês (Ushio et al., 2022) e de 78,6% para o português (Souza; Nogueira; Lotufo, 2020), essa eficácia não se replica em domínios especializados, como o financeiro, biomédico ou o jurídico, e em textos ruidosos. Nesses cenários, o desempenho dos sistemas é significativamente inferior, um obstáculo atribuído principalmente ao uso de terminologia específica não vista pelos modelos durante o pré-treinamento.

Com a explosão do Conteúdo Gerado por Usuários (CGU) (Krumm; Davies; Narayanaswami, 2008), a comunidade de PLN passou a reconhecer os desafios impostos por esses novos tipos de texto. Estudos como o de Derczynski, Bontcheva e Roberts (2016) estabelecerem os *tweets*[2] como um dos maiores desafios para a tarefa de REN. Como argumentam Esmaail et al. (2024), pesquisadores ainda se dedicam a investigar soluções para a tarefa de REN em mídias sociais, especialmente no Twitter. No entanto, conforme os autores, a tarefa permanece sem uma solução plenamente satisfatória, devido à natureza ruidosa desses textos. Essa dificuldade é evidenciada pelo estado-da-arte nesse contexto, que alcança uma Medida-F1 de apenas 53% (Nguyen; Vu; Nguyen, 2020).

O desafio da tarefa de REN em *tweets* se deve às características da linguagem desse gênero de CGU, que incluem (i) informalidade ortográfica e de pontuação, (ii) brevidade, podendo incluir truncamentos e fragmentações de diferentes tipos, (iii) elementos típicos da plataforma, como menções, URL, *hashtag* e truncamentos, e (iv) vocabulário especializado e neologismos diversos a depender do domínio (Sanguinetti; Bosco; Cassidy et al., 2023). Tais obstáculos somam-se às dificuldades intrínsecas da anotação manual de ENs, que, como aponta Freitas (2022), residem nas etapas de identificação, segmentação e classificação das entidades.

Mesmo diante desse cenário, o REN é uma etapa importante para a análise semântica das mensagens circulantes no Twitter (Liu; Zhang et al., 2011), que é hoje uma fonte de conteúdo/informação relevante para vários segmentos da sociedade, como o mercado financeiro. Nesse domínio, uma comunidade ativa de usuários debate e compartilha informações sobre o retorno de ações, a volatilidade de ativos, empresas que compõem a bolsa e pronunciamentos (Souza; Fernandes; Fernandes, 2021). A extração e classificação correta das entidades presen-

[1] A medida-F1, descrita na Seção 3.1, é dada pela média harmônica entre precisão e revocação.

[2] Embora a plataforma tenha sido renomeada para "X" e as mensagens para "*posts*" após a aquisição e reestruturação dessa mídia social por Elon Musk em 2022, optou-se por utilizar, neste trabalho, as denominações originais (ou seja, "Twitter" para a plataforma e "*tweets*" para as mensagens) em concordância com a época em que o *corpus* aqui utilizado foi compilado.

tes nesses *posts* são, portanto, relevantes para aplicações especializadas, como a predição dos movimentos do mercado com base nesse conteúdo.

O DANTEStocks, em particular, é um *corpus* de 4.048 *tweets* do mercado financeiro (Silva; Roman; Carvalho, 2020) e se destaca por ser um *tweebank*, ou seja, possui anotação sintática padrão-ouro segundo o modelo Universal Dependencies (UD) (Nivre et al., 2020). Além da camada sintática, o recurso possui anotação manual de emoções (Plutchik; Kellerman, 1986; Silva; Roman; Carvalho, 2020) e de Entidades Nomeadas (ENs) (Zerbinati; Roman; Di-Felippo, 2024). A robustez de sua anotação-UD já subsidiou o desenvolvimento de ferramentas de PLN de alto desempenho, como um *tagger* (Silva; Pardo et al., 2021) e *parsers* (Di-Felippo; Nunes; Barbosa, 2024a; Di-Felippo; Roman et al., 2024), consolidando o DANTEStocks, o primeiro até onde se sabe, como um *corpus* com grande potencial para análises *cross*-dimensionais (Zerbinati; Roman; Di-Felippo, 2024).

Sobre a primeira anotação de ENs no DANTEStocks, esta baseou-se nas 10 categorias genéricas do HAREM (Mota; Santos, 2008): ABSTRAÇÃO, ACONTECIMENTO, COISA, LOCAL, OBRA, ORGANIZAÇÃO, PESSOA, TEMPO, VALOR e OUTRO, o que, apesar de vantajoso para a generalização, resultou em uma grande perda de especificidade para o domínio financeiro. Classificar um *ticker* de ação apenas como COISA, por exemplo, é uma simplificação que limita a utilização do *corpus* para o desenvolvimento de aplicações especializadas mais precisas (Zerbinati; Roman; Di-Felippo, 2024). Diante desse cenário, a justificativa para esse trabalho será detalhada a seguir.

1.2 Justificativa

A justificativa para esse trabalho fundamenta-se na interseção de dois pontos: a necessidade contínua por recursos anotados de alta qualidade e a dificuldade de se aplicar o REN em domínios especializados. Mesmo na era dos LLMs, os *corpora* anotados permanecem como uma base empírica indispensável para o refinamento de modelos, para a avaliação de sistemas e para o próprio estudo científico da linguagem. Contudo, a eficácia da tarefa de REN, já consolidada para textos formais, decai significativamente em Conteúdo Gerado por Usuários, como os *tweets*, devido a fenômenos como a informalidade e o vocabulário específico.

Nesse sentido, a primeira anotação de Entidades Nomeadas do *corpus*, embora pioneira, utilizou 10 categorias genéricas, o que resultou em uma perda de informatividade para o domínio

e o gênero. Diante disso, o presente trabalho se posicionou como uma segunda perspectiva. O processo de reanálise ajudou a clarificar pontos ambíguos e a identificar decisões nas diretrizes de anotação originais de Zerbinati, Roman e Di-Felippo (2024) que careciam de maior precisão linguística, permitindo assim o seu refinamento e o estabelecimento de novas diretrizes e expansão da taxonomia para lidar com os fenômenos CGU e do domínio.

1.3 Objetivos

O objetivo geral deste trabalho foi caracterizar as Entidades Nomeadas do DANTEStocks por meio de uma anotação mais refinada. Para tanto, a taxonomia hierárquica do Segundo HAREM (categorias e tipos) foi adaptada ao gênero e ao domínio com o acréscimo de novos tipos, que se mostraram relevantes para a análise do mercado financeiro. A caracterização resultante baseou-se, especificamente, no levantamento da distribuição estatística das categorias, tipos e das formas de expressão linguística das ENs (unitárias ou multipalavra).

Com isso, a anotação realizada neste trabalho configurou-se como a segunda análise de ENs no DANTEStocks e propôs uma reavaliação da anotação originalmente apresentada por Zerbinati, Roman e Di-Felippo (2024). Essa reavaliação se materializou em um refinamento da taxonomia e das diretrizes de anotação, com todas as decisões fundamentadas em uma análise linguística dos fenômenos encontrados no *corpus*.

1.4 Procedimentos metodológicos

Para a realização do projeto, as seguintes tarefas foram desenvolvidas:

1. **Revisão da literatura fundamental:** essa etapa englobou a leitura e o estudo da bibliografia fundamental e demais referências pertinentes, publicadas no decorrer do projeto, sobre (i) UGC, o gênero *tweet* e suas características linguísticas, (ii) entidades nomeadas, (iii) reconhecimento de ENs (REN) e (iv) anotação de *corpus* e uma revisão sistemática sobre REN para CGU, com foco no português.

2. **Descrição do *corpus* selecionado:** essa tarefa consistiu em apresentar as características estruturais e lexicais do *corpus* selecionado para esta pesquisa, o DANTEStocks, e as suas anotações prévias, com especial atenção à anotação de ENs (Zerbinati; Roman; Di-Felippo, 2024).

3. **Definição de diretrizes de anotação para as ENs:** essa tarefa englobou a definição de: (i) o conjunto de etiquetas para rotular as ENs (isto é, o *tagset*); (ii) os critérios de delimitação e classificação das ENs em função das particularidades estruturais e lexicais dos *tweets*, e (iii) o esquema de anotação.

4. **Execução da anotação:** essa tarefa abrangeu a delimitação e a classificação efetiva das ENs que ocorrem no DANTEStocks com base das diretrizes estabelecidas no manual produzido na tarefa anterior. Essa etapa foi conduzida por um único anotador de modo semiautomático.

5. **Descrição ou caracterização linguística das ENs no *corpus* :** essa tarefa consistiu na descrição das entidades anotadas, por meio do levantamento estatístico da ocorrência de cada categoria e tipo no *corpus* para analisar sua distribuição e relevância no domínio. Adicionalmente, foram investigados os fenômenos estruturais e lexicais que dificultaram a anotação das ENs. Por fim, para validação, o trabalho foi contrastado com a anotação de Zerbinati, Roman e Di-Felippo (2024).

1.5 Estrutura da dissertação

Além da Introdução (Capítulo 1), esta dissertação está organizada em mais seis capítulos. O Capítulo 2 estabelece os conceitos fundamentais sobre CGU, o gênero *tweet*, *tweebank*, Entidade Nomeada e anotação de *corpus*. Em seguida, o Capítulo 3 apresenta a revisão da literatura sobre a tarefa de REN, com foco em uma Revisão Sistemática que mapeia os recursos disponíveis para o CGU em português.

O Capítulo 4 apresenta o *corpus* selecionado, o DANTEStocks, detalhando suas características estruturais, lexicais e as camadas de anotação pré-existentes, com especial atenção à primeira anotação de ENs de Zerbinati, Roman e Di-Felippo (2024). A metodologia da nova anotação é descrita no Capítulo 5. Este capítulo detalha o conjunto de etiquetas expandido, o formato de anotação e as diretrizes de delimitação e classificação desenvolvidas para lidar com os fenômenos do *corpus*.

O Capítulo 6 apresenta e analisa os resultados da anotação refinada, discutindo a distribuição das categorias e tipos de entidades. Finalmente, o Capítulo 7 traz as considerações finais, sintetizando as principais contribuições do trabalho, reconhecendo suas limitações e apontando direções para pesquisas futuras.

Capítulo 2

Conceitos Fundamentais

Este Capítulo apresenta os conceitos que fundamentam o desenvolvimento desta pesquisa. A exposição inicia-se, na Seção 2.1, com a contextualização do Conteúdo Gerado por Usuário (CGU) e do gênero *tweet*. Em seguida, na Seção 2.2, define-se a noção de *tweebank*. A Seção 2.3 foca nos conceitos de Entidade Nomeada e na tarefa de REN. Por fim, a Seção 2.4 detalha as metodologias de anotação de *corpus*.

2.1 Conteúdo Gerado por Usuário

Segundo Krumm, Davies e Narayanaswami (2008), o termo Conteúdo Gerado por Usuário (CGU) recobre todo conteúdo, seja na forma de vídeo, imagem, áudio ou texto, produzido por usuários da *web* em plataformas, como as redes sociais, como Facebook, Twitter, Whatsapp, e serviços diversos, como *chats*, *blogs*, *microblogs*, fóruns de discussão e seção de *reviews* de produtos/serviços. Também estão incluídos nesse cenário diálogos de *call center* (Kaplan, 2020; Davidson et al., 2021) e consultas em mecanismos de busca (Topçu; Durgar El-Kahlout, 2021). Isso evidencia que CGU não constitui um gênero textual específico, mas sim, conforme destacado por Sanguinetti, Bosco, Cassidy et al. (2023), um *continuum* de subgêneros que variam significativamente conforme as convenções e limitações do meio (ou plataforma) utilizado.

Em outras palavras, Sanguinetti, Bosco, Cassidy et al. (2023) quer dizer que a forma que o CGU assume é profundamente influenciada pelo meio em que é produzido (Eisenstein, 2013), refletindo as características linguísticas próprias desse contexto – desde o grau de aderência a uma linguagem mais formal ou padrão até os dispositivos linguísticos empregados para comunicar a mensagem (Sanguinetti; Bosco; Cassidy et al., 2023). Enquanto a linguagem dos

textos extraídos da Wikipédia, por exemplo, tende a ser mais próxima da norma padrão, a linguagem de conversas com *chatbots* (Kurniawan; Louvan, 2018), comentários em plataformas *online* (Derczynski; Nichols et al., 2017; Costa, 2023) ou publicações em redes sociais, como no Twitter, geralmente é mais informal.

A crescente popularidade nas últimas décadas de CGU transformou principalmente as mídias sociais em fontes importantes de dados para muitas áreas. E isso pode ser comprovado pelo fato de que empresas, governos e a comunidade, no geral, requerem cada vez mais conteúdo em tempo real advindo dessas mídias dinâmicas e de larga escala (Derczynski; Bontcheva; Roberts, 2016).

2.1.1 O gênero *tweet*

Sobre o Twitter, um estudo relativamente recente demonstrou que os jornalistas que utilizam essa rede como fonte de informação consideram seu conteúdo tão importante quanto as manchetes de fontes oficiais como a *Associated Press* (McGregor; Molyneux, 2020).

Atualmente, em 2025, a plataforma Twitter/X não divulgou o número total de usuários globais ou específicos do Brasil. Entretanto, estimativas de fontes externas fornecem algumas indicações. Segundo a Lifewire[3], a plataforma possui mais de 600 milhões de usuários em todo o mundo. No Brasil, os dados da Statista[4] indicam que, em 2024, o Brasil contava com aproximadamente 22,13 milhões de usuários. Assim, mesmo depois do bloqueio temporário no Brasil devido a questões legais, essa mídia social mantém uma posição de relativo destaque.

Do ponto de vista linguístico, o conteúdo do Twitter é considerado por alguns, como Freitas e Barth (2015), um gênero textual, enquadrando-se no referido *continuum* de subgêneros de Sanguinetti, Bosco, Cassidy et al. (2023). Nesse sentido, ele tem sido concebido como uma mistura de outros gêneros, como notícia, propaganda e bilhete, os quais foram modificados para atender às necessidades comunicativas da plataforma. Outros autores, como Eisenstein (2013), dizem que o *tweet* não pode nem mesmo ser considerado um gênero textual unificado, sendo formado por uma diversidade de estilos e registros, os quais variam, aliás, em função do domínio/assunto.

De forma geral, os *tweets* são caracterizados pela informalidade e brevidade, promovendo, assim, uma comunicação concisa e direta. Tal brevidade, aliás, advém da limitação de caracte-

[3]https://www.lifewire.com/bluesky-vs-x-8777189?utm_source=chatgpt.com.
[4]https://www.statista.com/search/?q=twitter&p=1

res imposta pela plataforma, a qual, atualmente, é de 280 caracteres. Essa brevidade influencia a linguagem do gênero em questão, pois estruturalmente os *tweets* podem apresentar (i) sequências de sintagmas curtos, (ii) sequências de elementos simplesmente justapostos (isto é, sem uma conexão sintática clara entre eles), (iii) orações ou fragmentos de orações, com ou sem problemas de pontuação.

Do ponto de vista lexical, os *tweets* são marcados por fenômenos como o uso de elementos próprios da plataforma (*hashtags*, menções, URLs), simplificações como acrônimos e inicialismos, marcas de expressividade (a exemplo do alongamento grafêmico), além de empréstimos e erros de digitação (Sanguinetti; Bosco; Cassidy et al., 2023).

O domínio também determina certos fenômenos. No mercado financeiro, por exemplo, há características específicas do domínio que diferenciam esses *tweets* dos de língua geral (Di-Felippo; Postali; Ceregatto; Gazana; Silva et al., 2021). Entre essas características estão o uso de *tickers* e *cashtags*, que representam ativos financeiros negociados, como "PETR4" e "$FIBR3", respectivamente. Além de expressões numéricas especializadas, como índices de valorização e desvalorização de ações, frequentemente escritos com símbolos e valores percentuais ("+2,09%"). Também há a ocorrência de formas reduzidas de expressões temporais, como "1T14" para "primeiro trimestre de 2014", e valores monetários aglutinados, onde o símbolo da moeda aparece junto ao número ("R$10,00"). Alguns desses fenômenos estão ilustrados na Figura 2.1.

Figura 2.1: Exemplos de *tweets* do Mercado Financeiro

#bbas3 *Depois acho meu post mas ainda aguardo 18,75* [***#bbas3*** *I will find my post latter but I still wait for (the stock price to reach) 18.75*]
Acordo da ***ALLL3*** *melou?* [*Is the* ***ALLL3*** *agreement gone?*]
Ai meu bolso.... ***Bbas3*** *caiu pra c****** *hoje* [*Ouch, my wallet...* ***Bbas3*** *fell as (expletive) today*]
Ativo c/ vol Financeiro Superior a sua MM21-16h: ***AEDU3 ALUP11 ARTR3 BBRK3 BBTG11 BHGR3 BISA3 BRAX11 BRKM5 BRPR3 CCRO3 COCE5 CPFE3 CPLE6 ELET3*** [*Stocks with financial volume over their MM21-16h:* ***AEDU3 ALUP11 ARTR3 BBRK3 BBTG11 BHGR3 BISA3 BRAX11 BRKM5 BRPR3 CCRO3 COCE5 CPFE3 CPLE6 ELET3***]

Fonte: Vieira da Silva, Roman e Carvalho (2018).

2.2 *Tweebank*

A popularidade atingida pelo Twitter desde o seu surgimento em 2006 motivou o desenvolvimento de várias aplicações para o processamento do conteúdo veiculado por ele. Tais aplicações visam, por exemplo, (i) medir o interesse por certos tópicos ou assuntos, (ii) detectar eventos imprevisíveis em tempo real, (iii) predizer o comportamento dos ativos nas bolsas, entre outras. Para tanto, elas pautam-se em métodos de análise de sentimentos e mineração de opinião.

Para o desenvolvimento dos referidos métodos e aplicações, construíram-se vários *corpora* anotados. No período de 2011 a 2019, Sanguinetti, Bosco, Cassidy et al. (2023) identificaram mais de 30 *corpora* de CGU com anotação sintática de referência construídos para diversas línguas europeias, inglês americano, árabe, chinês, híndi, etc. (Figura 2.2). Como se pode ver na Figura 2.2, a maioria dos recursos é parcial ou inteiramente composta por conteúdo extraído do Twitter e a anotação sintática tende a seguir o modelo gramatical *Universal Dependencies* (UD) (Nivre et al., 2020), o que é indicado na quinta coluna ("UD-based") da Figura 2.2 por meio do índice "Yes". O modelo UD, aliás, será descrito com mais detalhes na apresentação do *corpus* selecionado para este trabalho no Capítulo 4. Quando um *corpus* de *tweet* possui anotação sintática, dá-se a denominação *tweebank* (Sanguinetti; Bosco; Sarti, 2018).

Além do alcance das opiniões veiculadas na plataforma em diferentes segmentos da sociedade, outras razões para a proeminência dos *corpora* anotados de *tweets* foram a facilidade de obtenção dos dados via *Application Programming Interface* (API) e a política de uso dos dados para fins acadêmicos adotada até muito recentemente pela plataforma. Até 2023, pesquisadores acadêmicos tinham acesso gratuito à API para coletar grandes volumes de dados, sendo possível acessar/compilar todos os *tweets* públicos desde 2006, sem limite de tempo [5].

Para o português, tem-se alguns *corpora* de CGU disponíveis ao PLN, sendo a maioria composta por *tweets*, a saber: (i) *corpus* de 76.358 *tweets* que mencionam os candidatos à eleição presidencial de 2010 (Silva; Gomide et al., 2011), (ii) *Corpus* 7x1, que engloba 2.728 comentários postados no *Twitter* durante as semifinais da Copa do Mundo de 2014 (Moraes; Manssour; Silveira, 2015), (iii) *corpus* que contém 554.623 *tweets* com ocorrência de *emojis* positivos e 425.444 de *emojis* negativos (Junior et al., 2017), (iv) TweetSentBR, que engloba 15.000 *tweets* do domínio "show de TV" (Brum; Nunes, 2018), (v) *corpus* 4P (Silva; Pardo, 2019), que possui

[5] Em 2023, o Twitter revogou o acesso gratuito à API acadêmica, como parte das mudanças impostas pela nova gestão de Elon Musk. Com isso, os pesquisadores passaram a enfrentar altos custos para acessar os dados. O plano gratuito atual permite o acesso a apenas 1.500 *tweets* por mês.

Figura 2.2: Lista de *treebanks* contendo CGU.

Name	References	Source	Language	UD-based
ATDT	Albogamy and Ramsay (2017)	Twitter	AR	Yes
Hi-En-CS	Bhat et al. (2018)	Twitter	HI/EN	Yes
TwitterAAE (TAAE)	Blodgett et al. (2018)	Twitter	AAE, MAE	Yes
TWITTIRÒ-UD (TWRO)	Cignarella et al. (2019)	Twitter	IT	Yes
DWT	Daiber and Van Der Goot (2016)	Twitter	EN	No*
W2.0	Foster et al. (2011)	Twitter, sort fora	EN	No‡
Foreebank (Frb)	Kaljahi et al. (2015)	Technical fora	EN, FR	No‡
Tweebank (Twb)	Kong et al. (2014)	Twitter	EN	No*
Tweebank2 (Twb2)	Liu et al. (2018)	Twitter	EN	Yes
TDT	Luotolahti et al. (2015)	Various	FI	Yes
xUGC	Martínez Alonso et al. (2016)	Various	FR	Yes
Estonian Web Treebank (EtWT)	Martínez Alonso et al. (2016)	Various	ET	Yes
ITU	Pamay et al. (2015)	n.a.	TR	No*
WDC	Read et al. (2012b)	Various	EN	No‡
tweeDe	Rehbein et al. (2019)	Twitter	DE	Yes
PoSTWITA-UD (Pst)	Sanguinetti et al. (2018)	Twitter	IT	Yes
FSMB	Seddah et al. (2012)	Twitter, Facebook, discussions fora	FR	No‡
Narabizi (NBZ)	Seddah et al. (2020)	Newspaper fora	DZ/FR	Yes
EWT	Silveira et al. (2014)	Various	EN	Yes
LAS-DisFo (LDF)	Taulé et al. (2015)	Discussion fora	ES	No‡
MoNoise (MNo)	Van Der Goot and van Noord (2018)	Twitter	EN	Yes
STB	Wang et al. (2017)	Discussion fora	SgE	Yes
CWT	Wang et al. (2014)	Twitter, Sina Weibo	ZH	No*
GUM	Zeldes (2017)	Various	EN	Yes
HSE	n.a.	Various	BE	Yes
OOD	n.a.	Various	FI	Yes
TwittIrish (TwIr)	n.a. (Publication forthcoming)	Twitter	GA	Yes
Cadhan (Cdh)	n.a.	Various	GV	Yes
Taiga	n.a.	Various	RU	Yes
IU	n.a.	Various	UK	Yes

Fonte: Sanguinetti, Bosco, Cassidy et al. (2023).

comentários sobre produtos eletrônicos e resumos comparativos entre eles, e (vi) o *corpus* de 4.517 *tweets* do domínio do mercado financeiro (Silva; Roman; Carvalho, 2020). Nenhum deles, no entanto, é um *tweebank*, uma vez que não possuem anotação sintática.

O interesse particular em automatizar a predição do comportamento dos ativos nas bolsas de valores com base no conteúdo circulante no Twitter sobre as ações (Carosia; Coelho; Silva, 2020; Dhabe et al., 2023) foi motivado principalmente pelo trabalho de Bollen, Mao e Zeng (2011). Nele, os autores encontraram uma forte relação entre as mudanças no estado de ânimo dos usuários do Twitter e as flutuações no Índice *Dow Jones Industrial Average* (DJIA).

No que tange ao domínio do mercado financeiro, parece não haver *corpus* com anotação sintática manual (*gold-standard*) em língua inglesa. Os *corpora* FinTweet (Goh et al., 2021) (1 milhão de *tweets*), FinNER-Twitter (Yang; Lin et al., 2020) (12.000 *tweets*) e STSA (em inglês, *Stock Tweets Sentiment Analysis*) (Bello et al., 2021) (15.000 *tweets*), que comumente subsidiam a investigação de métodos de análise de sentimento nessa língua, possuem anotação sintática automática. Para o português, no entanto, tem-se o *tweebank* de referência DANTEStocks (Di-Felippo; Nunes; Barbosa, 2024a), que será descrito no Capítulo 4.

Além dos *tweebanks*, isto é, dos *corpora* com anotação sintática, recursos com anotação de entidades nomeadas também são essenciais, como mencionado, para subsidiar métodos de REN, os quais desempenham papel importante na interpretação semântica automática dos *tweets*.

2.3 Entidade Nomeada

Na literatura, há várias definições para o termo "Entidade Nomeada" (Silva, 2023), não havendo, portanto, uma definição consensual. As definições, segundo Marrero et al. (2013), pautam-se nas noções de (i) categoria gramatical (nome próprio), (ii) designador rígido, (iii) identificador único ou (iv) domínio ou propósito de aplicação, as quais nem sempre são empregadas de forma excludente.

Muitas dessas definições foram propostas no âmbito de eventos ou desafios de avaliação conjunta (em inglês, *shared task*) . Nesses desafios, diferentes equipes de pesquisa trabalham em uma mesma tarefa, utilizando conjuntos de dados (*corpora*) e métricas de avaliação padronizados, objetivando comparar e aprimorar técnicas e métodos para definir ou superar o estado-da-arte em tarefas de PLN (Santos; Freitas, 2024). Entre eles, destacam-se o MUC Grishman e Sundheim (1996), o CoNLL (Tjong Kim Sang, 2002; Tjong Kim Sang; De Meulder, 2003), ACE (Doddington et al., 2004) e HAREM (Santos; Cardoso, 2007; Mota; Santos, 2008).

O termo Entidade Nomeada foi cunhado na sexta edição do evento *Message Understanding Conference* (MUC), realizada em 1995. O MUC foi um marco para o PLN, em especial para a área de Extração de Informações, tendo como principal objetivo avaliar e aprimorar sistemas automáticos de extração de informações a partir de textos jornalísticos, com foco específico no REN e reconhecimento de relações e eventos (Grishman; Sundheim, 1996)[6].

No MUC, as expressões a serem anotadas foram definidas como "identificadores únicos" de entidades. Essa definição diz respeito particularmente aos nomes próprios das categorias PESSOA (PER), LOCAL (LOC) e ORGANIZAÇÃO (ORG), classificados como ENAMEX (em inglês, *entity named expression*) e amplamente tratados como o foco da tarefa de REN. O MUC também estava interessado no reconhecimento de expressões temporais (como datas e horas) e numéricas (como porcentagens e valores). Tais expressões foram classificadas como TIMEX (do inglês, *temporal expression*) e NUMEX (do inglês, *numerical expression*), respectivamente.

[6] https://catalog.ldc.upenn.edu/LDC2003T13

Adotando uma concepção fixa e unívoca de EN, a categoria de uma entidade no MUC não muda conforme o contexto e as categorias PER, LOC e ORG não podem coincidir. Tal concepção de EN recebeu várias críticas, pois ignora a polissemia e o uso metonímico dos nomes próprios. Por isso, competições posteriores como ACE, CoNLL e HAREM adotaram abordagens mais flexíveis, considerando o contexto para definir a categoria da entidade.

A *Conference on Computational Natural Language Learning* (CoNLL) adotou a mesma definição do MUC, segundo a qual ENs são nomes próprios das categorias PER, LOC e ORG. No entanto, ao incluir a categoria MISC (miscelânea), o CoNLL estendeu o conceito de EN para incluir nomes, próprios e comuns, que expressam conceitos considerados relevantes para a extração de informação e que não se encaixavam nas categorias anteriores. Diz-se isso porque MISC engloba evento (p.ex.: Segunda Guerra Mundial, Olimpíadas de 2024, etc.), nacionalidade (p.ex: brasileiro, catalão, etc.), afiliação religiosa/política (p.ex.: católico, muçulmano, etc.) e obra cultural (p.ex.: Harry Potter, Mona Lisa, etc.).

O *Automatic Content Extraction* (ACE), por sua vez, não se limitou à identificação de nomes próprios (p.ex.: Barack Obama), mas incluiu entidades mencionadas de forma descritiva, seja por pronomes ou expressões definidas (p.ex.: o presidente). Além disso, ao buscar avançar a extração de informações para segurança, inteligência e análise de eventos (como conflitos, ataques terroristas e segurança nacional), o ACE considerou, além de PER, LOC e ORG, outras 4 categorias de EN: GEO-POLITICAL ENTITY (GPE) (entidades geo-políticas), FACILITY (FAC) (instalações/estruturas, como prédios, rodovias, pontes, etc.) e VEHICLE (VEH) e WEAPON (WEA) (Doddington et al., 2004).

O HAREM seguiu, de certa forma, o MUC, pois a noção de EN está fortemente pautada na noção de nome próprio. Tanto é que a identificação de uma EN no HAREM é guiada pela ocorrência de ao menos uma letra maiúscula (ou algarismo). No entanto, o HAREM não restringe a classificação dos nomes próprios às 3 categorias mais tradicionais, mas considera 10 categorias (ABSTRAÇÃO, ACONTECIMENTO, COISA, LOCAL, OBRA, ORGANIZAÇÃO, PESSOA, TEMPO, VALOR e OUTRO) e 43 subtipos. Além disso, o HAREM considera os fenômenos da metonímia e vagueza inerentes às ENs. A metonímia ocorre quando um nome tipicamente usado para designar determinada entidade é empregado no lugar de outro com o qual mantém uma relação. Assim, ao contrário do MUC, que considera "Palácio do Planalto" em "Fontes próximas do Palácio do Planalto" uma EN da categoria LOC, o HAREM a classifica como PER, referindo-se ao "presidente da República". A vagueza ocorre quando há possibilidade de di-

versas interpretações simultaneamente, ambas possíveis; por exemplo, "Immanuel Kant" em "um filósofo seguidor de Immanuel Kant" pode ser classificado como PESSOA ou ABSTRAÇÃO, sendo ambas as interpretações igualmente aceitáveis, uma vez que provavelmente elas ocorrem simultaneamente em nosso sistema conceitual e estrutura discursiva. Assim, o que se percebe é que a definição de EN no HAREM é mais contextual que no MUC.

Alguns autores, como Nadeau e Sekine (2007), citam a noção de "designador rígido" para definir EN. Segundo Kripke (1980), um designador d de um objeto x é rígido se ele designa x em todos os mundos possíveis onde x existe, e nunca designa um objeto diferente de x em qualquer um desses mundos possíveis. Esse é o caso de "Richard Nixon" que, segundo o autor, é um designador rígido porque se refere a uma mesma pessoa em qualquer mundo possível. Em oposição, presidente dos "Estados Unidos" não é um designador rígido, pois o referente muda de tempos em tempos. A noção em questão advém da Filosofia da Linguagem e tende a ser bastante controversa, pois mesmo "Richard Nixon" pode se referir a muitas pessoas com esse nome (e não apenas ao ex-presidente americano). Além disso, Kripke diz que termos como "ouro", "água" e "quente" também são designadores rígidos, o que também torna a definição de designação rígida pouco clara.

Outros autores trabalham com a definição de EN em função do domínio do *corpus* ou propósito de aplicação do REN para o qual o *corpus* está sendo construído. Esse é o caso, por exemplo, do *corpus* GENIA (Kim et al., 2003; Thompson; Ananiadou; Tsujii, 2017) da área biomédica, cujo conjunto de categorias de domínio subsidiou uma anotação de EN que foi continuamente enriquecida, fazendo do GENIA um *corpus* de referência no treinamento e avaliação de diversos sistemas de extração de informações em textos científicos do domínio biomédico. Na mesma linha de trabalho, Freitas, Souza et al. (2023), ao anotar um *corpus* composto por boletins e relatórios técnicos do domínio do gás e petróleo, definiram categorias de ENs específicas com o auxílio de especialistas e recursos prévios de domínios relacionados. Entre as categorias do domínio do gás e petróleo, estão, por exemplo, BACIA e UNIDADE LITOESTRATIGRÁFICA[7]. Como destacado por Marrero et al. (2013), essa é a definição de EN mais comumente expressa pela literatura sobre o tema.

Em suma, a concepção de Entidade Nomeada em aplicações práticas de Processamento de Língua Natural evoluiu para além da noção estrita de nome próprio. A concepção mais funcional do termo abrange não apenas entidades *per se*, mas também elementos cruciais para a aná-

[7]Agrupamentos de rochas que compartilham características físicas e composicionais e são identificadas e mapeadas com base nessas características.

lise de um domínio, como datas, expressões numéricas e conceitos nominais específicos. Essa abordagem mais ampla é uma consequência direta da própria natureza da anotação. Como o termo Entidade Nomeada está fundamentalmente relacionado às conexões que se estabelecem entre a língua e o mundo extralinguístico, é necessário empregar estruturas intermediárias, como as categorias de ENs, para viabilizar análises e generalizações. Portanto, o modelo semântico que define o que é uma entidade em um *corpus* é delineado muito mais em função do propósito da aplicação e das especificidades do domínio do que por uma definição gramatical rígida. Nesse sentido, o presente trabalho alinha-se a essa concepção funcional de entidade.

2.3.1 Reconhecimento de Entidades Nomeadas (REN)

A tarefa de REN consiste em rotular automaticamente as expressões linguísticas denominadas Entidades Nomeadas (discutidas na Seção 2.3) com uma ou mais categorias predefinidas (Jurafsky; Martin, 2025). Ela é feita em duas etapas: (i) delimitação da entidade, identificando seu início e fim, e (ii) classificação da entidade conforme o contexto em uma ou mais categorias estabelecidas previamente.

Na história do REN, o tratamento dado à tarefa está intimamente ligado à definição do termo Entidade Nomeada estabelecida principalmente nos desafios de avaliação conjunta. Esses desafios são importantes no PLN porque definem a tarefa, disponibilizam *corpora* anotados (comumente, de padrão-ouro) para o treinamento dos modelos, assim como esquemas de anotação de *corpus*.

O MUC teve um papel de destaque no PLN porque definiu e padronizou tarefas específicas de extração de informação. A partir do MUC, a tarefa de REN ficou focada nas ENAMEX, isto é, nomes próprios que designam entidades das categorias PER, LOC e ORG, assim como expressões de tempo (TIMEX) e valor (ou numéricas) (NUMEX) (Grishman; Sundheim, 1996). No MUC-6 (1995), o foco da tarefa de REN estava em textos jornalísticos, uma vez que o *corpus* disponibilizado aos competidores era composto por 318 notícias (em inglês) extraídas do *Wall Street Journal*, contendo anotação manual de EN (padrão-ouro).

O CoNLL, por sua vez, destaca-se ao fomentar a investigação de sistemas de REN independentes da língua, usando textos em flamengo, espanhol, inglês e alemão, e ao considerar, além de PER, LOC e ORG, a categoria MISC (miscelânea ou diversos) (Tjong Kim Sang, 2002; Tjong Kim Sang; De Meulder, 2003).

O ACE, diferentemente do MUC e CoNLL, propôs a tarefa EDT (do inglês, *Entity Detection and Tracking*), que objetivou o reconhecimento de entidades expressas não somente por nomes próprios, mas também por nomes comuns, pronomes e/ou sintagmas nominais. Nesse sentido, o ACE parece que misturou a tarefa de REN com a de identificação de correferência. Além disso, o ACE considerou, além das tradicionais PER, LOC e ORG, as categorias GPE, FAC, VEH e WEA, o que ampliou consideravelmente a complexidade da tarefa de REN (Doddington et al., 2004). Nesse conjunto, destaca-se a supercategoria GPE, que é responsável por diferenciar locais (LOC) e organizações (ORG) comuns de outras entidades dessas categorias que envolvem aspectos políticos e administrativos. O ACE não se limitou ao gênero jornalístico. Embora incluíssem notícias jornalísticas, os *corpora* disponibilizados pelo ACE também continham transcrições de áudio (p.ex.: conversas telefônicas e entrevistas) e conteúdo extraído de fontes diversas, como fóruns e comunicações informais. Além disso, o ACE abrangeu três línguas: inglês, árabe e chinês.

Para o português, destaca-se a Avaliação de Reconhecimento de Entidades Mencionadas (HAREM), que foi a primeira dedicada à referida língua (Santos; Freitas, 2024). Organizado pelo centro de recursos distribuídos para o processamento computacional o português, Linguateca[8], essa avaliação teve duas edições: Primeiro HAREM (Santos; Cardoso, 2007) e o Segundo HAREM (Mota; Santos, 2008).

Além de ter sido pioneiro, o HAREM ocupa lugar destaque no PLN porque definiu as bases para a tarefa de REN em português, sendo a maioria delas empregada ainda hoje. Resumidamente, a tarefa de REN no HAREM pauta-se em: (i) identificação de toda EN que contém ao menos uma letra maiúscula e/ou algarismo; (ii) conjunto de 10 categorias genéricas (PESSOA, LOCAL, ORGANIZAÇÃO, DATA, VALOR, ABSTRAÇÃO, OBRA, EVENTO, COISA e OUTRO) e 41 subtipos; (iii) classificação das ENs em função do contexto, considerando, por exemplo, a metonímia para definir categoria/tipo; (iv) resolução dos casos de ambiguidade, isto é, anotação precisa ser o resultado da escolha do anotador ou sistema entre as várias opções mutuamente exclusivas; (v) anotação dos casos de vagueza, isto é, a anotação de EN pode conter mais de uma categoria (ou tipo) devido à possibilidade de diversas interpretações simultâneas. Além disso, o HAREM considera diferentes gêneros (escritos) de linguagem formal (jornalístico, literário, enciclopédico, técnico-científico e legislativo), assim como transcrições de fala (entrevistas, debates, conferências, testemunhos e julgamentos).

[8]Mais informações em: https://www.linguateca.pt/.

Mais recentemente, realizou-se o *Iberian Languages Evaluation Forum* (IberLef) (Collovini et al., 2019), que objetivou definir novos desafios de pesquisa e estabelecer novos resultados do estado-da-arte para várias tarefas de PLN envolvendo pelo menos uma língua ibérica (espanhol, português, catalão, basco ou galego). Quanto ao REN, a tarefa focou especificamente no português, sendo definida como a identificação de nomes próprios em textos de diferentes gêneros (e domínios) e a classificação dos mesmos em uma ou mais categorias específicas ou na genérica MISC. Para tanto, a avaliação forneceu 3 *corpora* anotados para o treinamento dos modelos a competir: (i) *corpus* geral, contendo notícias jornalísticas, memorandos, e-mails, entrevistas e artigos de revistas, com anotação das seguintes categorias PER, LOC, ORG, VAL e TME, (ii) *corpus* clínico, contendo notas clínicas[9] produzidas por trabalhadores hospitalares a respeito dos pacientes, cujas ENs da categoria PER foram anotadas manualmente, e (iii) *corpus* da polícia, contendo testemunhos, declarações e interrogatórios advindos da Polícia Federal do Brasil, também anotados manualmente com a categoria PER.

No que concerne à tarefa de REN para CGU, destacam-se as edições do *Workshop on Noisy User-generated Text* (WNUT), cujo objetivo era o de avançar as pesquisas sobre o processamento de textos ditos ruidosos (ou seja, de linguagem informal), sobretudo advindos de redes sociais. O WNUT-16, em particular, focou especialmente em *tweets* e no reconhecimento de nomes próprios que designam ENs das categorias PERSON, GEO-LOC, PRODUCT, COMPANY, FACILITY, MOVIE, MUSIC, ARTIST, SPORTS TEAM e TV SHOW (além de OTHER) (Strauss et al., 2016). Essa edição do evento foi a responsável por disponibilizar um importante *corpus* de 3.856 *tweets* em inglês com anotação manual de ENs, o qual já fora mencionado.

2.4 Anotação de *corpus*: metodologias

Como mencionado, *corpus* é uma coleção de textos produzidos naturalmente que foi coletada com um propósito específico e armazenada eletronicamente. Diz-se que um *corpus* foi anotado quando ele recebe alguma camada de informação linguística explícita. Precisamente, anotação significa a delimitação de um tipo de segmento textual e a atribuição a ele de uma etiqueta (*tag*) previamente definida que explicita uma análise humana sobre o segmento. Isso quer dizer que toda anotação é uma atividade interpretativa e, por isso, um *corpus* anotado não é necessari-

[9]Esses textos impõem desafios ao REN porque são repletos de termos técnicos, abreviaturas e ortografia relativamente informal.

amente uma fonte objetiva de informações linguísticas, mas o resultado da interpretação dos anotadores (Freitas, 2022).

Os referidos segmentos podem ser palavras, sintagmas, orações, sentenças ou mesmo parágrafos. A depender do segmento sob anotação, as etiquetas podem indicar diferentes análises linguísticas, sendo que as mais comuns são: (i) **morfossintática**, que diz respeito às classes das palavras ou *Part-of-Speech* (PoS) *tags*, (ii) **sintática**, que comumente se refere à relações sintagmáticas ou de dependências, (iii) **semântica**, que pode ser relativa aos papéis semânticos dos argumentos projetados por palavras predicadoras, entidades nomeadas, polaridade, emoção, etc., (iv) **pragmática**, que tende a explicitar os atos de fala, e (v) **discursiva**, que comumente diz respeito a relações discursivas estabelecidas entre orações, sentenças, parágrafos ou outras unidades de análise (Jurafsky; Martin, 2025).

A anotação pode ser conduzida de forma totalmente **manual** por humanos ou por máquina com posterior revisão manual (isto é, de forma **semiautomática**). Além dessas, a anotação também pode ser totalmente **automática**. Em função da execução, a anotação pode ser classificada como (i) padrão-ouro ou dourada (*gold standard*), quando ela é conduzida de forma manual ou revisada manualmente, ou (ii) prateada (*silver standard*), quando a anotação é feita automaticamente, sem auxílio ou revisão humana. A seguir, são detalhados os processos de anotação.

2.4.1 Anotação manual

A anotação manual é comumente aplicada no desenvolvimento de esquemas de anotação, o que envolve definir etiquetas e diretrizes de anotação, ou quando não há anotação automática confiável. Embora envolva um trabalho moroso, o método manual é, como aponta Stefanowitsch (2020), a única possibilidade a depender do fenômeno linguístico de interesse, uma vez que a automática ou não é possível ou resulta em uma qualidade tão baixa que simplesmente torna a revisão manual posterior inviável. Para ilustrar, cita-se o caso das metáforas, que são quase impossíveis de serem identificadas automaticamente, pois têm poucas ou nenhuma propriedade que as distingue sistematicamente da linguagem literal. Esse tipo de anotação, embora demorada, pode levantar questões que talvez não fossem identificadas em uma etapa de revisão, uma vez que os anotadores tendem a acatar a análise fornecida pela máquina (Freitas, 2022). Ademais, ressalta-se que, quando manual, a tarefa de anotação manual é comumente feita com o auxílio de ferramentas ou editores (isto é, *softwares* dedicados à tarefa em questão).

2.4.2 Anotação semiautomática

A anotação semiautomática é a mais frequente. Ela se caracteriza por englobar uma primeira etapa de anotação automática seguida por uma de revisão humana. Em outras palavras, esse tipo de anotação é a correção manual da saída produzida por uma ferramenta computacional. Nesse caso, o tempo gasto no referido processo é consideravelmente menor, mas ainda assim trata-se de uma tarefa custosa, demandando tempo e esforço por parte da equipe responsável pela execução da tarefa.

Uma das principais estratégias de anotação semiautomática se baseia no uso de regras ou padrões linguísticos. Nesse método, o processo frequentemente ocorre em etapas, como demonstrado na anotação do *corpus* PetroNer (Freitas; Souza et al., 2023). Primeiramente, pode-se aplicar uma anotação inicial de alta cobertura, geralmente a partir de um léxico ou *gazetteer* de domínio, que identifica um grande número de entidades potenciais, mas que pode incluir falsos positivos. Em seguida, um conjunto de regras linguísticas, frequentemente implementadas por meio de expressões regulares (regex) baseadas em padrões morfossintáticos, é desenvolvido e aplicado para refinar o resultado. Essas regras desempenham um papel duplo: por um lado, corrigem a anotação inicial, eliminando os erros, e por outro, identificam novas entidades não previstas no léxico.

O Algoritmo 1, implementado para a correção e anotação no *corpus* PetroNER [10], em formato CoNLL-U, ilustra a aplicação prática de uma regra linguística.

Algorithm 1 Exemplo de Regra de correção e anotação do PetroNER comentada

1: **Entrada:** token com atributos `lemma` e `deps`
2: **if** regex(“Neoco-miano”, token.lemma):
3: Seja um token de lemma “Neoco-miano”
4: **Corrigir** o lemma para “Neocomiano”
5: token.lemma = “Neocomiano”
6: **Atribuir** etiqueta de Entidade “B=UNIDADE_CRONO” à coluna DEPS no arquivo ConNLL-U
7: token.deps = “B=UNIDADE_CRONO”
8: **Saída:** token anotado (se condição satisfeita)

Conforme apontado por Freitas (2024), é fundamental notar que a eficácia da anotação baseada em regras reside na sua capacidade de capturar regularidades, uma vez que estas dependem da repetição de padrões no *corpus*. Embora essa abordagem seja excelente para tratar fenômenos recorrentes, ela encontra um limite na própria natureza da linguagem, descrita pela

[10] https://github.com/alvelvis/Regras-PetroNer/tree/main

Lei de Zipf (Freitas, 2024). Essa lei postula que a frequência de uma palavra é inversamente proporcional à sua posição no *ranking* de frequência, resultando em uma distribuição onde poucas palavras são extremamente comuns e uma cauda longa é composta por eventos raros ou únicos. É para essa cauda longa de fenômenos que a criação de regras se torna impraticável.

Apesar dessa limitação, regras linguísticas continuam sendo uma das estratégias mais eficazes e controláveis para a construção de um *corpus* padrão-ouro, superando em muito a viabilidade de um processo inteiramente manual (Freitas, 2024).

Uma outra opção atualmente no contexto da anotação semiautomática é realizar o primeiro passo automático por meio de um LLM, como o GPT-4o, disponível gratuitamente na *web*, explorando técnicas da *Engenharia de Prompt* (Liu; Deng et al., 2023).

Em linhas gerais, o termo *Engenharia de Prompt* indica o processo de criar e ajustar *prompts* (comandos ou instruções) fornecidos a um LLM para obter respostas desejadas ou específicas (Jurafsky; Martin, 2025; Larguesa, 2024). Isso envolve a elaboração de perguntas ou instruções de maneira precisa e clara para orientar o modelo a gerar uma resposta útil e relevante, que, no caso, seria a anotação de um *corpus*.

As técnicas de *Engenharia de Prompt* são várias, como *zero-shot*, *few-shot prompting*, *chain-of-thought prompting*, *zero-shot chain-of-thought prompting*, *least-to-most prompting* e outras. A *zero-shot*, por exemplo, refere-se à capacidade de um LLM de entender e executar uma tarefa sem ter recebido exemplos específicos dessa tarefa anteriormente. A técnica *few-shot prompting* envolve fornecer ao LLM um pequeno número de exemplos para ajudá-lo a entender o contexto e realizar uma tarefa específica. Já *chain-of-thought prompting* consiste em demonstrar ao modelo, no próprio *prompt*, a forma exata de raciocinar, dando uma resposta completa como exemplo, incluindo o passo a passo para chegar a ela. A *zero-shot chain-of-thought prompting*, por sua vez, considera que os LLMs não precisam necessariamente de toda essa explicação e exemplos para dar uma resposta correta, mas apenas da frase *"Let's think step by step."* (em português, "Vamos pensar passo a passo.") no final do *prompt*.

2.4.3 Anotação automática

Por fim, a anotação automática é aquela realizada integralmente por um sistema computacional, sem qualquer intervenção ou revisão humana. No entanto, como salienta Freitas (2022), a avaliação da qualidade desse tipo de anotação é feita pela comparação com o desempenho humano, evidenciando a necessidade de anotação humana.

Essa abordagem é frequentemente implementada por meio de *toolkits* de PLN, que funcionam como caixas de ferramentas computacionais com modelos pré-treinados para diversas tarefas, como REN. Um exemplo clássico é o Stanford NER (Finkel; Grenager; Manning, 2005a) [11], que utiliza modelos estatísticos como os Conditional Random Fields (CRF) (Lafferty; McCallum; Pereira, 2001), para identificar e classificar entidades em textos brutos. Outros exemplos são as bibliotecas spaCy [12] e o NLTK [13] , *toolkits* amplamente empregados para gerar uma primeira camada de anotação em novos *corpora* e que oferecem suporte em língua portuguesa.

2.4.4 Avaliação da anotação humana

Por fim, é fundamental ressaltar que a anotação não é um processo neutro, mas sim um ato inerentemente interpretativo. Essa interpretação ocorre tanto no plano das decisões de projeto, em que as categorias e os esquemas refletem escolhas teóricas e metodológicas, quanto no plano humano. Neste último, os anotadores aplicam as diretrizes a partir de seu conhecimento, de seu entendimento contextual e, de forma mais profunda, de suas próprias visões de mundo e posicionamentos, o que inevitavelmente introduz variação entre eles.

Considerando essa natureza interpretativa, a avaliação de uma anotação pode seguir dois caminhos principais. O mais difundido é o cálculo da concordância entre anotadores (IAA), que mede a consistência com que as diretrizes foram aplicadas por diferentes pessoas.

Alternativamente, a anotação pode ser avaliada por comparação com um gabarito de referência, embora essa abordagem seja mais adequada para medir consistência. No caso de um único anotador, a concordância entre anotadores não se aplica, mas podem ser empregadas outras estratégias. A partir de Freitas (2024), uma opção é a avaliação intrínseca, em que o *corpus* é usado para treinar e testar um modelo de aprendizado de máquina, e o desempenho do modelo indica possíveis inconsistências ou dificuldades de generalização. Outra opção é a avaliação extrínseca, que verifica a adequação do *corpus* ao medir se ele melhora o desempenho em uma tarefa mais complexa. Além dessas avaliações, podem ser empregadas estratégias de revisão, tanto lineares (frase a frase) quanto transversais (analisando um mesmo fenômeno em todo o conjunto de dados). Outra possibilidade é aplicar métricas de comparação entre anotações de tarefas similares, mesmo que não utilizem exatamente o mesmo conjunto de diretrizes, estratégia adotada neste trabalho e descrita no Capítulo 6.

[11] https://stanfordnlp.github.io/stanza/index.html
[12] https://spacy.io
[13] https://www.nltk.org/

Capítulo 3

Revisão da literatura

Este Capítulo apresenta a Revisão da Literatura sobre a tarefa de REN e os recursos linguísticos disponíveis. A exposição inicia-se com um panorama das abordagens de REN (Seção 3.1), seguido pela apresentação do estado-da-arte para textos formais e para CGU na Seção 3.2. Na sequência, as Seções 3.4 e 3.3 detalham os recursos linguísticos e os formatos de anotação já consolidados para textos formais. Por fim, a Seção 3.5 apresenta uma Revisão Sistemática da Literatura desenvolvida para mapear os *corpora* disponíveis especificamente para a tarefa de REN em CGU no português.

3.1 Abordagens de REN

Embora as abordagens e métodos de REN não sejam de fato o tópico deste trabalho, mesmo quando se trata de CGU, fornece-se aqui um panorama da tarefa, pois o desenvolvimento de sistemas de REN é uma das principais motivações para a construção de *corpora* anotados.

A tarefa REN para a língua portuguesa foi recentemente mapeada por duas Revisão Sistemática (RS) principais, que se concentraram em textos de linguagem formal (Albuquerque; Souza et al., 2023; Silva, 2023). Embora um desses estudos (Silva, 2023) tangencie a análise de *tweets* ao citar Peres, Esteves e Maheshwari (2017), o consenso é que o reconhecimento de ENs para CGU em português permanece um campo com pouca tradição. Essa lacuna motiva a revisão sistemática específica que será detalhada na Seção 3.5.

Com base nessas revisões em textos formais, as abordagens de REN podem ser classificadas segundo os paradigmas de PLN/IA em: (i) abordagens simbólicas (ou baseadas em regras), (ii)

abordagens estatísticas tradicionais, (iii) abordagens baseadas em redes neurais profundas, e (iv) abordagens híbridas.

Independentemente da abordagem adotada, a construção de modelos de PLN é indissociável de sua avaliação, pois é preciso compreender quando e por que eles funcionam para validá-los e aperfeiçoá-los (Madureira, 2024). Essa necessidade é o que fundamenta a comparação de um sistema com o **estado-da-arte**, que representa o mais alto nível de desenvolvimento e conhecimento alcançado sobre um determinado tópico até o momento. Para que essa comparação com o estado-da-arte seja realizada de forma sistemática, são definidas métricas de avaliação.

Para tarefas de classificação, como o REN, a principal métrica é a **Medida-F1**, em inglês, *F1-Score* (Madureira, 2024). Essa métrica é calculada para condensar, em um único valor, o desempenho do sistema em duas outras medidas fundamentais: a **Precisão** P e a **Revocação** ou *Recall* R (Claro et al., 2024). A precisão avalia, dentre todas as classificações que o sistema realizou, quantas estavam de fato corretas. Já a revocação mede a abrangência do sistema, ou seja, de todas as identificações de Entidade que deveriam ser feitas em um *corpus* de teste, quantas ele foi capaz de identificar. A Medida-F1, portanto, corresponde à média harmônica entre essas duas, conforme é definida pela Equação 3.1.

$$F1 = \frac{2 * P * R}{P + R} \tag{3.1}$$

No que tange à tarefa de REN, as principais abordagens serão detalhados a seguir.

3.1.1 Abordagens simbólicas

Essas abordagens, também conhecidas como baseadas em conhecimento ou regras, utilizam informações linguísticas explícitas para realizar duas subtarefas: (i) identificar e (ii) classificar as entidades no texto. Elas envolvem a criação manual de regras e padrões que descrevem como identificar entidades nomeadas com base em características linguísticas, como informações morfossintáticas, frequência lexical, similaridade com palavras em um dicionário de referência, entre outros (Jurafsky; Martin, 2025). Essas abordagens foram muito exploradas nos primeiros sistemas para REN, como o PALAVRAS-NER Bick (2006), o SIEMÊS (Sarmento, 2006), PAMPO (Rocha et al., 2016), o Rembrandt (Cardoso, 2008) e outros. Como as regras são criadas manualmente, é possível ter controle rigoroso sobre o comportamento do sistema

construído segundo essa abordagem. Por outro lado, criar e manter um grande conjunto de regras é dispendioso e difícil de escalar para diferentes domínios. Além disso, abordagem simbólica pode ter dificuldades em lidar com entidades novas, pois funciona com base em regras fixas.

3.1.2 Abordagens estatísticas

Sobre as abordagens estatísticas, destaca-se que se trata de sistemas baseados em AM estatístico, os quais definem REN como uma tarefa de classificação multiclasse. Existem vários tipos de AM, mas o mais adotado em REN é o supervisionado. No AM supervisionado, um algoritmo aprende (ou é treinado) a partir de um conjunto de dados rotulados ou anotados. Isso significa que, para cada entrada do conjunto de treinamento, há uma saída esperada (ou rótulo/etiqueta), permitindo que ele aprenda uma função que mapeia entradas para saídas de forma otimizada, gerando um modelo de classificação. Quando exposto a dados novos, o modelo resultante do AM é capaz de prever os rótulos adequados (Russell; Norvig, 2016).

Uma característica importante desses modelos é que a identificação e a classificação das ENs geralmente são tratadas juntas, sem a necessidade de módulos separados para cada tarefa. Seguindo essa abordagem, têm-se os trabalhos de Solorio (2007), Amaral e Vieira (2014), Júnior et al. (2015) e Lopes, Teixeira e Oliveira (2019), entre outros.

3.1.3 Abordagens neurais

As abordagens baseadas em redes neurais profundas, em inglês *deep neural networks*, revolucionaram o PLN ao empregar redes neurais com múltiplas camadas, onde cada uma transforma e refina a informação antes de passar para a próxima. As primeiras arquiteturas a ganhar destaque na tarefa de REN foram as *Recurrent Neural Network* (RNN), ou Redes Neurais Recorrentes, especialmente suas variantes como o *Long Short-Term Memory* (LSTM).

Um dos primeiros trabalhos a aplicar redes neurais profundas para REN em português foi o de Santos e Guimarães (2015). Nele, os autores utilizam a CharWNN, que combinava vetores de representação de palavras (*word embeddings*) e *character embeddings*. Segundo Silva (2023), outros estudos exploraram diversas configurações com redes BiLSTM, frequentemente combinadas com uma camada final de CRF. Aliás, com a popularização das redes neurais profundas, tem-se visto um crescimento considerável da exploração de métodos de REN também para domínios específicos, como o jurídico e o biomédico.

Um avanço fundamental nesse paradigma veio com a introdução da arquitetura *Transformer* (Vaswani et al., 2017). Ao utilizar um mecanismo de *self-attention*, os *Transformers* superaram as limitações sequenciais das RNNs, permitindo o processamento de texto de forma mais paralela e contextualizada. Essa inovação não apenas melhorou o desempenho em diversas tarefas, mas também viabilizou o treinamento de modelos em uma escala sem precedentes.

É a partir dessa arquitetura que surge a atual geração de *Large Language Models* (LLMs). Em português, esse termo tem sido traduzido para "Grandes Modelos de Linguagem" ou "Modelos de Linguagem de Larga/Grande Escala". Entre eles, estão os da família GPT, da OpenAI, e da família dos *Bidirectional Encoder Representations from Transformers* (BERT) (Devlin et al., 2019). Diferentemente dos modelos anteriores, que eram majoritariamente treinados de forma supervisionada em tarefas específicas, os LLMs são pré-treinados em enormes volumes de dados não rotulados por meio de aprendizado não supervisionado. Em linhas gerais, eles aprendem padrões linguísticos complexos de vastos repositórios de texto, como a *web*.

Para a tarefa de REN, esses modelos pré-treinados são então adaptados por meio de técnicas de refinamento. As duas principais são o *fine-tuning*, que retreina o modelo, ajustando seus pesos, com um *corpus* anotado menor e específico, e o *In-Context Learning* (ICL), que guia o modelo com exemplos de anotação fornecidos através da Engenharia de *Prompt*, descrita na Seção 2.4.2.

Um exemplo prático do uso dessas técnicas para o Reconhecimento de Entidades Nomeadas é o trabalho de Covas (2023). O estudo focou na identificação de produtos e serviços, unificados sob uma única entidade PRODUTO, a partir das descrições de empresas em suas páginas da Wikipédia. Para tanto, foi feita uma comparação entre os modelos da biblioteca spaCy e o GPT-3.5. Os resultados mostraram que o GPT superou significativamente os modelos do spaCy, alcançando uma Medida-F1 de aproximadamente 85%. Para guiar o GPT, o autor desenvolveu um *prompt* detalhado, que funcionava como um *template*. Esse *prompt* definia a entidade "PRODUTO" , especificava o formato da saída desejada e fornecia um exemplo de texto de entrada com a correspondente lista de produtos já extraída, caracterizando a técnica de *few-shot prompting*.

3.1.4 Abordagens híbridas

Já as abordagens híbridas incluem modelos que combinam duas ou mais abordagens distintas, como técnicas de AM e estratégias simbólicas. Essa combinação pode ser vantajosa para obter

alta precisão e cobertura, especialmente com poucos dados anotados. Milidiú, Duarte e Cavalcante (2007), por exemplo, exploram a combinação de diferentes algoritmos de AM com um modelo *baseline* baseado em *gazetteers* e regras lexicais relevantes para a tarefa.

3.2 Estado-da-arte para a Língua Portuguesa

Tendo sido apresentadas as principais abordagens de REN e a métrica de avaliação utilizada para compará-las, esta seção detalha o estado-da-arte da tarefa. A análise é dividida em dois cenários principais: primeiramente, o desempenho dos sistemas em textos de linguagem formal e, em seguida, o panorama para o Conteúdo Gerado por Usuário.

3.2.1 REN para textos formais

Para textos de domínio geral em linguagem formal, a tarefa de REN, como demonstrado, vem sendo há muito explorada no PLN. O estado-da-arte atual da tarefa foi atingido por modelos de AM profundo, especificamente baseados em LLMs pré-treinados com *fine-tuning* a partir de *corpus* com anotação humana e representações das palavras por vetores (em inglês, *word embeddings*), os quais atingiram Medida-F1 acima de 90% (Ushio et al., 2022).

Segundo Silva (2023), o modelo com o estado-da-arte para o português é o BERTimbau (Souza; Nogueira; Lotufo, 2020) com 78,6% de Medida-F1 obtida por Souza, Nogueira e Lotufo (2019) ao considerar as 10 categorias gerais do Primeiro HAREM. Embora os resultados tenham avançado, isso mostra que ainda existe muito trabalho a ser feito na área.

Aliás, analisando os trabalhos para o português, observa-se que a maioria deles usa os *corpora* dourados do HAREM no desenvolvimento dos métodos/modelos. A seguir, discorre-se sobre os principais *corpora* dourados (de linguagem formal) com anotação de EN para o português, com especial atenção aos do HAREM, além de recursos lexicais disponíveis para a tarefa.

3.2.2 REN para CGU

A complexidade da tarefa de REN em dados de CGU pode ser confirmada pelo estado-da-arte atual de 53% de Medida-F1 obtido pelo BERTweet (Nguyen; Vu; Nguyen, 2020). Valor esse bem menor que os obtidos para *corpus* com linguagem formal. O BERTweet é uma versão especializada do BERT, treinada especificamente para processar e compreender a linguagem

do Twitter. O modelo foi pré-treinado em 850 milhões de *tweets* em inglês, capturando melhor a linguagem informal das redes sociais. O resultado de 53% de Medida-F1 foi obtido pelo BERTweet em uma avaliação feita com base no *corpus* WNUT-16 (Strauss et al., 2016), que contém 3.856 *tweets* anotados com as categorias PERSON, GEO-LOC, PRODUCT, COMPANY, FACILITY, MOVIE, MUSIC, ARTIST, SPORTS TEAM e TC SHOW (além de OTHER).

Para lidar com esse desempenho inferior, Castro (2019) demonstrou que o "ajuste fino" em *corpus* de domínio anotado e o emprego de vetores estáticos pré-treinados em *corpus* de domínio contribuem para melhorar a tarefa de REN em textos especializados, como é o caso da área jurídica.

Quanto ao processamento do português, identificaram-se os seguintes trabalhos sobre REN para CGU/*tweets*: Peres, Esteves e Maheshwari (2017), Costa (2023) e Zerbinati, Roman e Di-Felippo (2024). Peres, Esteves e Maheshwari (2017) fizeram alguns experimentos utilizando uma variedade de modelos baseados em arquiteturas de LSTM. Entre eles, o BI-LSTM com *word embeddings* pré-treinados (*Glove*[14]) atingiu a melhor performance de Medida-F1 para *tweet* com uma pontuação de 52,78%. Valor esse próximo do obtido para o inglês (53%). Costa (2023) investigou a tarefa de REN para comentários sobre projetos de lei. Nesse cenário, o modelo que obteve melhor desempenho foi o *fine-tuning* do BERT, com uma Medida-F1 de 73,6%. Zerbinati, Roman e Di-Felippo (2024), por sua vez, estudaram a pertinência da informação morfossintática de PoS na tarefa de REN em *tweets* do mercado financeiro.

A investigação da tarefa de REN para *tweets* em inglês conta com vários *corpora* de referência. Além do WNUT-16, tem-se o TweeBank-NER (Jiang et al., 2022), de aproximadamente 3.000 *tweets*, o TweetNER7 (Ushio et al., 2022), com cerca de 2.500, e outros. Todos esses recursos de referência possuem em média 3.000 *tweets* sobre assuntos variados (normalmente, os *trending topics* semanais) e anotação manual dos tipos de EN mais gerais (isto é, pessoa, localização, organização, evento e produto).

Em português, tem-se os seguintes *corpora* de referência que serão apresentações na Seção 3.5: Twitter-NER (Peres; Esteves; Maheshwari, 2017), C-*corpus* (Costa, 2023) e DANTEStocks (Di-Felippo; Nunes; Barbosa, 2024a), anotado por Zerbinati, Roman e Di-Felippo (2024).

[14]https://nlp.stanford.edu/projects/glove/

3.3 Formatos de anotação e desafios linguísticos gerais

Segundo autores como Silva (2023) e Jurafsky e Martin (2025), há alguns esquemas de anotação para o REN na literatura, a saber: IO (*Inside-Outside*) (Pirovani; Oliveira, 2018), BIO (*Begin-Inside-Outside*) (Ramshaw; Marcus, 1999), BIOES (*Beginning, Inside, Outside, End, Single*) (Ratinov; Roth, 2009) e BILOU (*Begin-Inside-Last-Outside-Unit*) (Amaral; Vieira, 2014)

Entre eles, o formato BIO é o mais difundido. Na Figura 3.1, tem-se uma comparação bastante ilustrativa entre os formatos IO, BIO e BIOES, extraída de Jurafsky e Martin (2025). Vale destacar que, em qualquer um desses esquemas, as etiquetas de delimitação da sequência de *tokens* (ou *span*) que representa uma EN são subespecificadas pela categoria da EN em questão, como I-LOC para "Chicago".

Figura 3.1: Diferentes formatos de anotação de entidade nomeada.

Words	IO Label	BIO Label	BIOES Label
Jane	I-PER	B-PER	B-PER
Villanueva	I-PER	I-PER	E-PER
of	O	O	O
United	I-ORG	B-ORG	B-ORG
Airlines	I-ORG	I-ORG	I-ORG
Holding	I-ORG	I-ORG	E-ORG
discussed	O	O	O
the	O	O	O
Chicago	I-LOC	B-LOC	S-LOC
route	O	O	O
.	O	O	O

Fonte: Jurafsky e Martin (2025).

No esquema notacional BIO, anota-se qualquer *token* de uma sequência de interesse com a etiqueta "B-" (*Begin*), os demais *token* que ocorrem no interior do *span* são anotados com "I-" (*Inside*) e quaisquer *tokens* fora do *span* de interesse são anotadas com "O" (*Outside*). Dessa forma, esse tipo de anotação representa exatamente as mesmas informações que a anotação entre colchetes ilustrada no exemplo (1), apresentado na Seção 1.1, com a vantagem de ser mais explícita no tocante à delimitação das ENs, uma vez que envolve sequências de *tokens*.

Comparativamente, o esquema IO perde algumas informações ao eliminar a etiqueta B, distinguindo apenas entidades ("I-") de não-entidades ("O-"). O formato BIOES, por sua vez, adiciona a etiqueta "E" (*End*) para indicar o final de um *span* e a etiqueta "S"(*Single*) para indicar que se trata de *span* composto por apenas um *token*. Ao fazer essas adições, o formato BIOES introduz maior complexidade na anotação e no processamento automático em comparação com o formato BIO. Quanto ao processamento, esse formato requer que um algoritmo

ou modelo de AM, por exemplo, precise distinguir não apenas entre o início e o interior de uma entidade, mas também identificar corretamente o fim e os casos de *tokens* únicos.

Como salienta Freitas (2022), a anotação de ENs impõe alguns desafios linguísticos. O primeiro deles é decidir o que deve ser anotado, isto é, identificar uma "entidade". A resposta para essa questão, segundo a autora, está na tarefa que motiva a própria anotação. Na construção de *corpora* de domínios específicos ou técnicos, por exemplo, os especialistas da área podem desempenhar papel importante na definição do que anotar. No caso do PetroNer, por exemplo, os especialistas da Petrobras, motivados pela tarefa de busca em documentos técnicos realizada por geocientistas, indicaram quais tipos de entidades são importantes nesse cenário. Assim, a dificuldade na identificação foi contornada com a utilização de categorias de entidades (e lista de instâncias) definidas por um grupo de especialistas.

Além disso, a classificação de uma EN também é um desafio, uma vez que a polissemia está presente nessa tarefa. É fato que palavras como *Brasil* podem ser LOCAL (p.ex.: "O Brasil tem muitos rios.") ou ORGANIZAÇÃO (p.ex.: "O Brasil assinou o tratado." e "O Brasil é campeão da Copa América.") a depender do contexto. Mesmo quando "Brasil" é empregado para fazer menção a um time, pode-se interpretá-la como PESSOA, pois uma interpretação não exclui a outra (ORGANIZAÇÃO). Aliás, pode-se optar por uma anotação que explicite ambas as categorias PESSOA|ORGANIZAÇÃO.

Por fim, como ressalta Freitas (2022), a delimitação de uma EN pode ser realizada de diferentes formas, as quais resultam da decisão de se considerar entidades maiores ou entidades mais granulares (ou composicionais). Considerando a expressão "Secretaria de Educação de São Paulo", por exemplo, pode-se ter ao menos duas possibilidades de anotação, ambas corretas a depender da granularidade adotada:

(a) [Secretaria de Educação de São Paulo]ORG
(b) [Secretaria de Educação]ORG de [São Paulo]ORG ou LOC

Outro aspecto que dificulta ou gera hesitação no processo de identificação das EN é o uso pouco sistemático das maiúsculas (Freitas, 2022). E isso fica bastante latente quando se trata de CGU, como ilustrado nos exemplos da Seção 4.1.

3.4 Recursos linguísticos: *corpora* anotados e *gazzeters* em textos formais

Entre os *corpora* de língua geral destacados na Figura 3.2, encontram-se os da avaliação conjunta HAREM (Santos; Cardoso, 2007; Mota; Santos, 2008). Os *corpora* do HAREM foram elaborados com o objetivo de abranger variantes do português de diferentes regiões, reunindo materiais provenientes na sua maioria de Portugal e Brasil, mas também de Angola, Moçambique, Macau, Índia, Timor-Leste e Cabo Verde. A seleção dos textos também considerou a variação de gênero e estilo, incluindo conteúdo extraído da *web*, notícias jornalísticas, transcrições de entrevistas, documentos técnicos provenientes de relatórios *online* e textos políticos.

Figura 3.2: *Corpora* de textos formais em português com anotação dourada de ENs.

Corpus	Categoria	Entidade
Primeiro HAREM	10	5.132
Segundo HAREM	10	7.847
Mini-HAREM	10	3.758
SIGARRA	8	12644

Fonte: Baseada em Silva (2023).

O *corpus* padrão ouro ou "coleção dourada" denominado Primeiro HAREM contém 129 textos e 5.132 ENs anotadas e o Mini-HAREM contém outros 128 textos e 3.758 ENs (Santos; Cardoso, 2007).

O *corpus* Segundo HAREM passou a incluir, além dos já tradicionais textos de notícias e páginas da *Web*, novos gêneros textuais, como *blogs*, *wikis*, enciclopédias (Wikipedia) e questões usadas em avaliações de sistemas de resposta automática a perguntas (Freitas; Carvalho; Oliveira et al., 2010). Diferentemente do Primeiro HAREM, transcrições orais e textos literários foram bem menos utilizados. Embora esse *corpus* possua mais conteúdo advindo da *web*, como *blogs*, ainda assim seu conteúdo majoritariamente apresenta linguagem formal ou padrão. O *corpus* é constituído de 129 documentos, contabilizando um total de 147.991 palavras e 7.847 ENs anotadas (Carvalho; Oliveira; Santos et al., 2008).

O conjunto de etiquetas usado no Segundo HAREM não é significativamente distinto do usado no Primeiro HAREM. Na verdade, em ambos os *corpora*, empregou-se o mesmo conjunto de 10 categorias gerais, com pequena alteração terminológica (por exemplo, a categoria VARIADO do Primeiro Harem passou para OUTRO no Segundo HAREM). As mudanças mais significativas ocorreram no elenco de subtipos, particularmente das categorias LOCAL

e TEMPO. Uma descrição mais detalhada da taxonomia de ENs do HAREM é fornecida mais adiante, uma vez que se utilizam substancialmente suas diretivas de anotação neste trabalho.

Além dos recursos gerados no HAREM, destaca-se o SIGARRA (Pires, 2017). Esse *corpus* é composto por 1.000 notícias coletadas de divisões internas da Universidade do Porto, as quais, por isso, cobrem uma variedade de tópicos relacionados à universidade e suas atividades. As entidades nele presentes estão classificadas em 8 categorias: PESSOA, LOCAL, ORGANIZAÇÃO, DATA, HORA, EVENTO, CURSO e UNIDADE ORGÂNICA (p.ex.: nome de institutos). As 1.000 notícias totalizam aproximadamente 185 mil *tokens*, sendo que, da anotação manual de 905 delas, obteve-se o montante de 12.644 entidades anotadas.

Quanto aos *corpora* de domínio especializados de linguagem formal com anotação de EN de referência, há vários disponíveis, sendo o domínio jurídico um dos mais difundidos (Souza; Albuquerque et al., 2024). Entre os recursos com anotação de referência, citam-se o LeNER-Br (Araujo et al., 2018) e o UlyssesNER-Br (Albuquerque; Costa et al., 2022).

O *corpus* LeNER-Br (Araujo et al., 2018) é composto por 70 documentos, sendo 66 provenientes de tribunais judiciais brasileiros e 4 leis. Além das categorias PESSOA, LOCAL, TEMPO e ORGANIZAÇÃO, o *corpus* contém as categorias específicas LEGISLAÇÃO (para leis) e JURISPRUDÊNCIA (para decisões relacionadas a casos legais).

O *corpus* UlyssesNER-Br (Albuquerque; Costa et al., 2022) é composto por 150 projetos de lei e 800 consultas legislativas da Câmara dos Deputados brasileira. Esse material foi anotado manualmente em três fases e por três grupos de anotadores, considerando 7 categorias de ENs, algumas contendo tipos específicos. A avaliação da concordância entre anotadores alcançou coeficiente Kappa de 91% na média. Quanto às categorias, o *corpus* possui 5 genéricas, baseadas na taxonomia do HAREM (Santos; Cardoso, 2007) (PESSOA, LOCALIZAÇÃO, ORGANIZAÇÃO, EVENTO e DATA), e outras 2 categorias específicas para o domínio legislativo: FUNDAMENTO DE LEI (como resoluções e decretos) e PRODUTO DE LEI (como propostas de lei e consultas legislativas).

Sobre o domínio médico, cita-se o SemClinBr (Oliveira et al., 2022), que contém dados clínicos de hospitais brasileiros. O desenvolvimento do *corpus* envolveu 8 anotadores (estudantes de medicina), 2 adjudicadores (médico e enfermeiro) e 4 pesquisadores de Informática em Saúde, totalizando 14 pessoas. A anotação de EN foi feita em um processo cíclico de treinamento, anotação, adjudicação de discordância e refinamento das diretrizes de anotação de ENs que durou 14 meses. A anotação de ENs foi basicamente feita a partir do extenso elenco de

100 categorias do *Unified Medical Language System* (UMLS), que também possui um número bastante elevado de subcategorias. As ENs foram 100% duplamente anotadas e adjudicadas, resultando em um *corpus* composto por 1.000 documentos (148.033 *tokens*) e 65.129 entidades. O IAA médio estrito foi de aproximadamente 0,71 e o relaxado foi de aproximadamente 0,92.

Para as pesquisas de PLN relativas ao domínio petrolífero, tem-se o PetroNer (Freitas; Souza et al., 2023), que é composto por boletins e relatórios técnicos. Anotação de ENs foi feita em um processo semiautomático composto pelas seguintes etapas: (i) definição das categorias de EN e criação de listas de instâncias das categorias por especialistas de domínio, (ii) anotação automática das ENs com base em regras que utilizam informação morfossintática e o léxico inicial fornecido pelos especialistas e (iii) revisão manual da anotação automática por linguistas e supervisão de especialista da área. A anotação semiautomática de ENs considerou 19 categorias de domínio (como BACIA, ROCHA, POÇO, CAMPO, FLUIDO, TEXTURA, UNIDADE LITOESTRATIGRÁFICA[15], etc.) levou cerca de oito meses e resultou na anotação de quase 20 mil entidades no PetroNer. Todo o trabalho foi auxiliado pelo ambiente Interrogatório, que compõe a ferramenta ET (Souza; Freitas, 2021), uma estação de trabalho para busca, edição e avaliação de arquivos no formato CoNLL-U.

Como salienta Silva (2023), além dos *corpora* anotados, recursos lexicais, os chamados *gazetteers*, podem auxiliar os modelos de REN, sobretudo aqueles desenvolvidos segundo abordagens híbridas ou baseadas em regras. No geral, os *gazetteers* são repositórios lexicais empregados como fonte de conhecimento externo, fornecendo informações que não estão contidas no texto sob processamento e que podem ser úteis para a classificação da entidade.

Em português, tem-se o REPENTINO (acrônimo para REPositório para reconhecimento de ENTIdades NOmeadas para o português) (Sarmento; Pinto; Cabral, 2006), que é um léxico composto de nomes próprios extraídos do corpus WPT03 (isto é, coleção da *Web* portuguesa de 2003)[16]. O REPENTINO contém mais de 45.000 exemplos de entidades nomeadas. O sistema de classificação do REPENTINO é amplo e detalhado, organizado em 11 categorias principais, as quais abrangem 97 subcategorias.

[15]Esse termo está relacionado à litoestratigrafia, que é uma disciplina da geologia focada no estudo das unidades litológicas ou formações rochosas.

[16]https://www.linguateca.pt/WPT/WPT03.html

3.5 Revisão sistemática sobre recursos linguísticos de REN para CGU

A tarefa de REN para a língua portuguesa já foi mapeada em duas revisões anteriores relativamente recentes (Albuquerque; Souza et al., 2023; Silva, 2023), as quais deram maior foco para o reconhecimento de ENs em textos formais ou de linguagem canônica. A de Silva (2023) é a única que inclui CGU. A RS realizada neste trabalho, atualiza atualiza, expande e refina a revisão de Silva (2023), ao focar especificamente em recursos linguísticos, especificamente corpora, para a tarefa de REN em CGU.

Especificamente, as revisões do tipo "sistemática" são ferramentas importantes para sintetizar as evidências disponíveis sobre determinado tópico de pesquisa ou fenômeno de interesse. O objetivo desse tipo de revisão é identificar, analisar e interpretar essas evidências em estudos primários. Em outras palavras, uma RS é uma maneira de acessar o estado-da-arte sobre o tema pesquisado. A realização de uma RS segue etapas bem delimitadas, descritas em um "protocolo de RS", que possibilita a replicação da revisão, atestando a confiabilidade dos resultados obtidos, a redução de vieses durante a pesquisa, além de apontar lacunas no conhecimento atual.

3.5.1 Protocolo da Revisão Sistemática

O protocolo empregado neste trabalho segue as diretrizes de Kitchenham (2004). Com isso, ele é composto por objetivo, questões de pesquisa, palavras-chave, bases de dados utilizadas e critérios de seleção empregados.

a) Objetivo

Investigar na literatura os estudos primários que abordam ou descrevem *corpora* anotados com ENs, bem como modelos de REN para CGU, particularmente em *tweets* em língua portuguesa.

b) Questões de pesquisa

Diante do objetivo traçado, formularam-se as seguintes questões de pesquisa:

1. Quais os *corpora* de CGU com anotação de EN de referência em português?
2. Qual a metodologia empregada para a anotação desses *corpora*?
3. Quais as categorias de ENs consideradas e os critérios para identificá-las?

4. Quais trabalhos sobre REN os *corpora* de CGU com anotação dourada de EN subsidiaram?

c) Palavras-chave

O conjunto de palavras-chave empregado na RS é composto pelos termos mais frequentes da literatura em inglês e suas respectivas traduções em português, assim como variações terminológicas em português (como entidades "nomeadas" e "mencionadas") e siglas. Esse conjunto está subdividido em dois. Um deles engloba os termos mais gerais referentes às entidades e seu reconhecimento automático. O outro possui termos referentes a CGU e *tweets*.

- *named entity*; *name entity*; NE; entidade nomeada; entidade mencionada; EM; *named entity recognition*; NER; reconhecimento de entidade nomeada; REN; reconhecimento de entidade mencionada; REM.
- *user generated content*; UGC; conteúdo gerado por usuário; CGU;*user-generated text*; *noisy text*.

d) Bases de dados

Para a **busca automática**, realizada com base em uma *string* de busca, utilizou-se inicialmente a ACL *Anthology*[17], que é um repositório digital mantido pela *Association for Computational Linguistics* (ACL)[18]. Ele armazena trabalhos científicos da área do PLN e da Linguística Computacional, publicados em anais de conferências e *workshops* e em periódicos, todos eles organizados pela ACL e outras associações afiliadas. No entanto, o filtro por palavras-chave do buscador da ACL *Anthology* é instável, gerando resultados variados para a mesma consulta em momentos diferentes. Por esse motivo, utilizou-se o *Google Scholar*[19] para realizar a busca no repositório da ACL. Nesse caso, os resultados foram limitados às primeiras 10 páginas retornadas pelo buscador. A *string* de busca utilizada foi:

- **ACL:** site:aclanthology.org ((("named entity"OR "name entity"OR "named entities"OR "name entities") AND ("recognition"OR "classification"OR "identification"OR "resolution"OR "detection"OR "categorization")) OR "NER") AND ("user-generated content"OR "UGC"OR "user-generated text"OR "noisy text")

A **busca manual** foi feita nas *venues* específicas dos seguintes eventos indexados pela ACL:

[17]https://aclanthology.org/
[18]https://www.aclweb.org/portal/
[19]https://scholar.google.com/

- W-NUT - *Workshop on Noisy User-generated Text*;
- PROPOR - *International Conference on Computational Processing of Portuguese.*

Para compor a revisão, considerou-se a revisão sistemática de REN realizada por Silva (2023) para a língua portuguesa como uma fonte adicional, por meio da qual foi possível identificar outras publicações relacionadas a CGU que não foram retornadas diretamente pelas buscas, mas que demonstraram pertinência para o objetivo da pesquisa.

2) Critérios de seleção

- **Critérios de Inclusão:**

  1. Trabalhos sobre a tarefa de REN para CGU, independente da língua/domínio.
  2. Trabalhos publicados e disponíveis integralmente nas bases de dados científicas.
  3. Estudos primários publicados a partir de 2014.

- **Critérios de Exclusão:**

  1. Trabalhos que não apresentam descrição de recurso (*corpus*) e/ou método de REN.
  2. Trabalhos que não apresentam avaliação da pesquisa.
  3. Trabalhos que descrevem a tarefa de REN para CGU como parte de um sistema maior, em que a descrição do *corpus*/método é feita de forma superficial.
  4. Trabalhos sobre REN para CGU que focam em apenas uma categoria de EN.
  5. Trabalhos que descrevem tarefas correlatas ao REN (como *entity linking* e desambiguação de entidades), mesmo que utilizem CGU.
  6. Trabalhos escritos em idioma diferente do inglês ou português.
  7. Trabalhos não disponíveis gratuitamente.
  8. Trabalhos que não tenham sido revisados por pares.

3.5.2 Condução da Revisão Sistemática

As revisões sistemáticas podem ser classificadas em restritivas ou amplas, cada uma com suas vantagens e desvantagens (Kitchenham, 2004; Scannavino et al., 2017).

A revisão sistemática restritiva adota critérios de inclusão e exclusão mais rigorosos, o que resulta em um número menor de estudos retornados pela busca. Essa abordagem busca garan-

tir uma seleção mais aprofundada e confiável dos dados. No entanto, sua principal desvantagem é o risco de excluir trabalhos potencialmente relevantes, limitando a representatividade do tema. Além disso, pode introduzir viés na seleção, pois muitos estudos são descartados com base em critérios estritos. Por outro lado, a facilidade de reprodutibilidade é um ponto positivo, já que os critérios bem definidos tornam mais simples a replicação do estudo.

A revisão sistemática ampla tem critérios mais flexíveis, permitindo a inclusão de um número maior de estudos e abordagens. Essa característica torna a revisão mais representativa, captando diferentes perspectivas sobre o tema. No entanto, a análise pode ser menos aprofundada, pois há um grande volume de dados a serem processados, e a inclusão de estudos de menor rigor metodológico pode comprometer a confiabilidade dos resultados. Além disso, devido à variedade e quantidade de estudos considerados, a reprodutibilidade pode ser mais difícil e o tempo necessário para a realização da revisão tende a ser maior.

Neste trabalho, optou-se pela abordagem restritiva, utilizando principalmente o repositório da ACL e as *venues* de eventos científicos específicos. A utilização das referidas fontes oferece maior custo-benefício, uma vez que permite um levantamento mais preciso dos estudos pertinentes ao tema deste trabalho.

Antes, porém, de aplicar a abordagem restritiva, fizeram-se, como orientam Kitchenham (2004) e Scannavino et al. (2017), uma busca piloto de abordagem ampla via *Google Scholar* e uma avaliação das revisões sistemáticas previamente publicadas sobre a tarefa de REN para o português. Isso foi feito com o objetivo de (i) mapear o estado-da-arte de REN e *corpora* anotados, (ii) ajustar critérios de inclusão e exclusão, (iii) identificar palavras-chave e termos alternativos, (iv) definir fontes e bases de dados mais relevantes e (v) compreender a diversidade metodológica de uma RS. A busca piloto foi feita em 22/07/2024 e retornou 155 estudos.

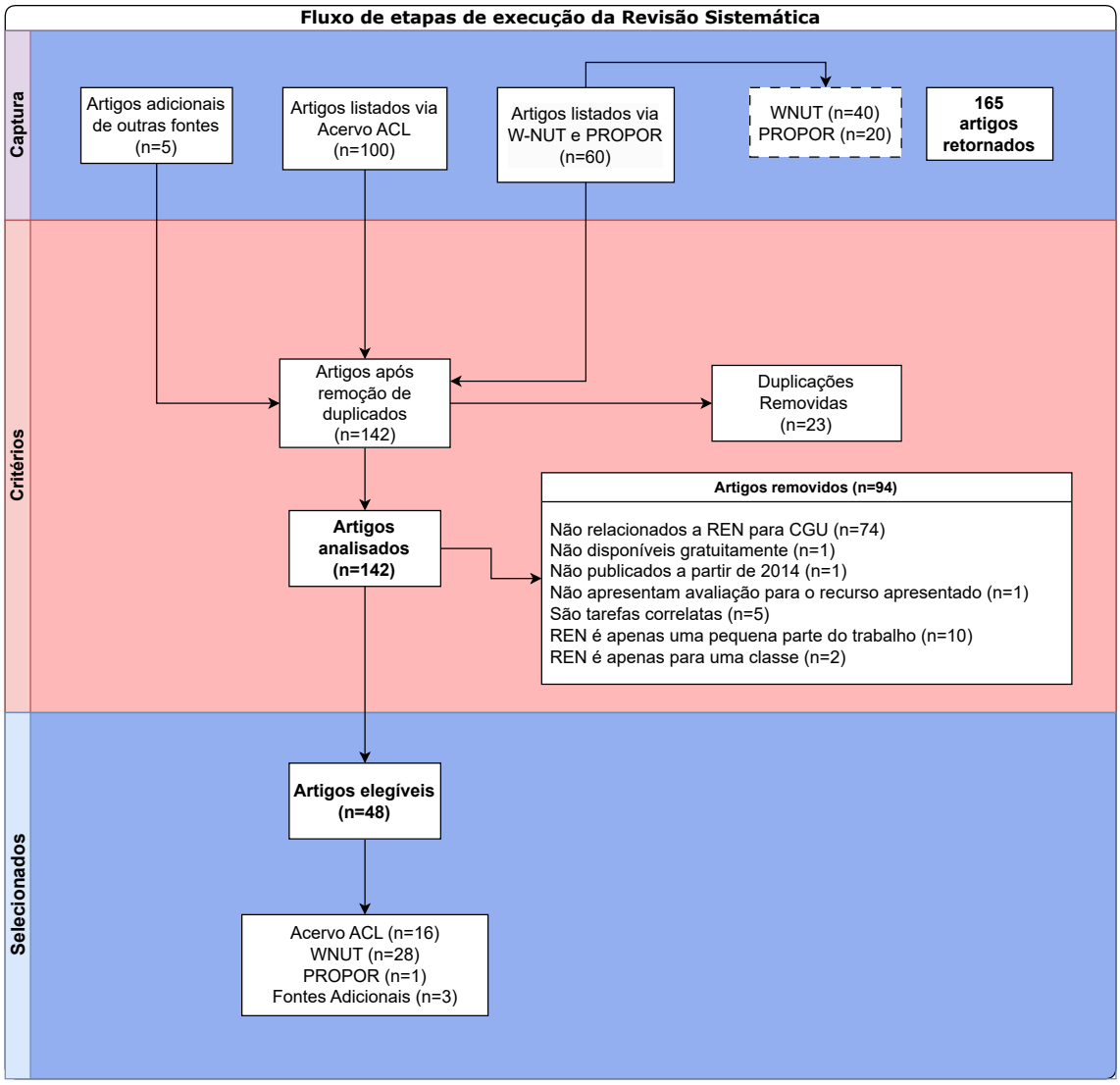
Uma vez que se selecionou a abordagem restritiva, a qual pode deixar estudos relevantes de fora, buscou-se atenuar esse problema lançando mão do "direcionamento de especialistas" como uma estratégia complementar. Segundo Kitchenham (2004), essa é uma prática importante no processo de revisão sistemática porque a consulta a especialistas na área de estudo pode ajudar a definir as questões de pesquisa, desenvolver o protocolo de revisão, e avaliar a qualidade e relevância dos estudos incluídos na revisão. Ainda sobre isso, Kitchenham (2004) ressalta que os especialistas podem ajudar a identificar estudos relevantes que puderam não ser recuperados, mesmo em buscas bem estruturadas, devido a limitações dos mecanismos de busca ou à abrangência das bases consultadas.

Além do acervo da ACL e dos eventos PROPOR e WNUT, este trabalho considerou outras publicações pertinentes (Teteo et al., 2019; Silva, 2023; Costa, 2023) advindas de fontes adicionais como a Biblioteca Digital Aberta da SBC, a revista Linguamática e o repositório de teses/dissertações da UFG, respectivamente (cf. Apêndice A.1). Nos casos em que alguma edição do evento, dentro do intervalo considerado (isto é, publicados a partir de 2014), não estivesse indexada, a busca foi realizada diretamente no *site* do evento. Os resultados foram reunidos em uma única base de dados e as duplicações foram removidas.

Após o levantamento inicial de artigos, aplicaram-se os critérios de inclusão e exclusão 6, 7 e 8, com base na análise do título, resumo e palavras-chave. Esse processo resultou em uma lista de artigos pré-selecionados. Em seguida, uma nova etapa de triagem foi realizada, na qual foram aplicados os demais critérios de exclusão. Nessa fase, analisaram-se as seções de introdução e metodologia, com especial atenção à apresentação dos *datasets* ou *corpora*, além da conclusão. Como resultado dessa filtragem, obteve-se a lista final de artigos elegíveis para a etapa de sumarização.

A revisão foi conduzida entre os dias 23/07/2024 e 24/08/2024. Dos 165 artigos inicialmente retornados, 48 foram selecionados, os quais estão listados no Apêndice A.1. O processo completo da RS está ilustrado na Figura 3.3. Na próxima seção, fornecem-se as respostas às questões de pesquisa com base na RS ora descrita.

Figura 3.3: Fluxo de etapas de execução da Revisão Sistemática.



Fonte: A autora, 2025.

3.5.3 Resultados: respostas às questões de pesquisa

1. **Quais os *corpora* de CGU com anotação de EN de referência em português?**

Como mencionado, a anotação dourada de um *corpus* foi conduzida (ou revisada) manualmente por especialistas para garantir a máxima precisão e qualidade. Diz-se, por isso, que a anotação contou com a curadoria humana.

Com base na revisão sistemática, identificaram-se 3 *corpora* de CGU em português com anotação dourada de ENs: Twitter-NER (Peres; Esteves; Maheshwari, 2017), C-*corpus* (Costa, 2023) e DANTEStocks (Zerbinati; Roman; Di-Felippo, 2024).

O Twitter-NER engloba 3.968 *tweets* sobre temas variados, coletados via API do Twitter em 2017. O total de *tokens* é 44.951 e o número médio de *tokens* por *tweet* é de 11,32. O C-*corpus* é um conjunto de 1.269 comentários de cidadãos brasileiros a respeito de projetos de lei brasileiros coletados do portal da Câmara dos Deputados. O DANTEStocks é um *corpus* de 4.048 *tweets* do mercado financeiro. Trata-se de um *tweebank* com anotação sintática segundo o modelo *Universal Dependencies* (UD) (Nivre et al., 2020).

Além desses recursos, Teteo et al. (2019), ao proporem um *framework* para extração, tratamento e identificação de eventos e localidades em *tweets* escritos em português, treinam um modelo estatístico de REN com base em um conjunto de *tweets*, cujo número total de *posts* não está claramente descrito no artigo. Esse conjunto foi automaticamente etiquetado e não passou por um processo de revisão humana. Dessa forma, sua anotação de EN é considerada prateada e, por isso, excluída deste trabalho. Tedeschi e Navigli (2022) citam um *corpus* multilíngue de CGU com anotação de EN que inclui a língua portuguesa. Esse recurso (denominado MultiNERD) é composto por material advindo da Wikipedia, considerada de linguagem mais formal que o *tweet*. Ademais, assim como o de Teteo et al. (2019), esse recurso possui anotação automática. Por essas razões, ele não foi considerado neste trabalho.

2. Qual a metodologia empregada para a anotação desses *corpora*?

Todos os 3 recursos de referência possuem anotação manual de ENs.

Sobre o Twitter-NER, a anotação foi realizada inteiramente de forma manual com o auxílio da ferramenta BRAT (*Rapid Annotation Tool*) (Stenetorp et al., 2012) e seguindo as diretrizes de Finin et al. (2010). O BRAT, que possui interface gráfica amigável, gera anotação de ENs no formato *.ann*, que se assemelha ao BIO, mas com diferenças. Em vez de usar as *tags* "B-", "I-" e "O" diretamente, a identificação das entidades é feita por intervalos de posição (dos caracteres) no texto, sem indicar explicitamente se uma entidade está dentro de outra. Assim, para "Barack Obama viajou.", tem que o *token* T1, correspondente ao intervalo de caracteres 0-12, é da categoria PERSON. Como mencionado na introdução deste trabalho, não foi possível identificar informações sobre o perfil e a quantidade de anotadores, assim como sobre a qualidade ou consistência da anotação.

A anotação manual do C-*corpus* foi conduzida por 3 equipes de 3 especialistas e em 2 fases, uma vez que os modelos treinados pelos autores com base no conjunto inicial de 969 comentários anotados obtiveram desempenho baixo para categorias mais raras, o que levou à condução

da segunda fase, similar à primeira, em que se anotaram mais 300 comentários, totalizando, assim, os 1.269 que compõem o *corpus*. Cada equipe era composta por dois alunos de graduação responsáveis pela anotação e um aluno de pós-graduação que atuava como curador. No geral, ambas as fases de anotação englobaram: (i) treinamento dos anotadores, (ii) anotação propriamente dita, (iii) reuniões periódicas entre anotadores e curadores para solução de dúvidas e (iv) cálculo, também periódico, da concordância via *Kappa* de *Cohen*. A anotação desse material foi feita com o auxílio de um manual de anotação e da ferramenta INCEpTION (Klie et al., 2018), que gera anotações no formato *.ann* (BRAT), BIO e outros. Embora Costa (2023) comente sobre a realização do cálculo de IAA em cada uma das 3 fases de anotação, a autora não fornece os respectivos valores e nem o tempo que a anotação levou para ser feita.

Quanto ao DANTEStocks, ressalta-se que a anotação de ENs foi inteiramente manual e conduzida por apenas um cientista da computação, sem nenhum tipo de avaliação da qualidade ou consistência da anotação. Assim como o C-*corpus*, a anotação do DANTEStocks também foi feita com o auxílio de um manual de anotação. Porém, diferentemente dos demais *corpora* de CGU, a delimitação e classificação das ENs não foram feitas por meio de uma ferramenta específica de anotação de *corpus* como BRAT ou INCEpTION. O anotador utilizou um editor de texto genérico e inseriu as etiquetas de ENs na 10 coluna de cada arquivo CoNLL-U, referente a cada *tweet* anotado segundo o modelo UD[20]. Para tanto, empregou-se o formato BIOES.

3. Quais as categorias de ENs consideradas e os critérios para identificá-las?

No Twitter-NER, considerou-se o elenco de categorias composto por PESSOA, LOCAL e ORGANIZAÇÃO (conhecido como PLO), sendo que as diretrizes gerais de identificação das ENs baseadas em Finin et al. (2010) foram: (i) anotar entidades que se referem exclusivamente a um objeto pelo seu nome próprio (“Barack Obama”), acrônimo (“IBM”), apelido (“Opra”) ou abreviação (“Minn.” para “Minnesota”) e (ii) classificar as entidades em função do contexto.

A categoria PESSOA ficou limitada a humanos (vivos, falecidos, fictícios, divindades, etc.). Para a identificação dessas entidades, os títulos e cargos (p.ex.: “Sr.”, “presidente” e “treinador”) que acompanham os nomes próprios são excluídos, ao passo que sufixos (p.ex.: “Jr.” e “III”) são considerados parte da EN. A categoria ORGANIZAÇÃO engloba corporações, instituições, agências governamentais e outros grupos de pessoas definidos por uma estrutura organizacional estabelecida, como empresas (p.ex.: “Bridgestone Sports Co.”), símbolos de ações

[20] A anotação sintática do DANTEStocks e o formato CoNLL serão apresentados no capítulo seguinte.

(ou *tickers*) (p.ex.: "NASDAQ"), organizações multinacionais (p.ex.: "União Europeia"), partidos políticos, entidades governamentais não-genéricas (p.ex.: "Departamento de Estado"), times esportivos e grupos militares. Já a categoria LOCAL inclui nomes de lugares definidos politicamente ou geograficamente, como cidades, províncias, países, regiões internacionais, montanhas, além de estruturas artificiais, como aeroportos, ruas, monumentos, entre outras formações.

O C-*corpus* emprega as mesmas categorias e tipos do UlyssesNER-Br (Albuquerque; Costa et al., 2022), sendo 5 delas baseadas no HAREM (Santos; Cardoso, 2007) (PESSOA, LOCAL, ORGANIZAÇÃO, EVENTO e DATA) e outras 2 criadas para atender às necessidades do domínio legislativo (FUNDAMENTO e PRODUTOS DE LEI) (Figura 3.4). Costa (2023) não fornece diretrizes claras de delimitação das ENs.

Figura 3.4: Categorias de entidades anotadas no C-*corpus*.

Categoria	Tipo	Descrição	Exemplo
DATA	—	Data	01 de janeiro de 2020
EVENTO	—	Evento	Eleições de 2018
FUNDAMENTO	FUNDlei	Norma legal	Lei no 8.666, de 21 de junho de 1993
	FUNDapelido	Apelido da norma legal	Estatuto da Pessoa com Deficiência
	FUNDprojetodelei	Projeto de lei	PEC 187/2016
LOCAL	LOCALconcreto	Local concreto	Niterói-RJ
	LOCALvirtual	Local virtual	Jornal de Notícias
ORGANIZAÇÃO	ORGpartido	Partido político	PSB
	ORGgovernamental	Organização governamental	Câmara dos Deputados
	ORGnãogovernamental	Organização não governamental	Conselho Reg. de Medicina (CRM)
PESSOA	PESSOAindividual	Indivíduo	Jorge Sampaio
	PESSOAgrupoind	Grupo de indivíduos	Família Setúbal
	PESSOAcargo	Cargo	Deputado
	PESSOAgrupocargo	Grupo cargo	Parlamentares
PRODUTO DE LEI	PRODUTOsistema	Sistema	Sistema Único de Saúde (SUS)
	PRODUTOprograma	Programa	Programa Minha Casa, Minha Vida
	PRODUTOoutros	Outros produtos	Fundo partidário

Fonte: Costa (2023).

Zerbinati, Roman e Di-Felippo (2024), para a anotação do DANTEStocks, seguiram os critérios de identificação do Segundo HAREM Mota e Santos (2008), assim como o elenco de 10 categorias gerais: ABSTRAÇÃO, ACONTECIMENTO, COISA, LOCAL, OBRA, ORGANIZAÇÃO, PESSOA, TEMPO, VALOR e OUTRO. Essa primeira anotação de ENs no *corpus* será apresentada com mais detalhes no Capítulo 4, em que se descreve o referido *corpus*, uma vez que é o foco deste trabalho, assim como suas características linguísticas e anotações prévias.

4. Quais são as ferramentas/sistemas de REN para CGU em português?

Quanto ao português, identificaram-se os seguintes trabalhos sobre REN para CGU/tweets: Peres, Esteves e Maheshwari (2017), Teteo et al. (2019) e Costa (2023). Entre eles, Peres, Esteves e Maheshwari (2017) e Teteo et al. (2019) são os únicos que focaram no gênero *tweet*.

Utilizando especificamente o *corpus* dourado Twitter-NER, Peres, Esteves e Maheshwari (2017) fizeram vários experimentos utilizando uma variedade de modelos baseados em LSTM. Entre eles, o BI-LSTM com *word embeddings* pré-treinado Glove atingiu a melhor performance de Medida-F1 para *tweet*, isto é, 52,78%. O resultado obtido por esses autores é bastante próximo ao valor de 53% obtido para o inglês por Strauss et al. (2016).

Teteo et al. (2019), empregando um *corpus* com anotação prateada, investigaram a identificação de entidades das categorias EVENTO e LOCALIDADE em *tweets* sobre trânsito. Para tanto, os autores adaptaram o modelo probabilístico CRF (*Conditional Random Field*) do Stanford-NER (Finkel; Grenager; Manning, 2005b) para o português, atingindo Medida-F1 de 96,76%.

Costa (2023), com base na coleção dourada C-*corpus*, investigou a tarefa de REN para comentários sobre projetos de lei. Nesse cenário, o modelo que obteve melhor desempenho foi o *fine-tuning* do BERT, com uma Medida-F1 de 73,6%.

Os sistemas de REN para outras línguas diferentes do português que resultaram na RS estão listados no Apêndice B.1.

3.5.4 Comparativo dos *corpora* anotados para CGU encontrados na RS

A análise dos *corpora* CGU anotados para a tarefa de REN revela a escassez de recursos disponíveis. Dos 148 artigos revisados, apenas cinco propuseram recursos para a língua portuguesa e, entre eles, apenas três contêm anotação dourada, a saber: Twitter-NER (Peres; Esteves; Maheshwari, 2017), C-Corpus (Costa, 2023) e DANTEStocks (Zerbinati; Roman; Di-Felippo, 2024).

No que diz respeito ao método de anotação, a abordagem manual foi predominante entre os três recursos identificados. Já a definição de entidade nomeada varia conforme as diretrizes adotadas em cada projeto.

A comparação dos conjuntos de categorias revela semelhanças e diferenças entre os *corpora*. O Twitter-NER adota três categorias básicas: PESSOA, LOCAL e ORGANIZAÇÃO. O C-*corpus* expande esse conjunto ao incluir EVENTO e DATA, além de duas categorias específicas para o domínio legislativo: FUNDAMENTO e PRODUTOS DE LEI. Já o DANTEStocks apresenta um esquema mais amplo, composto por ABSTRAÇÃO, ACONTECIMENTO, COISA,

LOCAL, OBRA, ORGANIZAÇÃO, PESSOA, TEMPO, VALOR e OUTRO. Apesar dessas diferenças, todos os *corpora* analisados incluem as categorias PESSOA, LOCAL e ORGANIZAÇÃO, consideradas prototípicas na tarefa de REN. Embora o C-*corpus* e o DANTEStocks compartilhem a influência do HAREM, o primeiro introduz categorias especializadas para o domínio legislativo e o segundo mantém as categorias genéricas.

Já as diretrizes para a identificação de ENs variam entre os *corpora*. O C-*corpus* não apresenta diretrizes claramente documentadas. Em contrapartida, o DANTEStocks segue basicamente as diretrizes do Segundo HAREM, enquanto o Twitter-NER baseia-se em Finin et al. (2010). Uma distinção relevante entre esses dois últimos é que, no Twitter-NER, títulos e pronomes de tratamento não são anotados como pertencentes à categoria PESSOA, diferentemente do HAREM. Além disso, a categoria de LOCAL no Twitter-NER limita-se a lugares concretos, como regiões geográficas e construções físicas, enquanto que, no DANTEStocks, ela engloba locais abstratos, como espaços virtuais.

Capítulo 4

O *tweebank* DANTEStocks

Neste Capítulo, apresenta-se o *corpus* selecionado para este trabalho. Trata-se do *tweebank* DANTEStocks (Di-Felippo; Nunes; Barbosa, 2024a) que é um *corpus* de *tweets* em língua portuguesa do domínio do mercado financeiro, originado a partir dos 4.517 *tweets* compilados por Silva, Roman e Carvalho (2020)[21] em 2014. A compilação dos dados foi automática a partir da ocorrência de pelo menos um *ticker* de uma das 73 ações que compunham, naquele ano, o índice Ibovespa, principal indicador de desempenho das ações negociadas na bolsa brasileira, a Brasil, Bolsa, Balcão (B3). Um *ticker* é uma sequência alfanumérica de cinco ou seis caracteres que representa um tipo específico de ação de uma empresa, como "PETR4" para a ação preferencial da Petrobras no exemplo (2). O emprego desse critério se justifica pelo fato de que os usuários ligados ao mercado financeiro geralmente usam esses termos para se referir às ações e também às empresas.

A exposição está organizada em duas seções principais. A Seção 4.1 detalha as características linguísticas e estruturais do DANTEStocks. A Seção 4.2, por sua vez, descreve as camadas de anotação pré-existentes, com especial atenção à primeira anotação de Entidades Nomeadas de Zerbinati, Roman e Di-Felippo (2024), que é o ponto de partida para o presente estudo.

4.1 Características estruturais e lexicais

Sobre as características estruturais dos *tweets* do DANTEStocks, destaca-se inicialmente a sua brevidade, uma vez que o limite de caracteres vigente à época da sua compilação era de 140[22]. Também, Di-Felippo, Postali, Ceregatto, Gazana, Silva et al. (2021) apontam que esse *corpus*

[21] https://www.kaggle.com/fernandojvdasilva/stock-tweets-ptbr-emotions/data.
[22] O limite de caracteres foi expandido para 280 em 2017.

apresenta uma combinação de linguagem padrão e não-padrão, além de a ocorrência de alguns fenômenos (lexicais) típicos de CGU, como ilustrados por meio dos *tweets* **(2)** ao **(5)**.

(2) #PETR4. **Petrobrás**? Deixa derreter. :-D http://t.co/tBP4xCQ6dq
(3) Cadê a #PETRD15? #PETR4 #IBOV **petrobras**.
(4) Acionistas votam hoje em fusão da ALL com Rumo, Santander, **Petrobras** e mais 6 no radar: **Petr**... http://t.co/PQ31oKjQ1m #infomoney #vale5.
(5) #PETR4 Mensal da **PETROFUMO**. (mensagem: 951011) http://t.co/cm13ALseDU.

No caso, esses fenômenos envolvem a organização Petrobras. Nos exemplos (2), (3) e (4), observa-se a ocorrência de informalidade ortográfica, marcada pelo uso pouco sistemático das maiúsculas e equívocos de acentuação. Além da grafia oficial em (4), isto é, "Petrobras", os demais *tweets* apresentam outras duas variações, "Petrobrás" (2) e "petrobras" (3). No *tweet* (4), tem-se ainda um caso de truncamento, que é "Petr", indicado pelas reticências após esse *token*. O exemplo (5), por fim, contém o neologismo "PETROFUMO", resultante da combinação de "Petrobras" e "fumo". No contexto do mercado financeiro brasileiro, "fumo" é uma gíria que pode significar prejuízo ou situação desfavorável. Portanto, "Petrofumo" é um apelido pejorativo usado por alguns usuários no Twitter à época da compilação do *corpus* DANTEStocks no ano de 2014 para expressar insatisfação ou críticas à Petrobras, especialmente no que se refere às ações dessa organização na bolsa de valores.

Já os exemplos **(6)** ao **(11)** ilustram características estruturais do *corpus*. Destaca-se que ele possui *tweets* formados por uma ou mais sentenças bem delimitadas (6) e (7), mas também por *tweets* que apresentam ausência de pontuação (8) ou pontuação equivocada (9). *Tweets* com disfluências (10) e justaposição de fragmentados (11) também ocorrem. Em (10), tem-se um truncamento (ou quebra), indicado pelas reticências, e que resulta, nesse caso, em uma estrutura sintática incompleta. Todas essas características impõem desafios para qualquer anotação linguística.

(6) No momento PETR4 respeita o suporte de R$ 15,42.
(7) Um motivo a menos para a alta da PETR4. Que venha a correção!
(8) O #PT conseguiu fazer propaganda eleitoral antecipada O que a @dilmabr tem a dizer sobre isso?
(9) Bom dia Marcos, Alguma previsão para petr4?!

(10) Petrobrás PN (PETR4), Gráfico Semanal. Estudo das... http://t.co/5bHkUTy8AC

(11) #OIBR4 (mensagem: 956643) http://t.co/VD2ApxqWqR

Sobre as idiossincrasias léxico-ortográficas, Scandarolli et al. (2023) fornecem uma taxonomia organizada em duas grandes classes de fenômenos, denominadas: (i) variação da norma padrão e (ii) norma inovadora. A Figura 4.1 mostra uma versão expandida da taxonomia.

Figura 4.1: Taxonomia de fenômenos lexicais e ortográficos do DANTEStocks.

Phenomenon	Type	Subtype	Attested example	Standard form	Gloss
Standard Norm Variation	Substitution	*Diacritic* (cedilla)	**lancamento** das notas	*lançamento*	`notes issuing´
		Other	**segunda feira** **Neh?**	*segunda-feira* Né? (não é)	`Monday´ `Right?´
	Omission	*Diacritic*	capital **proprio**	*capital próprio*	`equity capital´
		Other	**valu** ferris	*valeu ferris*	`thanks ferry´
	Insertion	*Diacritic*	#PETR4 **fêz** uma Onda 2	*#PETR4 fez uma Onda 2*	`#PETR4 made a Wave 2´
		Other	montar um **Streaddle**	*montar um Straddle*	`to set a Straddle´
	Transposition	-	vc se manteve na **comrpa**?	*vc se manteve na compra?*	`did you stick with stocks?´
Innovative Norm	Abbreviation	*Initialism*	ação de **LP**	*ação de longo prazo*	`long-term stock´
		Shortening	(eles) falam **q** por **enqt**	(eles) *falam que* *por enquanto*	`(they) say that´ `for now´
		Contraction	**pq** será?	*por que será?*	`I wonder why´
	Neologism	*Agglutination*	44.6k no **Ibolixo**	*44.6 mil no Ibolixo*	`44.6 thousand in Ibotrash´
		Derivation	**diretassa** do morgan	*diretaça do morgan*	`straight from morgan´
		Foreign influence	#itub4 **estopou**	*#itub4 estopou*	`#itub4 stopped´
	Expressiveness	*Graphemic stretching*	**choooooram**!	*choram*	`Cry!´
		Punctuation repetition	linda!!!!	*linda!*	`beautiful!´
		Dialectal variation	De **zóio**!	*De olho!*	`(I am) keeping an eye!´
		Pictogram	:) 😊 muito $	- *muito dinheiro*	`smile´ `much money´
		Capitalization	**LINNDAA**	*linda*	`beautiful´
		Disguise	essa **p******a**	*essa puta*	`this bitch´
	Homophone Writing	*Phonetization*	é **d+**	*é demais*	`(it) is awesome´
		Graphemic substitution	**xatiado**	*chateado*	`upset´
		Onomatopoeia	**hahaha**	-	-
	Medium-dependent token	*Hashtag*	Presidente da **#PETR4**	-	`President of #PETR4´
		At-mention	né, **@user**?	não é, @user?	`isn´t it, @user?´
		URL	http://t.co/OQ3rDdWilf	-	-
		RT	**RT** @user...	-	-
		Truncation	ação sobe **fo**...	*ação sobe forte...*	`Stock rises sharply´
		Code-switching	E ponto final! **PERIOD**!	-	`Full stop! PERIOD´
	Domain-specific token	*Ticker*	**PETR4** subiu	-	`PETR4 went up´
		Cashtag	**$PBR** testando	-	`$PBR (is) testing´
		Decimal number	de **18,xx** a 21,00	-	`from 18,xx to 21,00´
		Valuation rate	ELET6 **+2,09%**	*ELET6 + 2,09 %*	-
		Temporal expression	**1T14, jun/14**	-	`first quarter of 2014`
		Monetary value	perdeu só **R$20,00**	*perdeu só R$ 20,00*	`(it) only lost R$ 20,00

Fonte: Baseada em Scandarolli et al. (2023).

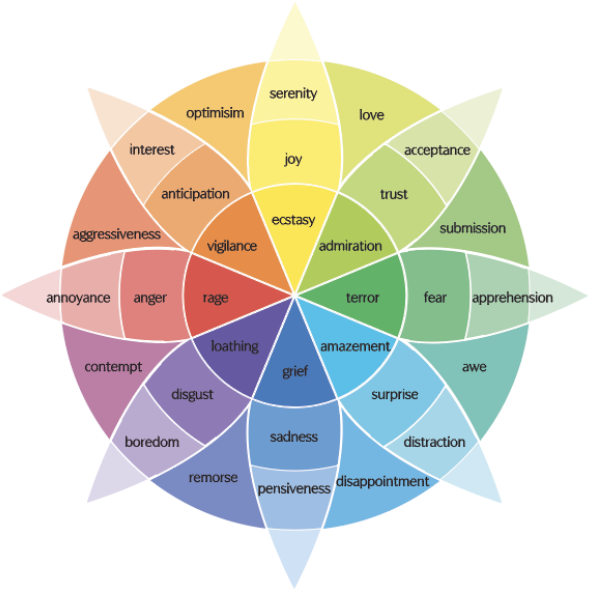
4.2 Anotações prévias

O DANTEStocks tem 3 tipos de anotação: emoção (Silva; Roman; Carvalho, 2020), gramatical (Di-Felippo; Nunes; Barbosa, 2024a) e entidades nomeadas (Zerbinati; Roman; Di-Felippo, 2024).

4.2.1 Emoções

As emoções no DANTEStocks foram anotadas com base na "Roda de Emoções" (em inglês, *Wheel of Emotions* de Plutchik e Kellerman (1980), que categoriza as emoções em 4 eixos fundamentais. Os eixos são compostos por pares de emoções opostas: (i) *joy* v. *sadness*, (ii) *anger* v. *fear*, (iii) *trust* v. *disgust* e (iv) *surprise* v. *antecipation* (cf. Figura 4.2).

Figura 4.2: "Roda de Emoções" de Plutchik.



Fonte: https://en.m.wikipedia.org/wiki/File:Plutchik-wheel.svg.

A anotação de emoção foi feita de forma colaborativa via *web* (*crowdsourcing*[23]), buscando garantir sobretudo a pluralidade dos anotadores. Todos os 4.517 *tweets* originalmente coletados por Silva, Roman e Carvalho (2020) foram submetidos à anotação por um conjunto de 442 voluntários, garantindo que todos os *tweets* fossem etiquetados por pelo menos 3 voluntários diferentes. Para cada *tweet*, os anotadores tinham de selecionar entre uma emoção de cada par oposto, *neutro* ou *não sei*. Do total de 4.517, 240 *tweets* foram descartados por terem sido anotados com *não sei* em todos os pares de emoções pela maioria dos anotadores, resultando em 4.277 *tweets* anotados.

Quando a maioria dos anotadores (isto é, 2/3) escolheu determinada emoção, o *tweet* foi então rotulado como tal. O *tweet* (12) , por exemplo, tem as seguintes etiquetas de emoções finais: *digust*, *sad* e *anger*. Como avaliação da confiabilidade das anotações, cada emoção etiquetada foi avaliada com base na proporção de anotadores que escolheram aquela emoção.

[23] Abordagem colaborativa na qual tarefas de anotação são distribuídas para um grande número de anotadores, muitas vezes anônimos, por meio de plataformas *online*.

Dos 4.277 *tweets*, 2.340 receberam uma etiqueta majoritária em, no mínimo, um par de emoções, enquanto o restante foi classificado como neutro. Isso quer dizer que mais de 50% dos *tweets* do *corpus* foram rotulados com pelo menos uma emoção, o que torna a anotação útil como padrão-ouro.

(12) vai, oibr4. um troux... ops... investidor precisa pagar as minhas férias.

4.2.2 Anotação gramatical

O DANTEStocks possui anotação segundo o modelo *Universal Dependencies* (Nivre et al., 2020), que resulta de uma iniciativa colaborativa internacional que objetiva criar padrões consistentes de anotação para representar a estrutura gramatical de diferentes línguas/gêneros. Desde 2021, o projeto POeTiSa[24] se dedica à construção de um *corpus* multigênero significativo para fomentar o desenvolvimento de ferramentas e sistemas de análise sintático-semântica com base no modelo UD. Esse *corpus*, nomeado *Porttinari*, engloba atualmente 2 porções de gêneros distintos (jornalístico e CGU) com anotação-UD. O DANTEStocks corresponde à porção de CGU do *corpus* multigênero *Porttinari*.

a) Pressupostos gerais da UD

Trata-se de um modelo de dependência, que prevê 2 níveis de representação. No nível morfológico, especificam-se lema, categoria morfossintática e traços gramaticais (*features*). No nível sintático, a anotação se dá por relações de dependência (*deprels*) binárias e assimétricas. Ademais, ressalta-se que o modelo possui 17 etiquetas de PoS[25] e 37 relações de dependência[26] (*deprels*). A anotação-UD é codificada no formato CoNLL-U, como ilustrado na Figura 4.3. Nele, tem-se a anotação-UD do *tweet* (13), extraído do DANTEStocks.

(13) Olá, a bolsa de valores hoje caiu com as ações PETR4.

Cada uma das 10 colunas do CoNLL-U é destinada a uma informação específica:

1. ID: o identificador da posição do *token* na sentença (índice numérico a partir de 1).
2. FORM: *token* na forma como ocorre na sentença.
3. LEMMA: lema ou forma canônica da palavra.

[24] https://sites.google.com/icmc.usp.br/poetisa
[25] https://universaldependencies.org/u/pos/
[26] https://universaldependencies.org/u/dep/

Figura 4.3: Formato CoNLL-U do modelo UD.

# sent_id = dante_01_457299748639477760l									
# text = Olá, a bolsa de valores hoje caiu com as ações PETR4									
Id	Form	Lemma	Upos Tag	Xpos Tag	Feats	Head	DepRel	Deps	Misc
1	Olá	olá	INTJ	_	_	8	discourse	_	SpaceAfter=No
2	,	,	PUNCT	_	_	1	punct	_	_
3	a	o	DET	_	Definite=Def|Gender=Fem|Number=Sing|PronType=Art	4	det	_	_
4	bolsa	bolsa	PROPN	_	_	8	nsubj	_	_
5	de	de	ADP	_	_	6	case	_	_
6	valores	valor	NOUN	_	Gender=Masc|Number=Plur	4	nmod	_	_
7	hoje	hoje	ADV	_	_	8	advmod	_	_
8	caiu	cair	VERB	_	Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin	0	root	_	_
9	com	com	ADP	_	_	11	case	_	_
10	as	o	DET	_	Definite=Def|Gender=Fem|Number=Plur|PronType=Art	11	det	_	_
11	ações	ação	NOUN	_	Gender=Fem|Number=Plur	8	obl	_	_
12	PETR4	PETR4	PROPN	_	_	11	nmod	_	SpaceAfter=No

Fonte: A autora, 2025.

4. POS: etiqueta de classe de palavra (ou *part-of-speech* (PoS) *tag*).
5. XPOS: etiqueta PoS específica da língua.
6. FEAT: atributos morfológicos do *token*.
7. HEAD: ID do *head* da *deprel* cujo token (dependente) que está sendo descrito.
8. DEPREL: relação de dependência que conecta o *token* ao seu *head*.
9. DEPS: relação de *enhanced dependency* do *token*.
10. MISC: informações adicionais sobre o *token*.

A partir de um arquivo no formato CoNLL-U, várias ferramentas de visualização geram representações em grafo. O grafo da Figura 4.4 foi gerado por uma dessas ferramentas[27] a partir do CoNLL-U da Figura 4.3, que contém a anotação da sentença (13). Nela, vê-se que apenas um *token* é o **root**[28] da árvore e que as *deprels* estão indicadas por setas rotuladas que se originam no *head* e se destinam ao dependente. O *token* destacado *caiu* é o *root* e suas informações morfológicas estão no retângulo cinza; ele é o *head* das *deprels* **advmod** (modificador adverbial), **nsubj** (sujeito), **discourse** (discurso) e **obl** (nominal oblíquo).

Seguindo a decisão do projeto POeTiSA, a anotação-UD do DANTEStocks foi fatorada nos níveis morfológico e sintático, pois, segundo Pardo et al. (2021), a separação dos níveis produz melhores resultados dado que a tarefa é sofisticada. Antes, porém, da anotação em si, o *corpus* passou por um pré-processamento.

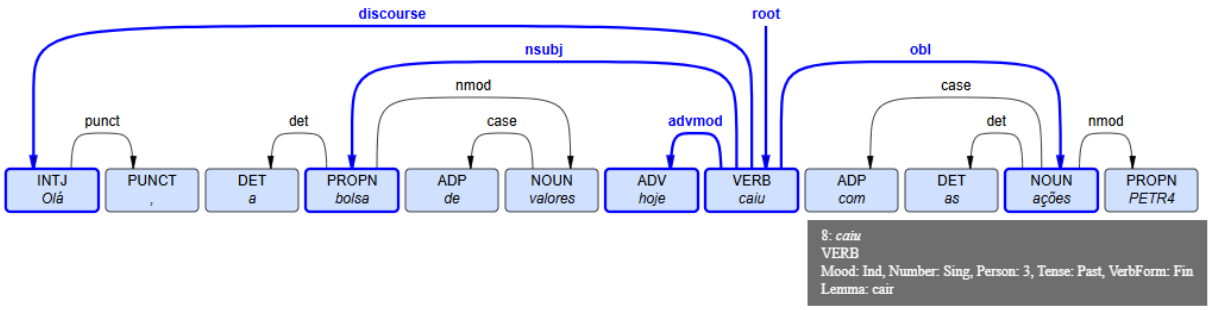
b) Pré-processamento do DANTEStocks

O conjunto inicial de 4.517 *tweets* de Silva, Roman e Carvalho (2020) passou por um refinamento, consistindo na exclusão de 469 *tweets* repetidos e/ou não pertencentes ao domínio.

[27] https://urd2.let.rug.nl/ kleiweg/conllu/

[28] Toda sentença tem uma raiz, normalmente o predicado da oração principal, como dependente da *deprel* **root**.

Figura 4.4: Exemplo de representação arbórea da anotação-UD.



Fonte: A autora, 2025.

O refinamento resultou em 4.048 *posts* que foram efetivamente submetidos à anotação-UD. Destaca-se que o *tweet* foi tomado como unidade de análise e, com isso, os *posts* não passaram por nenhum processo de segmentação em unidades estruturais menores, como sentenças ou sintagmas, assim como de normalização.

Na sequência, o *corpus* foi submetido à *tokenização* que, segundo o modelo UD, consiste em identificar as *palavras sintáticas*[29]. Para tanto, utilizou-se uma versão do NLTK TweetTokenizer enriquecida com regras para o DANTEStocks (Silva; Pardo et al., 2021) baseadas na taxonomia de Scandarolli et al. (2023). A ferramenta preservou a maioria dos *tokens* delimitados por espaços em branco, incluindo fonetização, como "d+" ("demais"), *hashtag*, *cashtag*, *at-mention*, *emoticon* e URL. Por outro lado, a ferramenta separou *tokens* ortográficos únicos que correspondem a várias palavras (sintáticas), como clíticos, contrações (canônicas e não-canônicas), sinais de pontuação (exceto abreviações), taxas de avaliação das ações na bolsa e valores monetários com ortografia não-convencional. Após a revisão manual, o *corpus* totalizou 81.037 *tokens*.

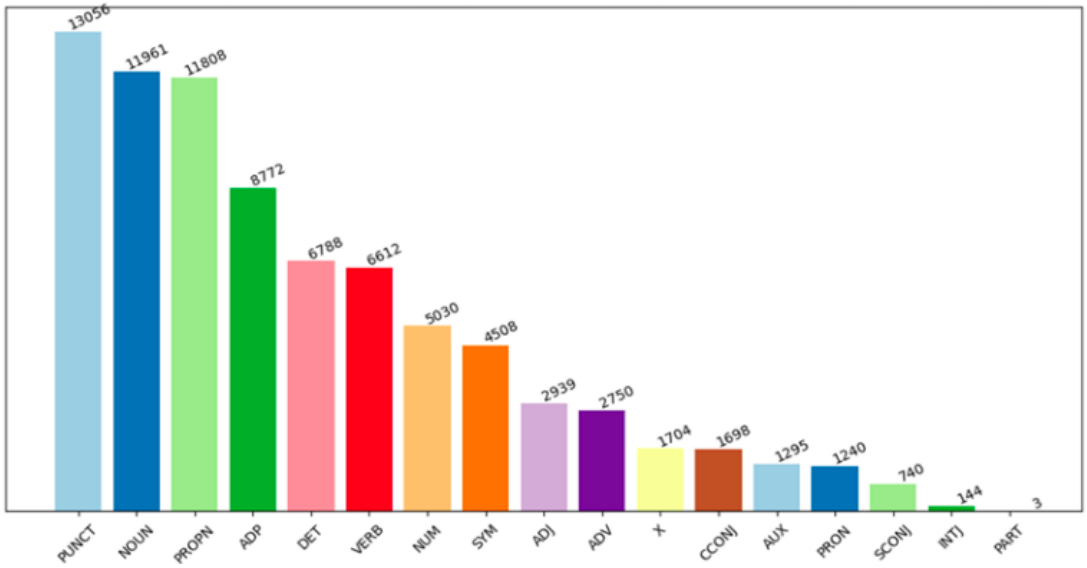
c) Anotação das informações morfológicas

Entre as informações morfológicas do modelo, as *tags* PoS foram as primeiras a serem anotadas. Em um processo semiautomático (Silva; Pardo et al., 2021), o *corpus* foi submetido ao *parser* UDPipe 2 (Straka, 2018), treinado incrementalmente sobre o *corpus* UD-Portuguese Bosque (Rademaker et al., 2017) e *tweets* do DANTEStocks. Os resultados do *parser* foram analisados manualmente por três anotadores com o auxilio de diretrizes para os *tweets* do DANTEStocks Di-Felippo, Postali, Ceregatto, Gazana e Roman (2022) e gerais da língua portuguesa (Duran, 2021). Os casos de discordância foram adjudicados por um linguista sênior. Com base na

[29]Tradução do termo em inglês *syntactic word*, que é definido como a unidade mínima a que corresponde uma função sintática. Na anotação-UD, palavras sintáticas são sinônimas de *tokens*.

distribuição das PoS no *corpus* (Figura 4.5), vê-se que todas as 17 *tags* UD ocorrem. PUNCT é a mais frequente (16%), seguida por NOUN (15%) e PROPN (14%). Juntas, essas 3 etiquetas somam quase metade de todas as *tags* PoS (45%). A partir da anotação PoS padrão-ouro do DANTEStocks, desenvolveu-se o primeiro *tagger* para CGU em português, o Porttagger (Silva et al., 2023), que obteve resultados do estado-da-arte.

Figura 4.5: Frequência de ocorrência das etiquetas PoS no DANTEStocks.



Fonte: (Barbosa, 2024).

De acordo com Di-Felippo, Nunes e Barbosa (2024a), os lemas e os traços (*features*) foram anotados de forma semiautomática a partir do PortiLexicon-UD (Lopes; Duran et al., 2022). Os dados gerados pelo léxico foram revisados manualmente devido à alta taxa de *tokens out-of-vocabulary* (isto é, não previstos no léxico). Quanto aos traços, ressalta-se que, uma vez que a extração das informações do PortLexicon tenha sido guiada pelas *tags* PoS e lemas já validados, o esforço de revisão manual dos traços foi menor, limitando-se à correção de erros decorrentes da ambiguidade dos traços da classe VERB.

d) Anotação sintática

A anotação dos *deprels* no DANTEStocks foi feita em duas etapas semiautomáticas (Di-Felippo; Nunes; Barbosa, 2024a). A primeira criou um *subcorpus* de referência e a segunda etapa ajustou um *parser* pré-treinado para *tweets*, usando o *subcorpus* de referência como parte de seu conjunto de treinamento inicial, e anotou o restante do *corpus*.

Para tanto, os 4.048 *tweets* foram agrupados automaticamente em 3 conjuntos em função do tipo de linguagem/estrutura: linguagem relativamente padrão, padrões estruturais recorrentes

e outros (*tweets* que não pertencem aos outros dois conjuntos). A organização dos *tweets* nos referidos conjuntos permitiu selecionar instâncias de cada um deles para compor um *subcorpus* de referência de 1.000 *tweets*, cobrindo, assim, a diversidade estrutural do DANTEStocks.

O 1.000 *tweets* foram submetidos ao UDPipe 2, treinado sobre o UD-Portuguese Bosque. A anotação gerada pelo *parser* foi posteriormente revisada de forma manual por um único especialista com o auxílio de diretrizes para *tweets* (Di-Felippo; Nunes; Barbosa, 2024b) e para a língua portuguesa no geral (Duran, 2022). Visando consistência na anotação, cada conjunto foi anotado e revisado manualmente em separado, começando pelos *tweets* com linguagem padrão, os quais foram seguidos pelos *tweets* com padrões estruturais recorrentes e, por fim, pelos outros. Ao final, obteve-se um *subcorpus* de referência com anotação padrão-ouro.

O restante do *corpus* foi anotado refinando o Stanza (Qi et al., 2020) para o DANTEStocks. Trata-se de um modelo de *parser* pré-treinado para o português que tem a vantagem de compor um *pipeline* mais amigável para análise de texto. O processo começou com a arquitetura base do Stanza, ajustada no *Porttinari-base* acrescido pelo *subcorpus* de referência de DANTEStocks. O modelo de *parser* resultante do treinamento inicial anotou um novo pacote de dados (proveniente dos 3.048 *tweets*), que foi revisado manualmente e incorporado ao conjunto de dados inicial, sendo então usado para iniciar uma nova execução de treinamento do Stanza. Esse ciclo continuou de forma incremental até que o último pacote de *tweets* tivesse sido anotado/revisado. Os pacotes de *tweets* foram adicionados na mesma ordem aplicada na anotação do *subcorpus* de referência: *tweets* de linguagem padrão, *tweets* de padrões estruturais e *tweets* com propriedades lexicais/estruturais variadas.

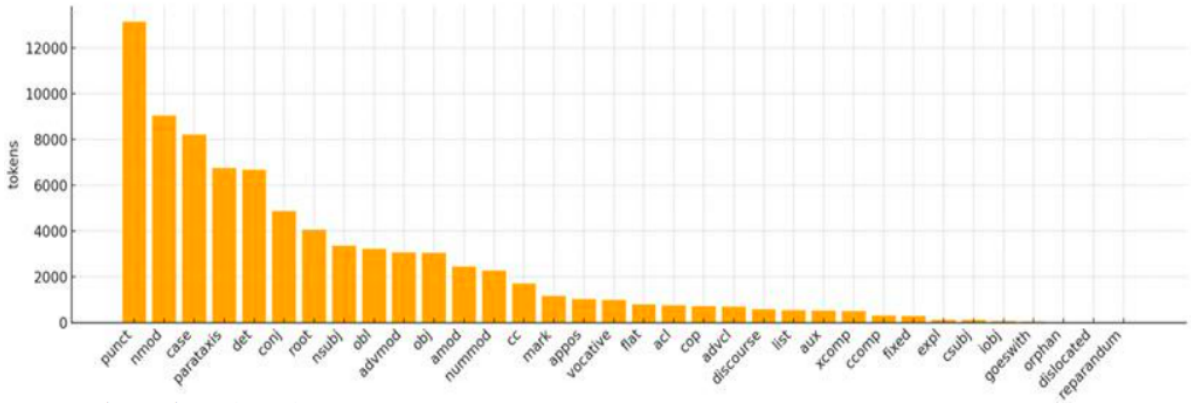
Como todo o processo de revisão foi feito por apenas um anotador, outro especialista em PLN revisou manualmente a anotação automática de 100 *tweets* aleatórios com base nas mesmas diretrizes do primeiro anotador. As árvores de dependência analisadas pelo anotador adicional poderiam ser do *subcorpus* de referência ou geradas pelo Stanza em uma de suas interações. A pontuação do IAA foi calculada usando o coeficiente Kappa (Carletta, 1996) em dois cenários diferentes (Di-Felippo; Nunes; Barbosa, 2024a). No primeiro, avaliou-se a anotação de *head* e *deprel* separadamente. Os resultados do Kappa para *head* e *deprel* foram 0,96 e 0,97, respectivamente. No segundo cenário, a avaliação visou a combinação de *head* e *deprel*, obtendo a pontuação Kappa de 0,95. A IAA por *deprel* foi medida usando a pontuação de concordância total, uma vez que o Kappa não é apropriado devido à distribuição desequilibrada

das relações. Nesse caso, obteve-se concordância total de 100% para mais da metade das 46 diferentes *deprels* (incluindo sub-relações) que ocorrem na amostra de 100 *tweets*.

Assim sendo, esses índices indicam que a anotação sintática-UD do DANTEStocks é útil como padrão-outro. Tanto é que ela já permitiu o desenvolvimento de 2 *parsers*, sendo um deles dedicado a *tweets* (Di-Felippo; Nunes; Barbosa, 2024a) e outro multigênero (Di-Felippo; Roman et al., 2024) (jornalístico, científico e *tweet*). Ambos com desempenho do estado-da-arte.

A Figura 4.6 ilustra a distribuição geral das relações de dependência no DANTEStocks. Ao final, esse treinamento gerou o UGC Parser (Barbosa, 2024), que é o primeiro do tipo voltado para CGU em português.

Figura 4.6: Frequência das *deprels* no DANTEStocks.



Fonte: (Barbosa, 2024).

A respeito da frequência das relações de dependência do modelo UD no DANTEStocks exibidas na Figura 4.6, observa-se que 34 das 37 previstas pelo modelo foram empregadas na anotação do corpus. As relações não empregadas foram **clf**, **compound** e **dep**. Das 34, ressalta-se que **parataxis** é a quarta mais frequente, com 6.733 ocorrências. Isso comprova aquilo que Di-Felippo, Postali, Ceregatto, Gazana, Silva et al. (2021) já tinham observado sobre a alta frequência de *tweets* fragmentados, uma vez que a **parataxis** se estabelece entre dois elementos que poderiam ter relação sintática entre si, porém essa relação não está explicitada.

4.2.3 Primeira anotação de Entidades Nomeadas

A anotação de ENs de Zerbinati, Roman e Di-Felippo (2024) foi manual e conduzida por apenas um anotador de forma linear, ou seja, *tweet* por *tweet*. Para tanto, elaborou-se um manual que se baseia nas categorias e em quase todas as diretrizes de anotação do Segundo HAREM (Zerbinati; Roman, 2023). Mais precisamente, os autores utilizam apenas as 10 categorias da taxono-

mia (ABSTRAÇÃO, ACONTECIMENTO, COISA, LOCAL, OBRA, ORGANIZAÇÃO, PESSOA, TEMPO, VALOR e OUTRO), desconsiderando os tipos e subtipos (Figura 4.7).

Figura 4.7: Categorias do Segundo HAREM na primeira anotação do DANTEStocks.

Categorias	Tipos	Subtipos
ABSTRACCAO (5)	DISCIPLINA	
	ESTADO	
	IDEIA	
	NOME	
	OUTRO	
ACONTECIMENTO (4)	EFEMERIDE	
	EVENTO	
	ORGANIZADO	
	OUTRO	
COISA (5)	CLASSE	
	MEMBROCLASSE	
	OBJECTO	
	SUBSTANCIA	
	OUTRO	
LOCAL (4)	FISICO (7)	ILHA, AGUACURSO, PLANETA, REGIAO, RELEVO, AGUAMASSA, OUTRO
	HUMANO (6)	RUA, PAIS, DIVISAO, REGIAO, CONSTRUCAO, OUTRO
	VIRTUAL (4)	COMSOCIAL, SITIO, OBRA, OUTRO
	OUTRO	
OBRA (4)	ARTE	
	PLANO	
	REPRODUZIDA	
	OUTRO	
ORGANIZACAO (4)	ADMINISTRACAO	
	EMPRESA	
	INSTITUICAO	
	OUTRO	
PESSOA (8)	CARGO	
	GRUPOCARGO	
	GRUPOIND	
	GRUPOMEMBRO	
	INDIVIDUAL	
	MEMBRO	
	POVO	
	OUTRO	
TEMPO (5)	DURACAO	
	FREQUENCIA	
	GENERICO	
	TEMPO_CALEND (4)	HORA, INTERVALO, DATA, OUTRO
VALOR (4)	CLASSIFICACAO	
	MOEDA	
	QUANTIDADE	
	OUTRO	
OUTRO (1)		

Fonte: Mota e Santos (2008).

A opção por classificar as entidades apenas no nível de categoria, que é o nível mais amplo ou genérico da taxonomia do HAREM, foi motivada pelos benefícios tanto para a eficiência do processo de anotação quanto para a aplicabilidade de modelos de AM. Diz-se isso porque categorias genéricas (i) simplificam a anotação, reduzindo ambiguidades e o tempo empregado pelos anotadores, (ii) melhoram a eficiência do treinamento de modelos de AM, uma vez que eles têm menos classes a distinguir, e (iii) permitem maior flexibilidade e aplicabilidade de mo-

delos de AM, pois modelos treinados em um domínio podem ser mais facilmente transferidos para outros domínios, sem necessidade de ajustes complexos.

A anotação de Zerbinati, Roman e Di-Felippo (2024) seguiu a seguinte diretriz do Segundo HAREM: a classificação de cada EN está baseada na sua interpretação em contexto. No HAREM, esse princípio implica que os casos de ambiguidade devem ser resolvidos com a escolha de uma única categoria sempre que o contexto permitir a desambiguação. A anotação com múltiplas categorias era um recurso reservado especificamente para os casos de vagueza, onde o contexto não favorece uma única interpretação e múltiplas leituras são igualmente válidas.

Ao anotar o DANTEStocks, os autores aplicaram a regra da desambiguação contextual, mas optaram por não utilizar a anotação múltipla para os casos de vagueza. Em vez disso, adotaram uma abordagem estritamente *single-label*, na qual o anotador deveria sempre eleger a categoria que julgasse mais proeminente no contexto, mesmo diante de casos potencialmente vagos.

Para os processos de delimitação e classificação das ENs não foi utilizado uma ferramenta de anotação específica. No caso do DANTEStocks, o anotador utilizou um editor de texto genérico e inseriu a anotação de EN na coluna 10 de cada arquivo CoNLL-U, referente a cada *tweet* anotado segundo o modelo UD, seguindo o padrão BIOES (Figura 4.8). Como ilustrado na Figura 4.9 essa abordagem integrou a nova camada de anotação de ENs ao arquivo que já continha as informações morfossintáticas do modelo UD.

Figura 4.8: Anotação de ENs no formato BIOES do DANTEStocks.

Fonte: Zerbinati e Roman (2023).

Embora Zerbinati, Roman e Di-Felippo (2024) não mencionem explicitamente a razão pela inserção das ENs diretamente no arquivo CoNLL-U, acredita-se que isso tenha sido feito pela motivação que levou os autores a conduzir a anotação de ENs, que foi investigar a relevância das informações de PoS para a classificação automática de ENs. Para isso, ter os dois tipos de anotação em um mesmo arquivo facilita essa investigação de diversas formas, sobretudo do ponto de vista do processamento dos dados. Além disso, vale ressaltar que, embora as ENs tenham sido inseridas na versão do *corpus* que continha apenas as etiquetas de PoS, o

Figura 4.9: Anotação de EN no formato BIOES adicionada ao arquivo CoNLL-U.

sent_id	dante_01_457299748639477760l								
text	Olá, a bolsa de valores hoje caiu com as ações PETR4								
Id	Form	Lemma	Upos Tag	Xpos Tag	Feats	Head	DepRel	Deps	Misc
1	Olá	olá	INTJ	_	_	_	_	_	SpaceAfter=No
2	,	,	PUNCT	_	_	_	_	_	_
3	a	o	DET	_	_	_	_	_	_
4	bolsa	bolsa	PROPN	_	_	_	_	_	ENTIDADE=B-ORGANIZAÇÃO
5	de	de	ADP	_	_	_	_	_	ENTIDADE=I-ORGANIZAÇÃO
6	valores	valor	NOUN	_	_	_	_	_	ENTIDADE=E-ORGANIZAÇÃO
7	hoje	hoje	ADV	_	_	_	_	_	ENTIDADE=S-TEMPO
8	caiu	cair	VERB	_	_	_	_	_	_
9	com	com	ADP	_	_	_	_	_	_
10	as	o	DET	_	_	_	_	_	_
11	ações	ação	NOUN	_	_	_	_	_	_
12	PETR4	PETR4	PROPN	_	_	_	_	_	ENTIDADE=S-COISA|SpaceAfter=No

Fonte: Zerbinati e Roman (2023).

DANTEStocks já possui atualmente os traços morfológicos e as relações de dependência do referido modelo gramatical.

A anotação resultou na identificação de 23,453 *tokens* como EN, o que equivale a mais de 28% do total de *tokens* do DANTEStocks. Uma possível explicação para isso é a elevada frequência de ocorrência dos *tickers*, que resulta, por sua vez, do fato de os *tweets* terem sido compilados com base na ocorrência de ao menos um código de uma ação da Ibovespa. Isso faz com que praticamente todo *tweet* tenha ao menos um *ticker* e, consequentemente, uma EN. Os exemplos de (6) a (11) ilustram esse fato. A proeminência dos *tickers*, aliás, também pode justificar parcialmente a alta frequência da PoS PROPN (14% de todas as *tags* do *corpus*), uma vez que eles foram etiquetados como tal.

Além disso, na caracterização linguística do DANTEStocks feita por Zerbinati, Roman e Di-Felippo (2024) com base na anotação de ENs, as categorias COISA e VALOR são as mais representativas, com 45.03% e 22,29% de frequência, respectivamente. A alta frequência de COISA pode mais uma vez ser justificada pela ocorrência dos *ticker*, pois, como mencionado, eles foram etiquetados como tal porque a categoria COISA engloba entidades que não têm nome individual, mas que são designadas pelo nome da classe a que pertencem, como é o caso dos *tickers*. A frequência proeminente de VALOR também está relacionada aos *tickers*, uma vez que eles tendem a ocorrer acompanhados de seus respectivos valores na bolsa.

Embora a anotação tenha sido realizada as boas práticas de anotação de *corpus*, sua condução a partir de uma única perspectiva introduz o risco de que as diretrizes aplicadas não sejam suficientemente robustas ou totalmente otimizadas do ponto de vista linguístico para o

domínio e gênero, o que representa uma limitação metodológica, mas que, ao mesmo tempo, abre espaço para o refinamento.

Além disso, a adoção das 10 categorias do HAREM, embora vantajosa porque simplifica o processo de anotação manual (tornando-o mais rápido e intuitivo) e aumenta a capacidade de generalização dos modelos desenvolvidos a partir delas, apresenta como desvantagem a perda de informatividade, o que diminui a utilidade de um modelo treinado a partir delas em aplicações de PLN para o domínio especializado do mercado financeiro. Diz-se isso porque anotar os ativos financeiros (ou ações) apenas com o rótulo COISA pode ser interessante para o desenvolvimento de um sistema geral de REN, mas, para domínios especializados como o do mercado financeiro, uma anotação mais detalhada é essencial para melhorar a precisão das aplicações e caracterizar mais adequadamente o gênero/domínio.

Capítulo 5

Metodologia

Neste Capítulo, descreve-se o processo de anotação de ENs no *tweebank* DANTEStocks realizado no âmbito deste trabalho. Embora uma anotação de ENs já existisse para o referido *corpus* (Zerbinati; Roman; Di-Felippo, 2024), ela se restringe às 10 categorias gerais da taxonomia do Segundo HAREM. Assim, a anotação descrita neste trabalho, realizada de forma independente à primeira, refina as diretrizes para o emprego das mesmas 10 categorias por meio de decisões linguisticamente motivadas, e expande a categorização ao empregar uma coleção de tipos adaptada do HAREM para atender às particularidades do *corpus*/domínio. Mais precisamente, detalham-se a seguir (i) o conjunto de etiquetas, (ii) o esquema de anotação, (iii) os critérios de delimitação e classificação da ENs, considerando as particularidades estruturais e lexicais dos *tweets*, e (iv) o formato de anotação.

5.1 Conjunto de etiquetas ou *tagset*

Para a anotação das ENs no DANTEStocks, optou-se pela taxonomia do Segundo HAREM (Mota; Santos, 2008) por duas razões estratégicas. A primeira é o fato dessa taxonomia ser um padrão consolidado no PLN para o português, como demonstra seu uso e adaptação em trabalhos como os de Souza, Nogueira e Lotufo (2020) e Costa (2023), o que garante o diálogo deste trabalho com a literatura da área. A segunda razão para a escolha foi a possibilidade de comparação da tarefa ora descrita na Seção 4.2.3 com a primeira anotação de ENs no DANTEStocks (Zerbinati; Roman; Di-Felippo, 2024).

Diferentemente da primeira anotação, pautada exclusivamente nas 10 categorias gerais, empregou-se aqui o elenco de tipos do Segundo HAREM. A inclusão dos tipos foi feita porque,

embora o emprego de categorias (genéricas) tenha os benefícios já citados de simplicidade e generalização, ele também apresenta limitações.

Uma delas é a perda de especificidade, pois categorias amplas não capturam nuances importantes entre diferentes tipos de entidades. Outras limitações são a (i) redução da precisão de modelos de REN baseados em categorias genéricas em domínios especializados e a (ii) dificuldade de expansão posterior, pois, se for necessário detalhar mais a taxonomia no futuro, a reanotação do *corpus* pode exigir tempo e esforço adicionais. Por fim, cita-se também a limitação da aplicação de modelos de REN baseados em categorias genéricas em tarefas que exigem maior precisão semântica, como extração de informações, em que categorias mais detalhadas podem fornecer respostas mais relevantes.

Reconhece-se, no entanto, que o conjunto de etiquetas do Segundo HAREM, mesmo com a inclusão dos tipos, é inerentemente genérico, pois foi concebido para a anotação de nomes próprios em um *corpus* de língua geral. Ao se trabalhar com domínios especializados, a literatura aponta para diferentes caminhos. Uma possibilidade é a criação de uma taxonomia totalmente específica, como fizeram Freitas, Souza et al. (2023) para o domínio do petróleo (PetroNer) e Oliveira et al. (2022) para o domínio clínico (SemClinBr). Outra abordagem, mais híbrida, consiste em mesclar categorias genéricas com categorias de domínio, como fizeram Araujo et al. (2018) no LeNER-Br e, de forma similar, Albuquerque, Costa et al. (2022) no Ulysses-NER; ambos, aliás, utilizaram as categorias do HAREM.

A decisão metodológica deste trabalho seguiu um terceiro caminho, que foi o de empregar um *tagset* consolidado e que garantisse a comparação com a anotação anterior do *corpus* e propor um refinamento e expansão de seus tipos em função do domínio.

Com isso, buscou-se definir um *tagset* que fosse ao mesmo tempo abrangente e informativo, contendo, então, categorias e tipos. A Figura 5.1 resume a taxonomia usada neste trabalho (para comparação, a taxonomia original está apresentada na Figura 4.7).

Com base na Figura 5.1, destaca-se que, aos tipos originais do Segundo HAREM, foram adicionados 4 novos tipos, os quais estão em negrito. Os tipos *certificado*, *ticker* e *indicador* foram propostos para classificar entidades da categoria COISA que são próprias do domínio do mercado financeiro. Por sua vez, o tipo *usuário* tem por objetivo distinguir entidades da categoria PESSOA que são específicas do gênero *tweet*. Tomando como base a anotação de ENs feita em *corpora* como o SemClinBr (Oliveira et al., 2022) e PetroNer (Freitas; Souza et al.,

Figura 5.1: Adaptação das classes e tipos do HAREM ao DANTEStocks.

ABSTRAÇÃO	ACONTECIMENTO	COISA	LOCAL	OBRA
Disciplina Estado Ideia Nome Outro	Efeméride Organizado Evento Outro	Objeto Classe MembroClasse Substância **Certificado** **Indicador** **Ticker** Outro	Físico Humano Virtual Outro	Arte Reproduzida Plano Outro
ORGANIZAÇÃO	**PESSOA**	**TEMPO**	**VALOR**	**OUTRO**
Administração Empresa Instituição Outro	Cargo GrupoCargo Individual GrupoInd Membro GrupoMembro Povo **Usuário** Outro	Duração Frequência Genérico TempoCalend Outro	Classificação Quantidade Moeda Outro	

Fonte: A autora, 2025.

2023), os tipos aqui propostos foram definidos com o auxílio de especialistas de domínio, que apontaram a relevância de se distinguir esses tipos dos demais das categorias.

Em especial, no que se refere à categoria COISA, a introdução dos novos tipos teve como objetivo ampliar a granularidade da anotação, complementando os tipos já existentes, *objeto, classe, membro_classe, substância* e *outro*. O tipo *ticker* contempla os códigos alfanuméricos das ações, que podem ocorrer sem ou com os símbolos de *hashtag* e *cashtag* (p.ex.: "PETR4", "#PETR4" e "$PETR4"). O tipo *indicador*, por sua vez, engloba índices acionários como "Ibovespa" e "S&P 500", além de métricas financeiras, como o "P/L" (Preço/Lucro).

Já o tipo *certificado* abarca títulos, certificações profissionais ou instrumentos financeiros registrados oficialmente, que representam direitos ou participação em ativos, como ações, cotas, títulos de dívida ou certificados de depósito. Ele corresponde a referências a ativos que não sejam via *tickers*, como pela combinação do nome da empresa e a classe da ação (p.ex.: "Eletrobrás ON"), menção a segmentos de listagem (p.ex.: "ADR", do inglês *American Depositary Receipt*) ou referência a credenciais importantes no setor (como "MBA").

A categoria PESSOA, que engloba os tipos *cargo, grupo_cargo, individual, grupo_ind, membro, grupo_membro, povo* e *outro*, foi refinada pela inserção do tipo *usuário*, empregado para etiquetar as menções (aos perfis de usuários) típicas do gênero *tweet*. Vale ressaltar que as menções constituem elementos metalinguísticos nos *tweets* (Di-Felippo; Postali; Ceregatto; Gazana; Silva et al., 2021) e o tratamento dado a elas varia conforme as decisões de projeto. Em alguns *corpora*, as menções são removidas de uma anotação de ENs, como observado por Topçu

e Durgar El-Kahlout (2021), e, em outros, elas são classificadas genericamente como PESSOA (Ritter et al., 2011) ou como potencialmente pertencentes a qualquer categoria, (Derczynski; Bontcheva; Roberts, 2016). Nesse último caso, os autores buscam categorias as menções em função das entidades do mundo extralinguísticos que elas representam.

A diretriz para a anotação de menções parte de uma abordagem na qual se busca, sempre que possível, categorizar o perfil com base na entidade do mundo que ele representa. Assim, seguindo a taxonomia da Figura 5.1, perfis que representam empresas, como "@petrobras", são classificados como ORGANIZAÇÃO-*empresa*, enquanto perfis de canais de comunicação, a exemplo de "@UOLEconomia", são anotados como LOCAL-*virtual*. Já menções que não correspondem claramente a uma empresa ou um canal midiático são categorizadas como PESSOA-*usuário*. Isso porque nem sempre é possível tipificar o perfil de forma inequívoca, mesmo após análise na plataforma. Este é o caso de "@chrinvestor", cuja distinção entre PESSOA-*individual* ou PESSOA-*grupo_ind* é inviável. Assim, para não ter de diferenciar entre perfis individuais e coletivos e padronizar a anotação, definiu-se o tipo *usuário*.

Além de expandir a taxonomia, refinaram-se algumas diretrizes de classificação em função do domínio financeiro. A primeira delas foi classificar órgãos de gestão (p.ex.: conselhos e comitês) como ORGANIZAÇÃO-*administração*. A segunda foi diferenciar entidade permanente de evento temporário. Por isso, termos como "AGO" (Assembleia Geral Ordinária) foram classificados como ACONTECIMENTO-*evento*, pois designam um acontecimento pontual, e não uma estrutura organizacional contínua. A diretriz original do HAREM é a de anotar setores internos de uma organização (como comissões, comitês e assembleias gerais) com o mesmo tipo da organização principal. A terceira foi a de padronizar a anotação das URLs como LOCAL-*virtual*. Este princípio se baseia na função que elas desempenham nos *tweets*: servir como um elo para fontes externas de informação (notícias, relatórios) que contextualizam e ajudam a interpretar outras entidades no texto.

A Tabela 5.1 apresenta o conjunto final de categorias e tipos identificados no DANTEStocks, cada um acompanhado de sua respectiva descrição e exemplos. É importante notar que a tabela inclui exclusivamente os tipos que tiveram ocorrência no *corpus*. Por essa razão, os tipos OBRA-*arte* e COISA-*substância* foram omitidos. Exemplos para estes podem ser consultados no Exemplário do Segundo HAREM (Carvalho; Freitas, 2008). Da mesma forma, a categoria OUTRO, prevista para entidades não contempladas pelas demais classes, não precisou ser utilizada e, portanto, também não consta na tabela.

Tabela 5.1: Taxonomia do HAREM adaptada para tweets do mercado financeiro.

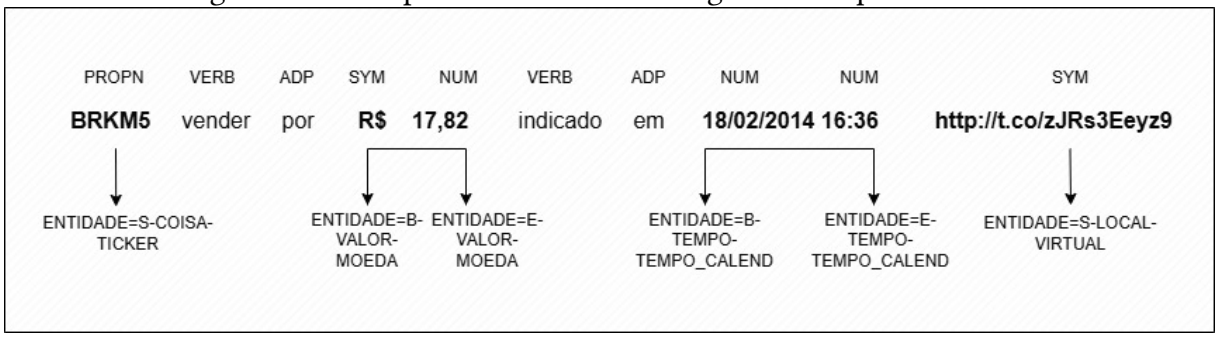
Categoria	Tipo	Descrição	Exemplos
Abstração	Disciplina	Práticas e estratégias de investimento	Análise #Ichimoku, Daytrade, Streaddle
	Estado	Estados físicos e condições	VACA LOUCA, mal da vaca louca
	Ideia	Conceitos e princípios do Mercado de ações	Mercado, mão invisível
	Nome	Objetos linguísticos	Graça
Acontecimento	Efeméride	Acontecimento único em seu contexto histórico	#Lavajato, Pasadenagate
	Organizado	Evento programado	Copa, carnaval
	Evento	Ocorrência pontual	Reunião do Conselho de Administração, ago/e
Coisa	Classe	População de objetos	candlesticks, Valemax, Boeing
	Membroclasse	Instância de classe	PDF
	Objeto	Padrões gráficos específicos	OCO, OCOI, shooting star
	Certificado	Ativos, cotas, certificados	MBA, Eletrobrás ON, bbas-nm
	Indicador	Índices, indicadores e métricas	IBOV, S&P500, Ibovespa/VPa
	Ticker	Códigos de ativos	ESTC3, PSSA3, VALE5, PETR4
Local	Físico	Local geográfico	Reserva da Cantareira, Bacia de Santos
	Humano	Local político/construído	Brasil, São Paulo, Refinaria Pasadena
	Virtual	Endereço ou canal virtual	@JornalOGlobo, http://t.co/LJluyesRk5
Obra	Reproduzida	Obra com várias cópias	Formulário 20-F, Relatório Anual 2013,
	Plano	Plano/Medida Oficial	Plano de demissao voluntária, MP 627
Organização	Administração	Órgão Gestor	Conselho de Ministros, C.a., Conselho da Petrobras
	Empresa	ORG com fins lucrativos	Cielo, Bando do Brasil, Bovespa
	Instituição	ORG sem fins lucrativos	PT, Organização Mundial de Saúde, PMDB
Pessoa	Cargo	Posto ocupado por uma pessoa	CEO da Vale, Vice do Bradesco, Presidente da República
	Grupocargo	Indivíduos representados por um cargo	ministros da Justiça
	Individual	Indivíduo específico	Ministra Rosa Weber, Graça Foster
	Grupoind	Indivíduos sem um nome fixo como grupo	governo Dilma, Gov FHC
	Membro	Indivíduo mencionado pela organização	ex-Itaú, ex-TAM
	Grupomembro	Conjunto de pessoas como membros de uma organização	BTG, Povo Brasileiro
	Povo	Grupo personificado para ações coletivas	China, Rússia
	Usuário	Perfil de Usuário	@Live_Trade, @PaiRico @frfontanella
Tempo	Duração	Quantificação temporal	20-30 anos, 7 dias
	Frequência	Frequência/Repetição	Todos os dias, poucas vezes
	Genérico	Tempo Genérico	hoje
	Tempo_calend	Datas,horas e intervalos.	28/04/2014, 18/03/2014 15:23, INTRADAY
Valor	Classificação	Ordem/Ranking	15ª, oitavo
	Quantidade	Percentuais e números isolados	+ 3,64 %, 70,2 milhões, 4
	Moeda	Valores monetários	R$ 52,38, R$ 14,81, dez reais

5.2 Formato ou marcação de anotação

Como mencionado, a anotação de ENs é dividida em duas etapas: (i) identificação (e delimitação) da expressão linguística correspondente à EN e (ii) classificação dessa expressão, atribuindo-lhe categoria/tipo. No que tange ao formato ou marcação para a delimitação das ENs, optou-se pelo BIOES, o mesmo emprego na primeira anotação do *corpus*. O formato BIOES estende o BIO ao adicionar as etiquetas "E" (fim de uma entidade multipalavra) e "S" (entidade de *token* único), conforme visto na Seção 3.3.

Com base nesse formato, a delimitação das ENs anotadas no *tweet* ilustrado na Figura 5.2 pode ser assim descrita. O *ticker* "BRKM5" e a URL "http://t.co/zJRs3Eeyz9" são entidades expressas por *tokens* únicos, uma vez que foram anotados por "S". O símbolo "R$" é marcado como "B", pois inicia uma entidade monetária, enquanto"17,82" recebe "E", indicando o fim dessa entidade. O mesmo pode ser observado para a entidade TEMPO-*tempo_calend*, composta por 2 *tokens*; a data "18/02/2014" recebe "B" e o horário "16:36" recebe "E".

Figura 5.2: Exemplo de *tweet* anotado segundo o esquema BIOES.



Fonte: A autora, 2025.

A anotação foi incorporada ao arquivo CoNLL-U na coluna 5, destinada originalmente a XPOS, isto é, categorias morfossintáticas específicas de uma língua. Como essa informação não foi considerada no projeto POeTiSA e, consequentemente, não anotada no DANTEStocks, optou-se por armazenar as anotações de ENs nessa coluna sem comprometer a estrutura padrão do formato CoNLL-U. Outra possibilidade seria a de Zerbinati, Roman e Di-Felippo (2024), que utilizou a coluna 10. No entanto, tendo em vista que a coluna 10 dos arquivos que compõem a versão do *corpus* anotada neste trabalho já possui informações ditas adicionais ou MISC (para as quais ela se destina), optou-se pela coluna 5. Com isso, a marcação das entidades foi preservada dentro do arquivo CoNLL-U sem interferir em outras informações linguísticas. Um exemplo dessa anotação pode ser visto na Figura 5.3.

Figura 5.3: Anotação BIOES adicionada ao arquivo CoNLL-U.

Id	Form	Lemma	Upos	Xpos
1	BRKM5	BRKM5	PROPN	ENTIDADE=S-COISA-TICKER
2	vender	vender	VERB	-
3	por	por	ADP	-
4	R$	R$	SYM	ENTIDADE=B-VALOR-MOEDA
5	17,82	17,82	NUM	ENTIDADE=E-VALOR-MOEDA
6	indicado	indicar	VERB	-
7	em	em	ADP	-
8	18/02/2014	18/02/2014	NUM	ENTIDADE=B-TEMPO-TEMPO_CALEND
9	16:36	16:36	NUM	ENTIDADE=E-TEMPO-TEMPO_CALEND
10	http://t.co/zJRs3Eeyz9	http://t.co/zJRs3Eeyz9	SYM	ENTIDADE=S-LOCAL-VIRTUAL

Fonte: A autora, 2025.

5.3 Diretrizes de anotação

5.3.1 Diretrizes gerais de segmentação e classificação

1. **Capitalização e variações lexicais:** dada a natureza do *corpus* de CGU DANTEStocks, o critério de obrigatoriedade de ocorrência de uma letra maiúscula para a identificação de uma EN definido pelo HAREM (Santos; Cardoso, 2007; Mota; Santos, 2008) não se aplica. Ao contrário de textos formais em que a capitalização corresponde a um nome próprio, o uso da maiúscula em CGU, principalmente *tweets*, é pouco sistemático. A maiúscula é apenas uma pista, pois muitas ENs no *corpus* não são grafadas segundo a linguagem padrão. Para lidar com essa variação ortográfica e lexical, optou-se por uma abordagem baseada em contexto, que se adapta à natureza ruidosa e criativa do conteúdo gerado por usuários (CGU). Dessa forma, diferentes grafias de um mesmo termo, como “Petrobras”, “Petrobrás”, “#petrobras” e “PETROBRAS”, recebem uma anotação padronizada. Da mesma forma, variações lexicais com viés de humor ou sentimento (p. ex.: “PETROFUMO”, ”Petrobomba”, “PeTebrás”) também são anotadas, desde que o contexto confirme a referência à Petrobras. A mesma regra se estende a abreviações informais como “Petro” e a termos truncados como “Petr”.

2. **Pistas morfossintáticas:** as etiquetas de PoS do modelo UD PROPN (nome próprio) e NUM (numeral) servem como indícios para a identificação de uma EN, mas outras categorias morfossintáticas também podem ser sinalizadoras. Um exemplo é a etiqueta X, que na maioria das vezes indica *hashtags* ou *cashtags* (que funcionam como indexadores do Twitter e não possuem função sintática), as quais são, muitas vezes, *tickers* precedidos pelos símbolos“#” e“$” e, por isso, alvo de anotação.

3. **Tratamento de fenômenos CGU:** fenômenos CGU como *hashtags*, *cashtags*, menções e URLs são considerados na anotação, dada a importância desses elementos linguísticos na caracterização do gênero e do domínio, e classificados conforme a taxonomia e em decorrência de sua interpretação em contexto.

4. **Anotação com rótulo único (*single-label*):** cada EN é anotada com uma única categoria e tipo com base na interpretação da mesma no *tweet*; isso quer dizer que, entre as classificações possíveis de uma EN no *tweet*, o anotador deverá optar por aquela que julga mais adequada ao contexto. Com isso, a anotação adota a abordagem *single-label*.

5. **Anotação de metáforas:** as metáforas são anotadas de acordo com a entidade à qual fazem referência. Quando uma expressão metafórica é usada para representar um ativo ou outro elemento, por exemplo, ela recebe a mesma categoria da entidade referida. Para ilustrar, no *tweet* (14), a expressão "King Kong" é usada metaforicamente para os ativos da Petrobras. A expressão deriva da gíria de mercado "mico" (um mau investimento), e "King Kong" funciona como uma hipérbole para o "maior mico", enfatizando a percepção negativa sobre os ativos. Dessa forma, como os ativos são classificados como COISA-*certificado*, a metáfora "King Kong" recebeu a mesma anotação.

   (14) King Kong me acordem quando bater em 12,50 que tenho interesse ....rsrsr

6. **Tratamento de entidades encaixadas:** no caso de uma entidade encaixada, ou seja, aquela contida em outra entidade maior e que dessa forma permite mais de uma interpretação de delimitação, deverá ser delimitada a entidade maior, assim como é feito no HAREM. Essa decisão tem como principal objetivo evitar a proliferação excessiva de entidades com identificações alternativas. No *tweet* (15) , a expressão "Ex-presidente da Petrobras" pode ser anotada de formas diferentes a depender da granularidade. Pode-se assumir a expressão inteira como uma EN da categoria PESSOA-*cargo* ou decompô-la, considerando a ocorrência de "Ex-presidente" (PESSOA-*cargo*) e "Petrobras" (ORGANIZAÇÃO-*empresa*). No caso, optou-se por delimitar a expressão maior, anotando-a como PESSOA-*cargo*.

   (15) #petr4 RT **Ex-presidente da Petrobras** nomeou primo para estatal nos EUA na época da compra de Pasadena http://t.co/0vRJaMtKH3 #mercados_IM

7. **Delimitação do núcleo da entidade:** a diretriz para a delimitação de Entidades Nomeadas (ENs) deve priorizar o núcleo da entidade, excluindo termos satélites como preposições, determinantes e quantificadores. Seguindo esse princípio, a delimitação das categorias VALOR e TEMPO também exclui modificadores. Por exemplo, modificadores escalares como "mais de", ilustrado no *tweet* (16) com "mais de R$ 1 bilhão", não são anotados. Aceita-se essa exclusão, pois, embora o modificador adicione uma nuance de imprecisão quantitativa, ele não altera a identidade referencial da entidade. O objetivo principal dessa diretriz é evitar a proliferação de anotações não comparáveis entre si. Como é o caso de "mais de 1 bilhão", "menos de 1 bilhão" e "aproximadamente 1 bilhão". Pela mesma lógica, a expressão "dia de" é excluída em casos como "dia de ontem" (*tweet* 17), visto que "ontem" já constitui uma referência temporal completa. Em contrapartida, modificadores temporais relacionais como "depois" na expressão "um ano depois", no *tweet* (18) são considerados partes integrantes da entidade, pois formam uma unidade temporal coesa. Portanto, nesses casos, a expressão completa é anotada como uma única entidade.

   (16) 13 empresas perdem mais de **R$ 1 bilhão** na Bolsa em fevereiro: Petrobras lidera lista com perdas significativas... http://t.co/t3uG25y3gj #infomoney #vale5

   (17) A LIGHT S.A. fechou o dia de **ontem** ao preço de R$ 16,87 (+1,20%) com volume de R$ 26,32 mm. $LIGT3

   (18) @coroneldoblog @o_colecionador problema nao é esse. Problema é o bilhao de dolares (3x valuation pago pela PETR4) **um ano depois**.

5.3.2 Diretrizes específicas de delimitação ou segmentação

1. **Variações de valor**: as variações de valor (valorização ou desvalorização) das ações são compostas por um dos símbolos "+/-" (PoS SYM), seguido de um número (PoS NUM) e do sinal de porcentagem (PoS SYM), como indicado por Scandarolli et al. (2023). Dessa forma, trata-se de entidades multipalavra. No *tweet* já *tokenizado* em (19), há os índices"+ 2,09" e"+ 2," relacionados aos ativos "Petr3" e "Petr4", respectivamente. Esses índices são classificados como entidades da categoria VALOR e do tipo *quantidade*. A anotação deve abranger toda a sequência, desde o símbolo "+" até o símbolo "%", de forma contínua. O primeiro *token* (SYM) é anotado como início da entidade (B-VALOR), o número (NUM) é

anotado como interior da entidade (I-VALOR), e o sinal de porcentagem (SYM) é anotado como final da entidade (E-VALOR), como ilustrado na Figura 5.4.

(19) RT @Ary_AntiPT: kkkk Coro de P**a RT @garimpodeacoes: Quem puxa para valer o IBOVESPA para cima é a Petrobrás. Petr3, **+ 2,09** % e Petr4, **+ 2,** ...

Figura 5.4: Diretriz de segmentação das variações de valor das ações.

23	Petr3	Petr3	PROPN	ENTIDADE=S-COISA-TICKER
24	,	,	PUNCT	_
25	+	+	SYM	ENTIDADE=B-VALOR-QUANTIDADE
26	2,09	2,09	NUM	ENTIDADE=I-VALOR-QUANTIDADE
27	%	%	SYM	ENTIDADE=E-VALOR-QUANTIDADE
28	e	e	CCONJ	_
29	Petr4	Petr4	PROPN	ENTIDADE=S-COISA-TICKER
30	,	,	PUNCT	_
31	+	+	SYM	ENTIDADE=B-VALOR-QUANTIDADE
32	2,	2,	NUM	ENTIDADE=E-VALOR-QUANTIDADE
33	...	...	PUNCT	_

Fonte: A autora, 2025.

2. **Truncamento**: o truncamento ocorre quando uma entidade textual é cortada, o que geralmente é indicado por reticências. A diretriz geral é anotar uma entidade truncada somente se sua categoria e tipo puderem ser determinados, seja pelo contexto, por outros *tweets*, por análise de especialista ou por fontes externas. Em todos os casos, as reticências são sempre excluídas da anotação, e a segmentação da entidade segue o formato BIOES. A aplicação dessa diretriz varia conforme o caso. As URLs truncadas, como no exemplo(20), são anotadas como S-LOCAL-*virtual* sempre que for possível identificá-las como tal. Já para os índices de (des)valorização das ações, como"+ 2," em (19), são anotados apenas se o valor (ainda que parcial) estiver presente; casos em que apenas o símbolo ocorre (sem um número que o acompanhe), como em "+...", não são anotadas. As *hashtags* e menções truncadas que contêm apenas os prefixo ("#..." e "@...", respectivamente) não são anotadas, pois as entidades não podem ser de fato identificadas.

(20) RT ubrals: Me lembra os junk bonds de a PDVSA edmilsonpapo10 #PETR4 fundo de o poço RT Petrobras prepara megacaptação de US$ 12 bi **http://t.c**...

Figura 5.5: Diretriz de anotação de URL truncada.

26	http://t.c	http://t.c	SYM	ENTIDADE=S-LOCAL-VIRTUAL

Fonte: A autora, 2025.

3. **Contração**: No *tweet tokenizado* (21), por exemplo, a contração "neste" foi descontraída para "em" (ADP) e "este" (DET). No arquivo CoNLL-U, ilustrado na Figura 5.6, a contração é mantida na linha 12-13, com indicação de que foi decomposta nos *tokens* correspondentes das linhas 12 e 13. Nos casos em que a contração ocorre no início de uma entidade, a linha correspondente à contração **não** é anotada e a entidade tem início no determinante (excluindo-se a preposição). Se a contração estiver **no interior** de uma EN (isto é, não na posição inicial), toda a sequência resultante da contração é anotada, respeitando os limites da entidade.

   (21) TOV aposta em blue chips, confira 7 ações para comprar **neste mês**: Os ativos que permanecem n... http://t.co/XKwFI1bsEg #infomoney #vale5

Figura 5.6: Exemplo de anotação de EN composta por contração.

12-13	neste	_	_	_
12	em	em	ADP	_
13	este	este	DET	ENTIDADE=B-TEMPO-TEMPO_CALEND
14	mês	mês	NOUN	ENTIDADE=E-TEMPO-TEMPO_CALEND

Fonte: A autora, 2025.

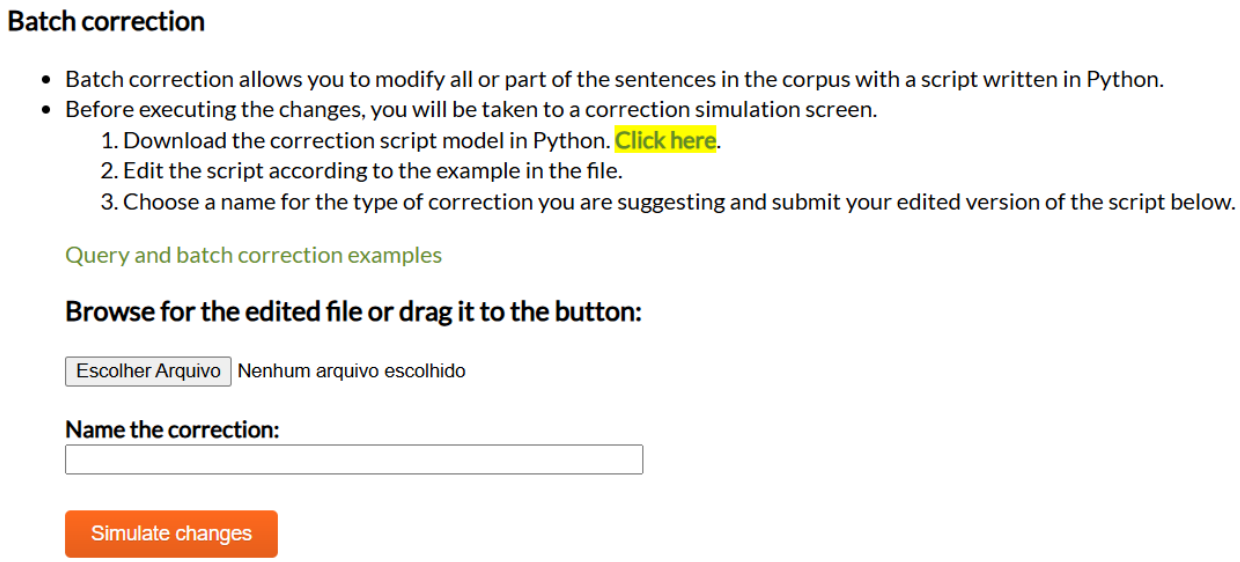
5.4 Método de anotação

A anotação deste trabalho foi conduzida por uma única pesquisadora por meio de um método semi-automático. Este método combina duas abordagens complementares: uma **manual e linear**, na qual a anotação é feita sequencialmente (*tweet* por *tweet*), e outra **automática e transversal**, que aplica regras de forma abrangente para anotar os casos correspondentes em todo o *corpus*.

A primeira fase do trabalho consistiu na anotação exclusivamente manual de uma amostra inicial de 400 *tweets*. O objetivo dessa etapa foi estabelecer as diretrizes de delimitação e classificação das entidades, que serviram de base para a criação de estratégias de anotação automática para o restante do *corpus*. Para essa tarefa, foi utilizado um editor de texto genérico para modificar os arquivos no formato CoNLL-U.

Concluída essa fase inicial, o processo se tornou iterativo e semiautomático. À medida que a anotação manual avançava, eram desenvolvidas e aplicadas regras para automatizar a classificação de fenômenos frequentes e passíveis de generalização. Essa etapa foi realizada com a ferramenta Interrogatório, que é um dos ambientes constitutivos da ET[30], escolhida por suas funcionalidades de busca e de anotação em lote, como ilustra a Figura 5.7. A partir de um *corpus* em formato CoNLL-U, a ferramenta utiliza expressões regulares em *Python* simplificado para executar duas funções principais: (i) buscas em todo o *corpus* e (ii) modificações ou anotações em massa.

Figura 5.7: Ambiente de anotação em lote do Interrogatório.

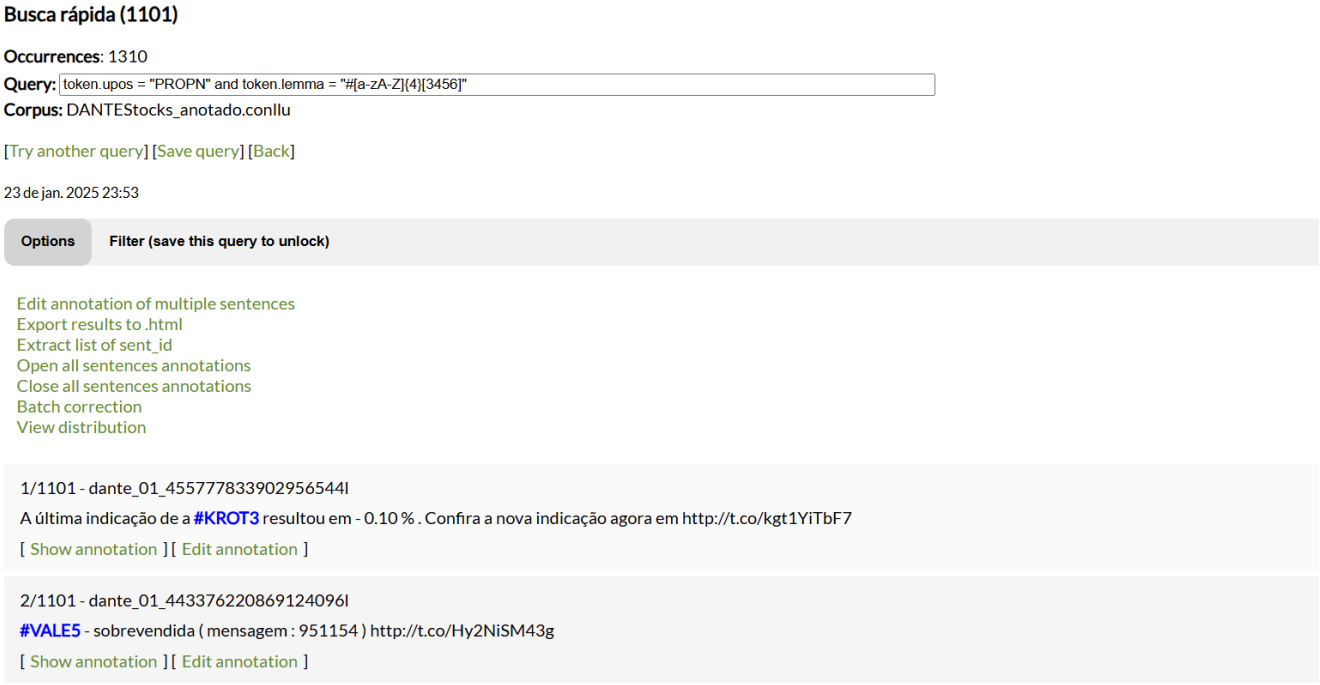


Fonte: A autora, 2025.

Ao total, formularam-se 87 regras distintas. Vale ressaltar que, embora essas regras visassem principalmente a anotação em lote do *corpus*, elas também foram empregadas para alterar anotações manuais anteriores conforme as diretrizes foram sendo definidas. Como o Interrogatório foi desenvolvido para processar *corpora* anotados segundo o modelo gramatical UD (via arquivos CoNLL-U), as informações morfossintáticas ou *tags* PoS puderam ser usadas para formular as regras de anotação.

[30] A ET é um conjunto de ferramentas desenvolvido para apoiar pesquisas linguísticas e tarefas de processamento de língua natural utilizando *corpora* anotados no formato CoNLL-U (Souza; Freitas, 2021).

Figura 5.8: Ambiente de busca da ferramenta Interrogatório.



Fonte: A autora, 2025.

Para anotar os casos de *hashtag* composta por *ticker* (como #KROT3), por exemplo, fez-se inicialmente a busca por todas as ocorrências desse tipo no *corpus*. A expressão regular formulada para isso está descrita na Figura 5.8. Trata-se de: token.upos =“PROPN” and token.lemma =“#[a-zA]4[3456]”. A mesma Figura fornece dois *tweets* de exemplo retornados pela ferramenta. Diante da definição da diretriz de anotar essas *hashtags* como S-COISA, formulou-se a regra apresentada no Algoritmo 2, que foi aplicada a todos os *tweets* retornados pela busca inicial.

Algorithm 2 Exemplo de sintaxe de regra para entidades da categoria S-COISA

```
1: if regex("PROPN", token.upos) and regex("#[a-zA-Z]4[3456]", token.lemma):
2:     token.xpos = "ENTIDADE=S-COISA"
```

Nesse caso, a sintaxe da regra pode ser interpretada formalmente da maneira como indicado no Algoritmo 3. Mais precisamente, as regras foram empregadas principalmente para a anotação automática das ENs dos tipos *ticker*, *hashtag*, *cashtag* e URL. Todas as instâncias anotadas pelo Interrogatório foram posteriormente revisadas de forma manual.

Já para decifrar certos termos da linguagem informal e os termos específicos do mercado financeiro nos *tweets*, recorreu-se também à análise de especialistas e a fontes *online*, pois,

sem esse conhecimento, seria difícil saber, por exemplo, que “CS”, em “CS indicando PETR4 x PETR3”, refere-se à empresa Credit Suisse (ORGANIZAÇÃO).

Algorithm 3 Regra de anotação para entidades da categoria `S-COISA`

Entrada: token com atributos `upos` e `lemma`
if regex(“PROPN”, token.upos) **and** regex(“#[a-zA-Z]4[3456]”, token.lemma) **then**
    Seja o **token** for um nome próprio (PROPN); e
    Seja o **lemma** iniciado pelo símbolo #; e
    contenha exatamente quatro letras (maiúsculas ou minúsculas); e
    seja seguido pelos números 3, 4, 5 ou 6;
    Atribuir anotação à coluna XPOS
    token.xpos ← “ENTIDADE=S-COISA”
end if
Saída: token anotado (se condição satisfeita)

Capítulo 6

Resultados e discussão da anotação

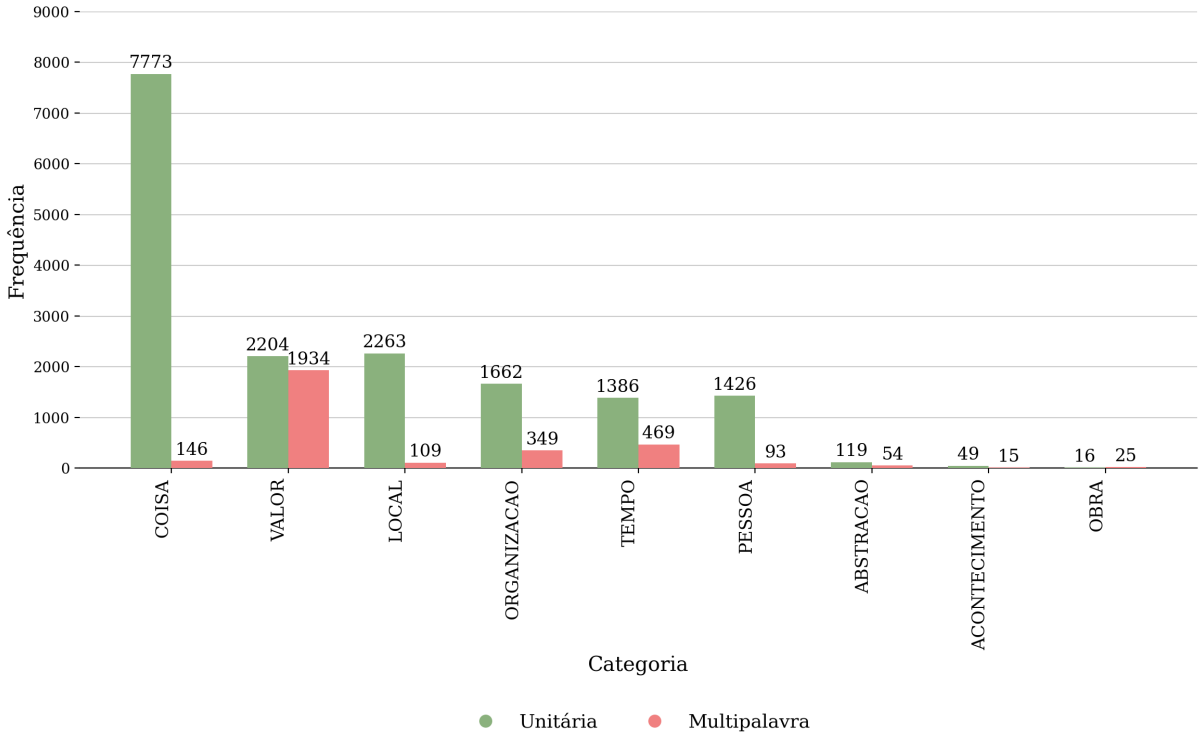
Neste Capítulo, apresentam-se os resultados da anotação de ENs no *corpus* DANTEStocks realizada com base na taxonomia refinada e nas diretrizes metodológicas descritas no Capítulo 5. A análise dos resultados busca fornecer uma caracterização quantitativa e qualitativa do *corpus*, detalhando a distribuição das entidades por categoria, tipo e expressão linguística. Para tanto, ressalta-se que todos os 4.048 *tweets* do *corpus* foram anotados.

Além das 10 categorias gerais, empregou-se o elenco de 43 tipos da taxonomia hierárquica do Segundo HAREM, ao qual foram acrescidos 4 novos tipos específicos para o domínio do mercado financeiro: *certificado*, *indicador* e *ticker* na categoria COISA, e *usuário* na categoria PESSOA. Com isso, o esquema de anotação final totaliza 47 tipos (Figura 5.1).

A expansão da taxonomia teve um impacto direto na complexidade da tarefa. Com a aplicação do esquema BIOES, o número de etiquetas de anotação possíveis saltou para 192. Ao final do processo, foram identificadas 20.092 entidades, que correspondem a 24.825 *tokens* anotados. Esse montante representa aproximadamente 29% do total de 84.396 *tokens* que compõem o *corpus*.

A análise da forma de expressão linguística das 20.092 entidades evidencia uma predominância expressiva de *tokens* unitários (16.898 ocorrências) em comparação às formas multipalavra (3.194 ocorrências). Conforme ilustrado na Figura 6.1, essa característica está presente em quase todas as categorias que ocorrem no DANTEStocks e é um reflexo direto da natureza do *corpus*, que combina fenômenos de redes sociais (*hashtags*, menções e URLs) com elementos do domínio financeiro (*tickers*, *cashtags*) frequentemente expressos por *tokens* unitários. A categoria OBRA representa a única exceção a essa regra geral, ainda que com uma diferença pouco expressiva.

Figura 6.1: Distribuição de *tokens* individuais e multipalavra por categoria.

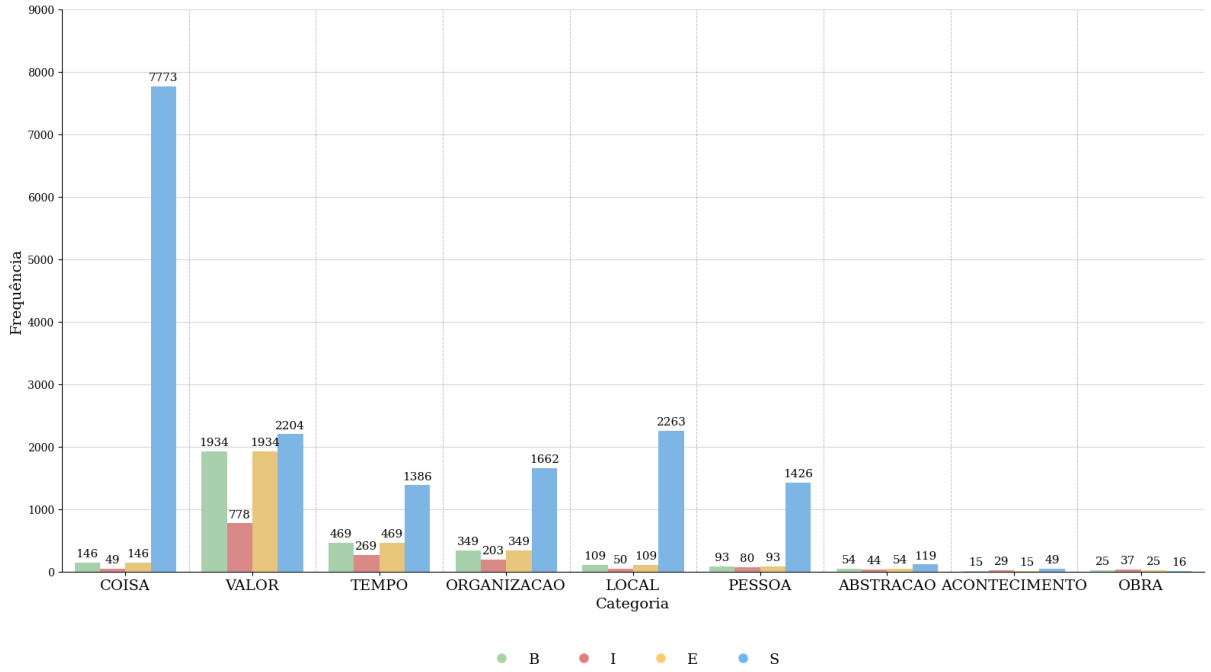


A categoria COISA é o exemplo mais proeminente desse padrão, pois, sendo composta quase que exclusivamente por códigos e siglas, 7.773 do total de 7.919 ocorrências são de *token* unitários. Entre eles estão, por exemplo, *tickers* (p.ex.: "PETR4" e "VALE"), indicadores (p.ex.: "#IBOV", "S&P" e "DJ") e certificados (p.ex.: "elet-n1" e "rent-nm"); todos anotados com "S" no esquema BIOES. Outras categorias, como LOCAL, ORGANIZAÇÃO e PESSOA, também demonstram comportamento similar, com predominância evidente de entidades unitárias.

Na categoria VALOR, entretanto, há certo equilíbrio entre as diferentes formas de expressão das ENs, com 2.204 casos de *token* único e 1.934 de multipalavra. Essa característica se dá principalmente pela ocorrência das indicações de mudança no valor de uma ação, como "-2.44%", *tokenizada* em três elementos (sinal de polaridade, valor numérico e porcentagem), configurando expressão do tipo multipalavra. Considerando a composição das entidades quanto à marcação BIOES (Figura 6.2), observa-se que a distribuição da etiqueta "I" (*Inside* ou "Intermediário") contribui para a caracterização estrutural das entidades. Na categoria VALOR, a distribuição em termos absolutos de *tokens* intermediários se destaca, com 778 ocorrências, refletindo a elevada frequência de entidades multipalavra na categoria.

A categoria PESSOA revela-se, proporcionalmente, como aquela que concentra as entidades nomeadas de maior extensão. Tal inferência decorre da elevada razão entre tokens do tipo

Figura 6.2: Frequência de etiquetas BIOES por categoria



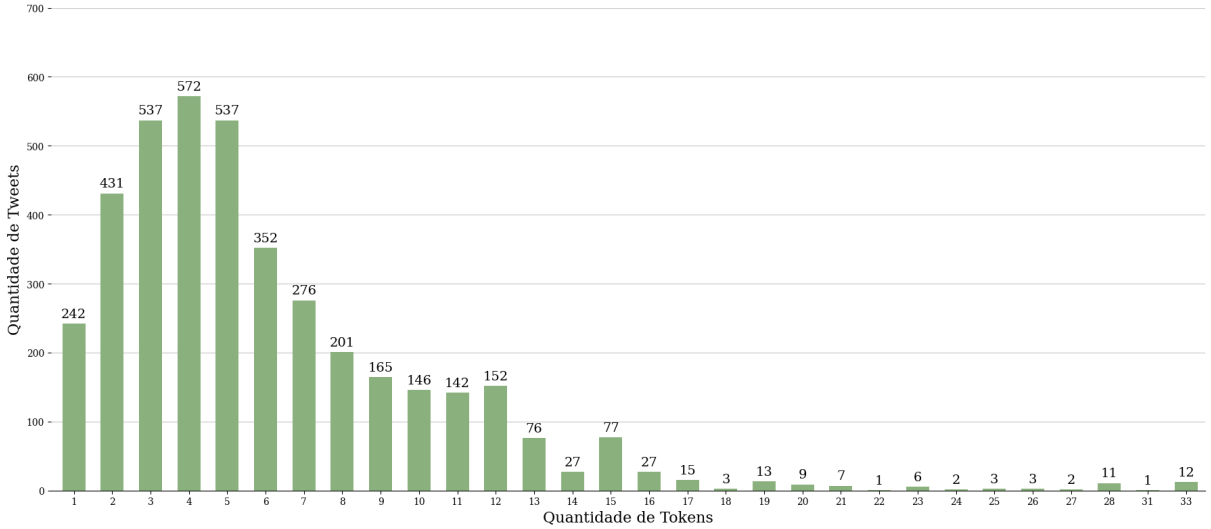
"I" (80) e "B" (93), indicando que, nos casos em que a entidade PESSOA é expressa por múltiplas palavras, ela tende a ultrapassar dois *tokens*. Esse alongamento lexical deve-se, em grande parte, à incorporação de títulos, cargos ou funções ao nome próprio, como em "Ministra Rosa Weber". Padrão análogo de extensão também é observado na categoria ORGANIZAÇÃO, frequentemente composta por denominações corporativas longas, a exemplo de "LPP Empreendimentos e Participações". A categoria TEMPO igualmente apresenta entidades mais extensas, sobretudo quando se referem a expressões temporais compostas pelo dia da semana e do mês (p.e.: "terça-feira, 22 de abril").

Quanto à quantidade de *tokens* anotados como EN por *tweet* (Figura 6.3), observa-se que ela varia de 1 (242 *tweets*) até 33 *tokens* (12 *tweets*). O pico de frequência corresponde a 4 *tokens* por *tweet*, como se observou em 572 *posts*. A ocorrência de 3 e 5 *tokens* também se destaca, com 537 *tweets* cada. Os 12 *tweets* com a maior quantidade de *tokens* anotados como ENs (isto é, 33) são similares ao Exemplo (22). Eles apresentam uma estrutura recorrente composta por data, hora e uma sequência ou lista de *tickers* associados às suas mudanças de valor no mercado.

(22) 05/03/14- 17:20: Maiores Baixas: BRAP4 - 4,50 % R$ 20,35, RSID3 - 4,09 % R$ 1,64, VALE5 - 3,50 % R$ 28,07, CSNA3 - 3,40 % R$ 9,60, BRKM5 - 3,28 % R$ 15,62.

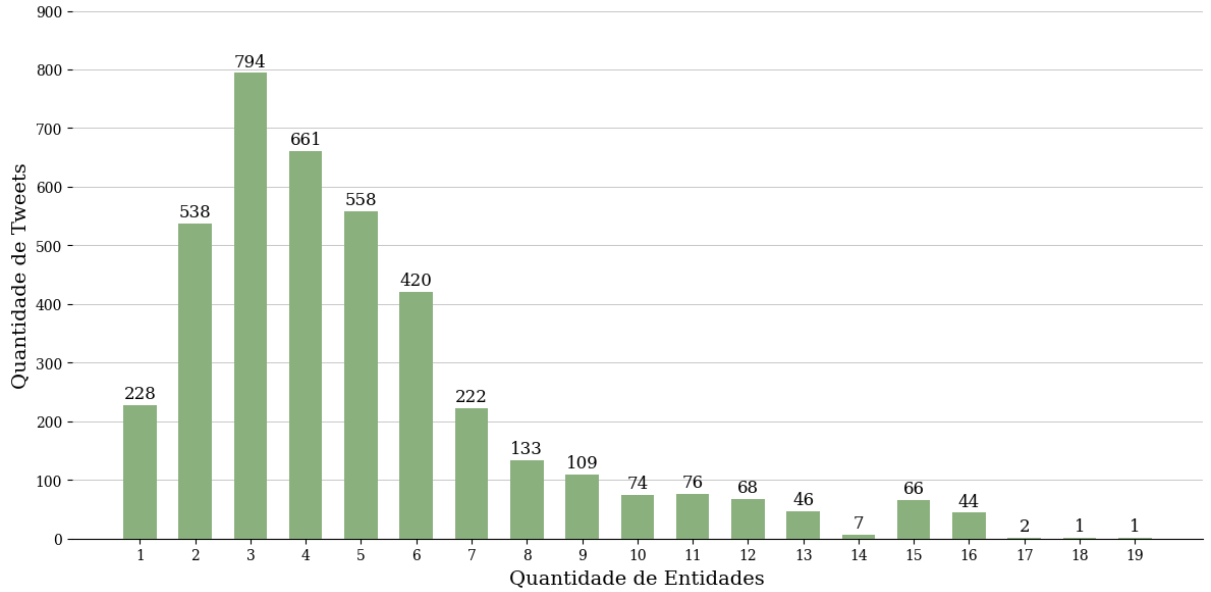
No que diz respeito à quantidade de ENs por *tweet* (Figura 6.4), observa-se que ela varia de 1 (228 *tweets*) até 19 entidades (1 *tweet*). O pico de frequência ocorre com 3 entidades por *tweet*, o

Figura 6.3: Quantidade de *tokens* anotados por *tweet* e a frequência correspondente.



que ocorre em 794 *tweets* diferentes. Os *posts* que contêm 3 entidades comumente apresentam o padrão: *cashtag*, nome da empresa e *ticker* (como "$PETR3 - Petrobras (petr)").

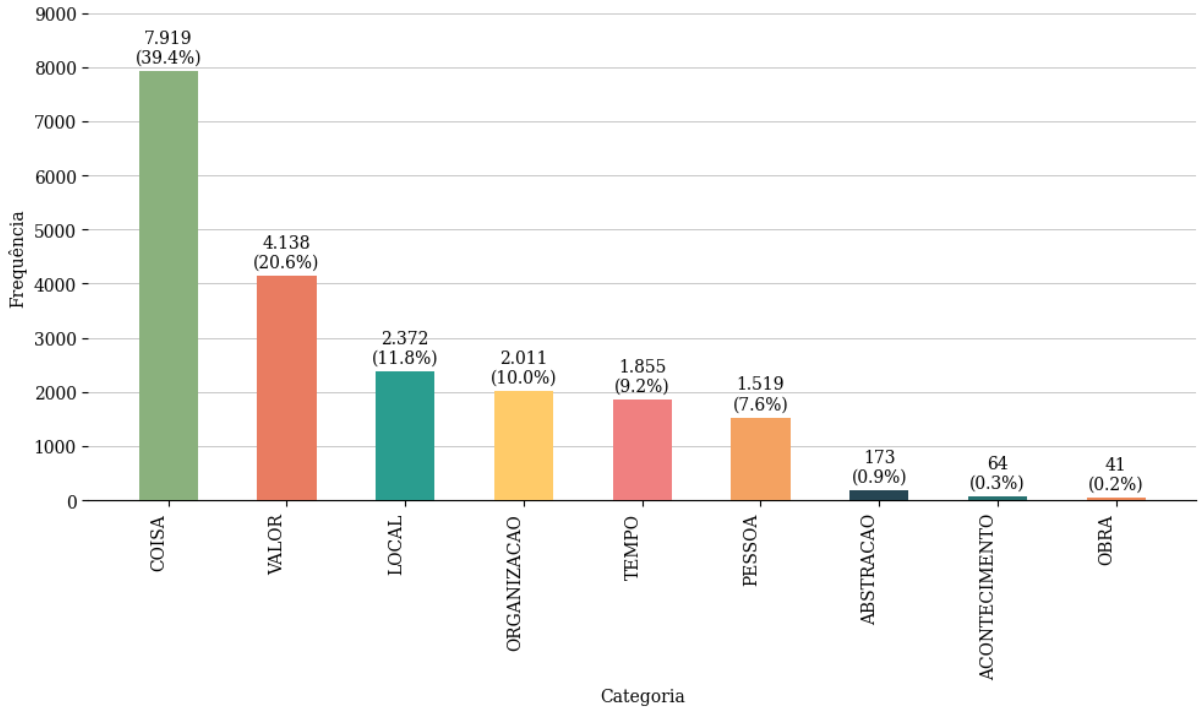
Figura 6.4: Quantidade de entidades por *tweet* e a frequência correspondente.



Quanto à distribuição das 10 categorias, a Figura 6.5 revela um padrão típico de dados do mundo real, em que algumas entidades são numerosas e outras aparecem raramente. Como esperado, a categoria COISA é a mais frequente, com quase 40% de todas as entidades. A segunda mais frequente é VALOR, com 20.6% das entidades. As categorias LOCAL, ORGANIZAÇÃO, TEMPO e PESSOA aparecem na sequência, como porcentagens próximas. Já ABSTRAÇÃO,

ACONTECIMENTO e OBRA são as menos frequentes. Das 10 categorias originais, apenas 9 aparecem no *corpus*, pois OUTRO, a classe restante, não foi necessária.

Figura 6.5: Quantidade de entidades por categoria.



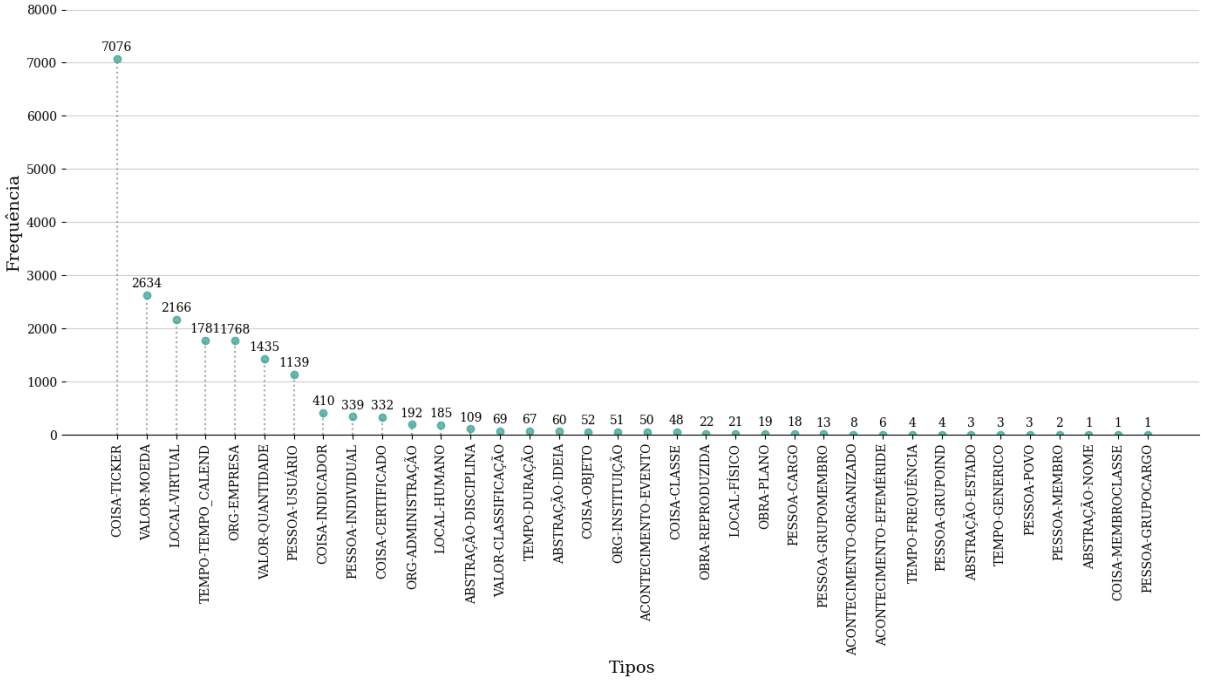
Analisando a distribuição dos tipos (Figura 6.6), vê-se que a categoria COISA é composta principalmente pelos tipos específicos do domínio acrescidos à tipologia original, isto é, *ticker* com 7.076 ocorrências, *indicador* com 410 e *certificado* com 332. Os demais tipos dessa categoria que ocorrem no *corpus* (*classe*, *membro_classe* e *objeto*), considerados mais gerais, possuem frequência mínima de apenas um dígito.

O tipo COISA-*ticker* é de longe o mais frequentes, seguido por VALOR-*moeda* (2.634 ocorrências) e LOCAL-*virtual* (2.166 ocorrências) (que se refere principalmente a URLs). A alta frequência de *ticker* e *moeda* é justificada pela ocorrência das ações comumente associadas a seus valores de mercado, enquanto a do tipo *virtual* justifica-se pela necessidade de compartilhar conteúdo que esteja fora da plataforma, até mesmo devido ao limite de caracteres. Em contraste, tipos mais abstratos, como ABSTRAÇÃO-*ideia* e TEMPO-*genérico*, são raros.

A Tabela 6.1, sistematiza-se a frequência das categorias, tipos e formas de expressão linguística (unitária e multipalavra). Vale ressaltar que os dados são referentes apenas às categorias e tipos que efetivamente ocorreram no *corpus*, isto é, 9 categorias e 36 tipos.

A análise da frequência das classes gramaticais (PoS *tags* da anotação-UD) das ENs anotadas revela uma distribuição concentrada em três categorias, as quais correspondem a mais de 90%

Figura 6.6: Quantidade de entidades por tipos.



das ocorrências. São elas: nome próprio (PROPN, 56.8%), numeral (NUM, 24.4%) e símbolo (SYM, 20.9%). Isso evidencia que as ENs do *corpus* são, prototipicamente, da categoria PROPN, sobretudo porque *tickers* e menções foram anotados como tal. A alta frequência de NUM e SYM advém do caráter informativo do mercado financeiro.

Figura 6.7: Quantidade de POS tags anotadas.

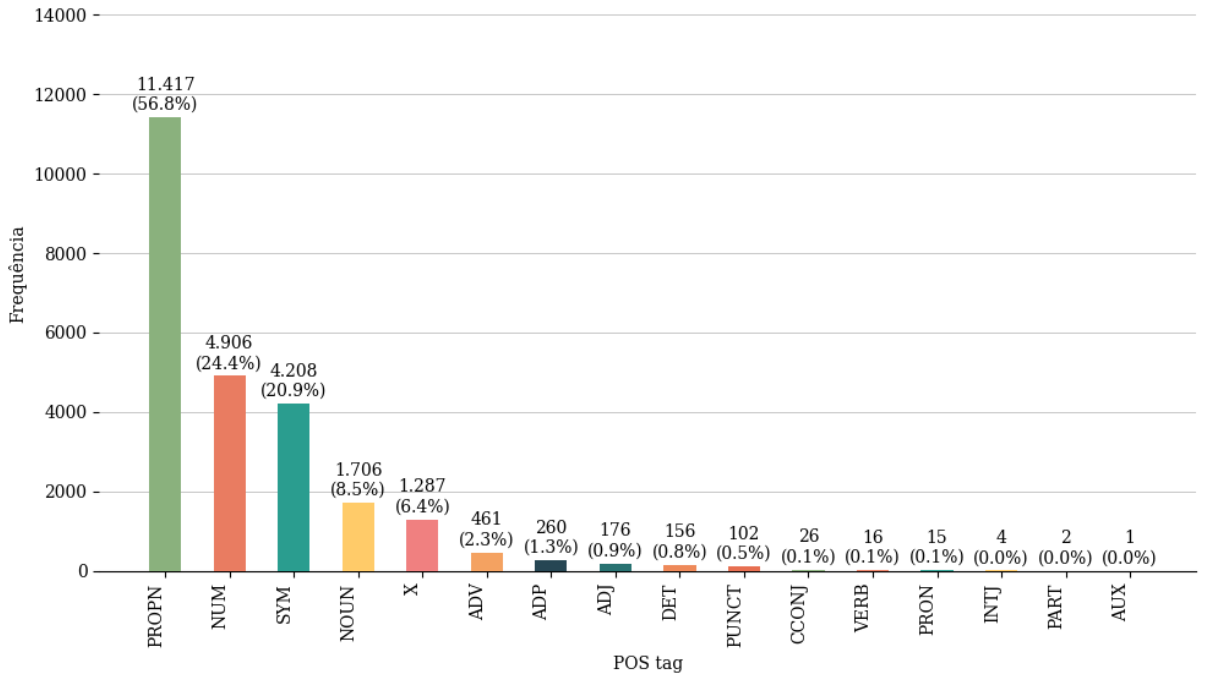


Tabela 6.1: Frequência de categorias, tipos e expressões linguísticas.

Categoria	**Qt.**	**Tipo**	**Expressão linguística**	
			Unitária	**Multipalavra**
Abstração	173	Disciplina	79	30
		Estado	-	3
		Ideia	39	21
		Nome	1	-
Acontecimento	64	Efeméride	4	2
		Evento	38	12
		Organizado	7	1
Coisa	7.919	Classe	40	8
		MembroClasse	1	-
		Objeto	12	40
		Certificado	255	77
		Indicador	389	21
		Ticker	7.076	-
Local	2.372	Físico	8	13
		Humano	146	39
		Virtual	2.109	57
Obra	41	Plano	11	8
		Reproduzida	5	17
Organização	2.011	Administração	170	22
		Empresa	1.443	325
		Instituição	49	2
Pessoa	1.519	Cargo	-	18
		GrupoCargo	-	1
		Individual	271	68
		GrupoInd	-	4
		Membro	2	-
		GrupoMembro	12	1
		Povo	3	-
		Usuário	1.138	1
Tempo	1.855	Duração	8	59
		Frequência	1	3
		Genérico	3	-
		TempoCalendário	1.374	407
Valor	4.138	Classificação	69	-
		Moeda	1.792	842
		Quantidade	343	1.092
Total	**20.092**	-	16.898	3.194

Fonte: A autora, 2025.

A classe dos nomes comuns (NOUN) corresponde a 8.5% dos *tokens* anotados (isto é, 1.706 ocorrências) e expressam entidades em todas as 9 categorias que ocorrem no *corpus*. Além de expressarem EN da categoria TEMPO, como hora, mês e dia da semana, os nomes comuns tam-

bém representam conceitos importantes do domínio do mercado financeiro, como "mercado de capitais" e "mão invisível[31]" (ambos ABSTRAÇÃO-*ideia*), por exemplo.

Vale ressaltar ainda que as 1.287 ENs (6.4%) expressas pela etiqueta X (Outros) correspondem basicamente aos indexadores de tópicos na plataforma. Trata-se das *hashtags* (p.ex.: "#BancodoBrasil" e "#usim5") e *cashtags* (p.ex.: "$CSNA3") que ocorrem no início ou final dos *posts* e que não compõem a estrutura sintática dos *tweets*.

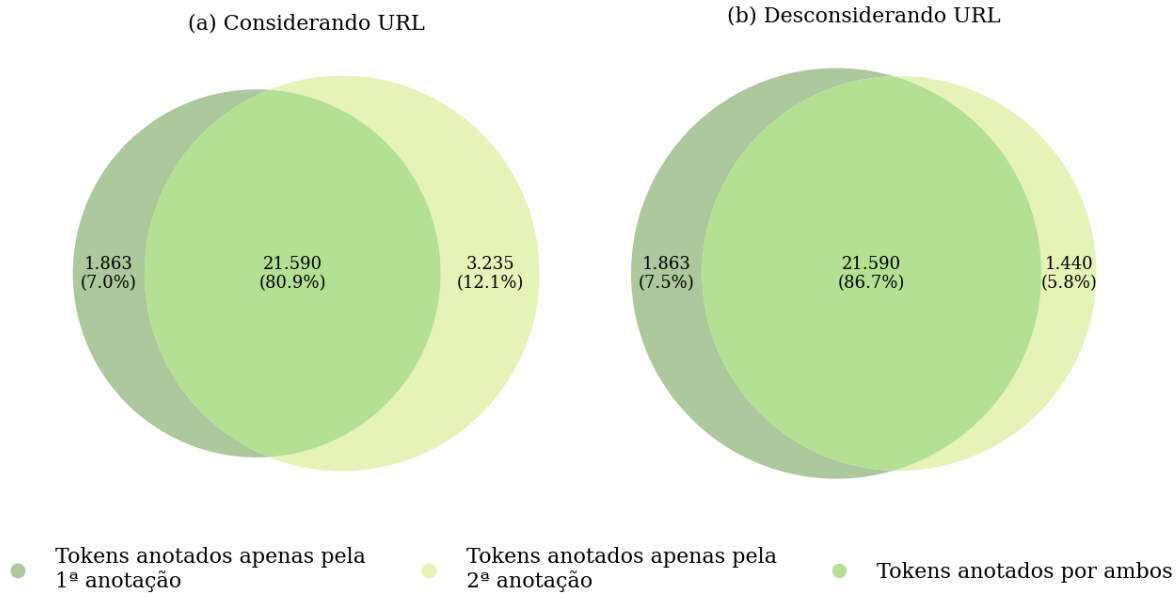
Por fim, realizou-se uma comparação entre a anotação desenvolvida neste trabalho e a de Zerbinati, Roman e Di-Felippo (2024). O diagrama de Venn (Figura 6.8a) revela uma expressiva interseção entre as anotações, com 21.590 *tokens* compartilhados por ambos os trabalhos, equivalentes a 80,9% do total combinado de *tokens*. Tal índice de sobreposição indica não apenas a aderência a critérios comuns de identificação e segmentação, mas também a existência de uma base metodológica convergente, ancorada nas diretrizes do Segundo HAREM. Esse alinhamento reforça a consistência dos procedimentos adotados e a compatibilidade entre os protocolos de anotação empregados nos estudos comparados.

No que se refere às divergências, observa-se que a presente anotação incorporou, de forma exclusiva, 3.235 *tokens* (12,1%), um volume superior aos 1.863 *tokens* exclusivos da anotação anterior (7,0%). Tal expansão decorre, em grande medida, da inclusão das URLs, classificadas como tipo LOCAL-*virtual*, pois elas representam 1.795 dos 3.235 exclusivos. Esse impacto torna-se evidente na análise do diagrama de Venn da Figura 6.8b. Nele, as URLs foram desconsideradas e, com isso, o índice de sobreposição entre as anotações aumentou para 86,7%, enquanto a taxa de divergência relativa à nova anotação foi reduzida para 5,8%. As demais discrepâncias observadas resultam, predominantemente, da adoção de novas diretrizes de delimitação e da reclassificação de algumas categorias.

É necessário ressaltar, contudo, que esta não é uma análise de concordância inter-anotadores (IAA). O cálculo do IAA pressupõe que diferentes anotadores apliquem o mesmo conjunto de diretrizes para medir a consistência de suas interpretações, o que não ocorre aqui. Isso se deve ao fato de que as duas camadas de anotação foram construídas sob premissas distintas. Enquanto a primeira utilizou 10 categorias genéricas, a segunda emprega uma taxonomia expandida para 47 tipos e propõe novas diretrizes para identificação e delimitação de entidades.

[31] A expressão "mão invisível" é uma metáfora de origem econômica, atribuída ao filósofo e economista Adam Smith. Ela se refere ao mecanismo pelo qual o mercado se autorregula por meio das ações individuais dos agentes econômicos, mesmo sem uma coordenação central explícita.

Figura 6.8: Interseção de *tokens*: comparação entre a 1ª e a 2ª anotação de EN.



Sendo assim, o objetivo não é observar a concordância entre as anotações, mas quão distantes ou similares elas são.

Para comprar a similaridade entre o conjunto de entidades anotado por (Zerbinati; Roman; Di-Felippo, 2024) e o anotado neste trabalho, aplicou-se o coeficiente ou índice Jaccard (Jaccard, 1901; Madureira, 2024), o qual, mais precisamente, quantifica a semelhança entre dois conjuntos de dados.

Conforme definido pela Equação 6.1, para dois conjuntos, A e B, o índice J é calculado dividindo o número de elementos em comum entre eles pelo número total de elementos únicos presentes em ambos. Mais precisamente, o índice determina (i) os elementos que são comuns aos dois conjuntos (interseção), (ii) os conjunto que contém todos os elementos únicos de ambos os conjuntos (união) e (iii) divide o número de elementos na interseção pelo número de elementos na união.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{6.1}$$

Nesse índice, o valor 1 indica sobreposição total entre os conjuntos, enquanto 0 representa uma disjunção completa, assim quanto mais próximo de 1 mais simulares os dados. A análise foi conduzida por categoria de entidade nomeada, em dois cenários, considerando URL ou desconsiderando URL. Os resultados estão detalhados na tabela 6.2.

Tabela 6.2: Análise Comparativa de Tokens Anotados por Categoria (Índice Jaccard).

Categoria	Índice Jaccard (Sem URL)	Índice Jaccard (Com URL)
VALOR	0.945	0.945
COISA	0.854	0.854
PESSOA	0.832	0.832
ORGANIZAÇÃO	0.816	0.816
TEMPO	0.667	0.667
LOCAL	**0.593**	**0.211**
ACONTECIMENTO	0.124	0.124
OBRA	0.101	0.101
ABSTRAÇÃO	0.005	0.005

Fonte: A autora, 2025.

A análise de similaridade aponta alta convergência para as categorias VALOR (0.945), COISA (0.854), PESSOA (0.832) e ORGANIZAÇÃO (0.816), enquanto TEMPO (0.667) exibe similaridade moderada. Para esse grupo de categorias, os pontos de distanciamento são, em grande parte, resultado de um refinamento nas diretrizes de identificação e delimitação do escopo das entidades.

As mudanças refletem a exclusão de modificadores e o tratamento de contrações, principalmente para VALOR e TEMPO, o refinamento na categoria PESSOA com a inclusão consistente de cargos e títulos como "Ministro", aplicando uma diretriz do HAREM que não era seguida de forma consistente na anotação anterior, a anotação sistemática de termos do domínio em COISA, como "*shooting star*", e a classificação de órgãos de gestão como "C.A." em ORGANIZAÇÃO.

Embora a divergência na categoria LOCAL seja a mais acentuada em valores absolutos, caindo de 0.593 para 0.211 com a inclusão das URLs, esse era um resultado previsto, decorrente da nova diretriz metodológica que passou a anotar esses elementos. Dessa forma, as baixas similaridades observadas em ACONTECIMENTO (0.124), OBRA (0.101) e ABSTRAÇÃO (0.005) são mais significativas, pois apontam para uma reeinterpretação distinta dos critérios de identificação em cada anotação. Isso ocorre possivelmente por serem categorias de natureza mais subjetiva e, portanto, mais propensas a variações na anotação.

Como observa Freitas (2022), a anotação de *corpus* constitui, em sua essência, uma atividade de natureza interpretativa, e não um processo meramente mecânico de aplicação de regras. Trata-se de uma prática analítica conduzida por anotadores cujas decisões são inevitavelmente influenciadas por seu contexto sociocognitivo e por suas perspectivas teóricas. A própria dificuldade em se estabelecer um esquema de anotação consensual, mesmo no âmbito

do Segundo HAREM, evidencia o caráter não trivial e, em certa medida, subjetivo dessa tarefa. Essa compreensão fundamenta a leitura crítica das anotações de ENs no DANTEStocks. Conforme antecipado pelos próprios organizadores do Segundo HAREM (Santos; Freitas et al., 2008), a aplicação de uma taxonomia por novas equipes, em novos conjuntos de dados, tende inevitavelmente a gerar ajustes, refinamentos e reinterpretações das diretrizes originais, como feito neste trabalho.

Destaca-se que os recursos desenvolvidos nesta pesquisa, como os dados anotados, o manual de diretrizes e as regras de anotação semiautomática podem ser encontrados no seguinte repositório Github: https://github.com/laispiai/DANTENER---Anotacao.

Capítulo 7

Considerações finais

Neste trabalho, anotaram-se as ENs no *tweebank* do mercado financeiro DANTEStocks. Esse recurso já possuía uma versão prévia desse tipo de anotação, feita para subsidiar a exploração da informação de PoS na tarefa de REN. Essa primeira anotação baseou-se exclusivamente nas 10 categorias genéricas e nas diretrizes de anotação do Segundo HAREM, sem alterações.

A análise da aplicação das diretrizes do Segundo HAREM ao DANTEStocks revelou a necessidade de refinamentos. Uma das principais limitações foi o emprego da capitalização como heurística primária para a identificação de entidades. Em textos escritos segundo a norma padrão, esse critério tende a ser eficaz, uma vez que a maioria das entidades é expressa por meio de nomes próprios. No entanto, em postagens de mídias sociais, essa abordagem frequentemente leva à omissão de entidades, já que os usuários desse gênero textual costumam se afastar das convenções de capitalização (Finin et al., 2010).

Outro problema observado nas diretrizes do HAREM refere-se à exclusão de URLs como EN. Considerando sua alta frequência em *tweets*, a ausência de anotação desses elementos pode resultar na perda de informações referenciais relevantes, comprometendo a capacidade de sistemas de PLN rastrearem fontes ou desambiguarem entidades (Liu; Zhu et al., 2018).

Como limitação adicional, a análise das anotações de ENs no DANTEStocks evidenciou um tratamento inconsistente de certos fenômenos típicos de CGU, como truncamento de palavras, além de particularidades do domínio financeiro, como abreviações informais.

Por fim, uma observação crítica refere-se à limitação imposta pelo uso exclusivo das 10 categorias gerais da taxonomia do HAREM (ABSTRAÇÃO, ACONTECIMENTO, COISA, LOCAL, OBRA, ORGANIZAÇÃO, PESSOA, TEMPO, VALOR e OUTRO). Embora essa estratégia contribua para a confiabilidade da anotação e favoreça o desenvolvimento de ferramentas de

PLN interoperáveis, ela pode também simplificar em excesso o processo de anotação, dificultando a identificação de nuances do domínio financeiro e de tipos de entidade mais específicos, que são cruciais para uma análise textual precisa nesse contexto (Sekine; Nobata, 2004).

Diante disso, fez-se a reanotação do *corpus* de forma independente, refinando as diretrizes para o emprego das mesmas 10 categorias por meio de decisões linguisticamente motivadas, e expandindo a categorização ao empregar uma coleção de 47 tipos adaptada do HAREM para atender às particularidades do corpus/domínios.

A anotação dos 4.048 *tweets*, conduzida de forma semiautomática por uma única anotadora, resultou na identificação de 20.092 ENs. Esse elenco de ENs foi caracterizado com base na distribuição estatística das categorias e tipos da taxonomia hierárquica, assim como da forma ou expressão linguística das entidades.

Tal distribuição evidenciou uma predominância de entidades nomeadas unitárias, com especial destaque para as categorias COISA e VALOR. Os tipos COISA-*ticker* e VALOR-*moeda* foram os mais frequentes, o que se justifica não apenas pelo domínio do mercado financeiro, mas também pelo critério de compilação do *corpus*, que se baseou na ocorrência dos *tickers*. No que tange ao domínio, a predominância desses dois tipos pode ser justificada pela natureza informativa e estatística dos *tweets* do DANTEStocks, que tendem a veicular as ações associadas a seu valor no mercado. Ressalta-se também a expressividade dos tipos LOCAL–*virtual* e PESSOA–*usuário*, bastante característicos do gênero "*tweet*". Com isso, pode-se dizer que a taxonomia hierárquica capturou de forma mais granular (que a primeira anotação) características próprias do domínio e do gênero em questão.

A principal dificuldade foi exatamente propor diretrizes para o tratamento de certos fenômenos característicos de CGU, sobretudo truncamento, menção e variação ortográfica, do domínio do mercado financeiro, principalmente as expressões de valor e o vocabulário especializado, e do próprio pré-processamento do *corpus*, como a *tokenização* das contrações.

Buscou-se tratar esses fenômenos com base em fundamentos linguísticos, sendo a noção de EN um deles. No caso, pautou-se em uma concepção mais ampla de entidade, que abrange não apenas o que pode ser referido por um nome próprio, mas também elementos que foram considerados relevantes para a caracterização das entidades do domínio, como datas, expressões numéricas e conceitos nominais diversos. Dessa forma, pode-se dizer que a definição de EN adotada neste trabalho é pragmática, motivada pela aplicação (REN) e pelo domínio.

Nesse sentido, as contribuições consistem não apenas em oferecer uma anotação mais refinada para as ENs do DANTEStocks, mas também em estabelecer diretrizes para o tratamento de fenômenos CGU (particularmente de *tweets*) e específicos do domínio do mercado de ações. Mais especificamente, destacam-se as seguintes contribuições:

- Adaptação da taxonomia hierárquica de ENs do Segundo HAREM ao *tweebank* DANTEStocks, por meio na inclusão de tipos específicos que capturam conhecimento do domínio do mercado financeiro;
- Proposição de diretrizes para anotação de ENs em fenômenos linguísticos de CGU e do domínio, consolidadas em um manual didático publicado pela Biblioteca do ICMC/USP-São Carlos (Piai; Di-Felippo; Roman, 2025).
- Enriquecimento do *corpus* DANTEStocks com uma camada de anotação de ENs segundo a taxonomia adaptada do Segundo HAREM; tal camada, por ser resultante de boas práticas de anotação e ter confiabilidade comprovada indiretamente pela comparação ao trabalho de Zerbinati, Roman e Di-Felippo (2024), pode servir como recurso de referência para o PLN;
- Emprego de uma taxonomia hierárquica de classes de tipos que pode contribuir para que a tarefa de REN capture nuances específicas do domínio com maior granularidade, as quais podem ser relevantes para aplicações de PLN que requeiram uma análise mais precisa das ENs;
- Caracterização linguística do *tweebank* DANTEStocks por meio da análise distribucional das categorias, tipos e formas de expressão das ENs, que fornece insumos linguísticos para o desenvolvimento de métodos/modelos de REN.

Reconhece-se como limitação o fato de a anotação ter sido realizada por uma única pesquisadora, o que impede a mensuração formal de concordância interanotador. Ainda assim, a comparação com a anotação anterior permite avaliar ganhos de cobertura e refinamento.

Para futuros trabalhos, propõe-se (i) a adjudicação dos casos de anotação divergentes entre as versões. (ii) Adicionalmente, o *corpus* anotado permitirá uma análise multidimensional, explorando a relação entre as Entidades Nomeadas e outras anotações já existentes, como dependências sintáticas e emoções. Essa compreensão aprofundada do recurso abrirá caminhos

para diversas aplicações de PLN, como sistemas mais precisos de análise de sentimento relacionada aos ativos negociados, extração de informação em tempo real para eventos de mercado e o ajuste fino de Modelos de Linguagem para tarefas como a predição de tendências e a criação de sistemas de recomendação de investimentos no mercado financeiro brasileiro.

Referências bibliográficas

AGHAJANI, M.; BADRI, A.; BEIGY, H. ParsTwiNER: A Corpus for Named Entity Recognition at Informal Persian. *In:* WORKSHOP ON NOISY USER-GENERATED TEXT, 7. **Proceedings [...].** Online: Association for Computational Linguistics, nov. 2021. P. 131–136. DOI: 10.18653/v1/2021.wnut-1.16. Disponível em: https://aclanthology.org/2021.wnut-1.16.

AGUILAR, G. et al. A Multi-task Approach for Named Entity Recognition in Social Media Data. *In:* WORKSHOP ON NOISY USER-GENERATED TEXT, 3. **Proceedings [...].** Copenhagen, Denmark: Association for Computational Linguistics, set. 2017. P. 148–153. DOI: 10.18653/v1/W17-4419. Disponível em: https://aclanthology.org/W17-4419.

AKHTAR, M. S.; SIKDAR, U. K.; EKBAL, A. IITP: Multiobjective Differential Evolution based Twitter Named Entity Recognition. *In:* WORKSHOP ON NOISY USER-GENERATED TEXT. **Proceedings[...].** Beijing, China: Association for Computational Linguistics, jul. 2015. P. 61–67. DOI: 10.18653/v1/W15-4308. Disponível em: https://aclanthology.org/W15-4308.

ALBUQUERQUE, H. O.; SOUZA, E. et al. Named Entity Recognition: a Survey for the Portuguese Language. **Procesamiento del Lenguaje Natural**, v. 70, n. 0, p. 171–185, 2023. ISSN 1989-7553. Disponível em: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6488.

ALBUQUERQUE, H. O.; COSTA, R. et al. **UlyssesNER-Br: a corpus of brazilian legislative documents for named entity recognition**. [S.l.]: Springer, 2022. DOI: 10.1007/978-3-030-98305-5_1.

AMARAL, D.; VIEIRA, R. NERPCRF: uma ferramenta para o reconhecimento de entidades nomeadas por meio de conditional random fields. **Linguamática**, v. 6, n. 1, p. 41–49, 2014.

ARAUJO, P. H. L. d. et al. LeNER-Br: A Dataset for Named Entity Recognition in Brazilian Legal Text. *In:* COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE. **Proceedings [...].** Cham: Springer International Publishing, 2018. Disponível em: https://github.com/peluz/lener-br.

BALDWIN, T. et al. Shared Tasks of the 2015 Workshop on Noisy User-generated Text: Twitter Lexical Normalization and Named Entity Recognition. *In:* WORKSHOP ON NOISY USER-GENERATED TEXT. **Proceedings [...].** Beijing, China: Association for Computational Linguistics, jul. 2015. P. 126–135. DOI: 10.18653/v1/W15-4319. Disponível em: https://aclanthology.org/W15-4319.

BARBOSA, B. **Descrição sintático-semântica de nomes predicadores em tweets do mercado financeiro em português**. 2024. F. 208. MSc Dissertation – Universidade Federal de São Carlos, São Carlos, SP.

BELLO, T. et al. STSA: A Stock Tweets Sentiment Analysis Corpus for Financial Forecasting. *In:* INTERNATIONAL CONFERENCE ON MACHINE LEARNING (ICML). **Proceedings [...].** Online: International Machine Learning Society, 2021. P. 2048–2056.

BICK, E. Functional aspects in Portuguese NER. *In:* INTERNATIONAL WORKSHOP ON COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE (PROPOR). **Proceedings [...].** [S.l.: s.n.], 2006. P. 80–89.

BOLLEN, J.; MAO, H.; ZENG, X. Twitter mood predicts the stock market. **Journal of Computational Science**, Elsevier BV, v. 2, n. 1, p. 1–8, mar. 2011. ISSN 1877-7503. DOI: 10.1016/j.jocs.2010.12.007. Disponível em: http://dx.doi.org/10.1016/j.jocs.2010.12.007.

BRUM, H. B.; NUNES, M. G. V. N. Building a sentiment corpus of tweets in Brazilian Portuguese. *In:* INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC),11. **Proceedings [...].** Miyazaki, Japan: [s.n.], 2018. P. 4167–4172.

CARDOSO, N. REMBRANDT: reconhecimento de entidades mencionadas baseado em relações e análise detalhada do texto. *In:* DESAFIOS NA AVALIAÇÃO CONJUNTA DO RECONHECIMENTO DE ENTIDADES MENCIONADAS: O SEGUNDO HAREM. **Proceedings [...].** [S.l.]: Linguateca, 2008. P. 195–211.

CARLETTA, J. Assessing Agreement on Classification Tasks: The Kappa Statistic. Edição: Julia Hirschberg. **Computational Linguistics**, MIT Press, Cambridge, MA, v. 22, n. 2, p. 249–254, 1996. Disponível em: https://aclanthology.org/J96-2004.

CAROSIA, A. E. d. O.; COELHO, G. P.; SILVA, A. E. A. d. Analyzing the Brazilian Financial Market through Portuguese Sentiment Analysis in Social Media. **Applied Artificial Intelligence**, Taylor & Francis, v. 34, n. 1, p. 1–19, 2020. DOI: 10.1080/08839514.2019.1673037. eprint: https://doi.org/10.1080/08839514.2019.1673037. Disponível em: https://doi.org/10.1080/08839514.2019.1673037.

CARVALHO, P.; OLIVEIRA, H. G.; SANTOS, D. et al. Segundo HAREM: Modelo geral, novidades e avaliação. *In:* MOTA, C.; SANTOS, D. (Ed.). **Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM**. 1. ed. Porto, Portugal: Linguateca, 2008. cap. 1.

CARVALHO, P.; FREITAS, C. Apêndice E: Exemplário do Segundo HAREM. *In:* MOTA, C.; SANTOS, D. (Ed.). **Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM**. [S.l.]: Linguateca, 2008. P. 323–338.

CASTRO, P. V. Q. d. **Aprendizagem Profunda para Reconhecimento de Entidades Nomeadas em Domínio Jurídico**. 2019. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Goiás, Goiânia. Orientadora: Profa. Dra. Nádia Félix Felipe da Silva; Co-orientador: Prof. Dr. Anderson da Silva Soares. Disponível em: http://repositorio.bc.ufg.br/tede/handle/tede/10276.

CHAPLYNSKYI, D.; ROMANYSHYN, M. Introducing NER-UK 2.0: A Rich Corpus of Named Entities for Ukrainian. *In:* UKRAINIAN NATURAL LANGUAGE PROCESSING WORKSHOP (UNLP), 3. **Proceedings[...].** Torino, Italia: ELRA e ICCL, mai. 2024. P. 23–29. Disponível em: https://aclanthology.org/2024.unlp-1.4.

CHERRY, C.; GUO, H.; DAI, C. NRC: Infused Phrase Vectors for Named Entity Recognition in Twitter. *In:* WORKSHOP ON NOISY USER-GENERATED TEXT. **Proceedings[...].** Beijing, China: Association for Computational Linguistics, jul. 2015. P. 54–60. DOI: 10.18653/v1/W15-4307. Disponível em: https://aclanthology.org/W15-4307.

CLARO, D. B. et al. Extração de Informação. *In:* CASELI, H. M.; NUNES, M. G. V. (Ed.). **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. 3. ed. [S.l.]: BPLN, 2024. cap. 22. ISBN 978-65-01-20581-6. Disponível em: https://brasileiraspln.com/livro-pln/3a-edicao/parte-aplicacoes/cap-ie/cap-ie.html.

COLLOVINI, S. et al. IberLEF 2019 Portuguese Named Entity Recognition and Relation Extraction Tasks. *In:* IBERIAN LANGUAGES EVALUATION FORUM CO-LOCATED CONFERENCE OF THE SPANISH SOCIETY FOR NATURAL LANGUAGE PROCESSING, 35. **Proceedings [...].** [S.l.: s.n.], 2019. P. 390–410. Disponível em: https://ceur-ws.org/Vol%202421/NER_Portuguese_overview.pdf.

COSTA, R. P. d. **Reconhecimento de entidades nomeadas em textos informais no domínio legislativo**. 2023. F. 70. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Goiás, Goiânia.

COVAS, E. **Named entity recognition using GPT for identifying comparable companies**. [S.l.: s.n.], 2023. arXiv: 2307.07420 [cs.CL]. Disponível em: https://arxiv.org/abs/2307.07420.

DÄNIKEN, P. von; CIELIEBAK, M. Transfer Learning and Sentence Level Features for Named Entity Recognition on Tweets. *In:* WORKSHOP ON NOISY USER-GENERATED TEXT, 3. **Proceedings [...].** Copenhagen, Denmark: Association for Computational Linguistics, set. 2017. P. 166–171. DOI: 10.18653/v1/W17-4422. Disponível em: https://aclanthology.org/W17-4422.

DAVIDSON, S. et al. Improved Named Entity Recognition for Noisy Call Center Transcripts. *In:* WORKSHOP ON NOISY USER-GENERATED TEXT, 7. **Proceedings [...].** Online: Association for Computational Linguistics, nov. 2021. P. 361–370. DOI: 10.18653/v1/2021.wnut-1.40. Disponível em: https://aclanthology.org/2021.wnut-1.40.

DERCZYNSKI, L.; BONTCHEVA, K.; ROBERTS, I. Broad Twitter Corpus: A Diverse Named Entity Recognition Resource. *In:* INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS: TECHNICAL PAPERS, 26. **Proceedings [...].** Osaka, Japan: The COLING 2016 Organizing Committee, dez. 2016. P. 1169–1179. Disponível em: https://aclanthology.org/C16-1111.

DERCZYNSKI, L.; NICHOLS, E. et al. Results of the WNUT2017 Shared Task on Novel and Emerging Entity Recognition. *In:* WORKSHOP ON NOISY USER-GENERATED TEXT, 3. **Proceedings [...].** Copenhagen, Denmark: Association for Computational Linguistics, set. 2017. P. 140–147. DOI: 10.18653/v1/W17-4418. Disponível em: https://aclanthology.org/W17-4418.

DEVLIN, J. et al. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. [S.l.: s.n.], 2019. arXiv: 1810.04805 [cs.CL]. Disponível em: https://arxiv.org/abs/1810.04805.

DHABE, P. et al. Stock Market Trend Prediction Along with Twitter Sentiment Analysis. *In:* INTELLIGENT COMPUTING AND NETWORKING. **Proceedings [...].** Singapore: Springer Nature Singapore, 2023. P. 45–59. ISBN 978-981-99-0071-8.

DODDINGTON, G. et al. The Automatic Content Extraction (ACE) Program – Tasks, Data, and Evaluation. *In:* INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC), 4. **Proceedings [...].** Lisbon, Portugal: European Language Resources Association (ELRA), mai. 2004. Disponível em: http://www.lrec-conf.org/proceedings/lrec2004/pdf/5.pdf.

DUGAS, F.; NICHOLS, E. DeepNNNER: Applying BLSTM-CNNs and Extended Lexicons to Named Entity Recognition in Tweets. *In:* WORKSHOP ON NOISY USER-GENERATED TEXT (WNUT), 2. **Proceedings [...].** Osaka, Japan: The COLING 2016 Organizing Committee, dez. 2016. P. 178–187. Disponível em: https://aclanthology.org/W16-3924.

DURAN, M. S. **Manual de Anotação de Relações de Dependência - Versão Revisada e Estendida: Orientações para anotação de relações de dependência sintática em Língua Portuguesa, seguindo as diretrizes da abordagem Universal Dependencies (UD)**. São Carlos, 2022. P. 166.

DURAN, M. S. **Manual de Anotação de Relações de Dependência: Orientações para Anotação de Relações de Dependência Sintática em Língua Portuguesa, seguindo as Diretrizes da Abordagem Universal Dependencies (UD)**. São Carlos, 2021. P. 79.

DURAN, M.; PARDO, T. Anotação de córpus, um lugar privilegiado de observação linguística: o estudo das aposições do português brasileiro segundo o modelo Universal Dependencies. *In:* ENCONTRO DE LINGUÍSTICA DE CORPUS (ELC), 16. **Anais [...].** Porto Alegre: [s.n.], 2024.

EISENSTEIN, J. What to do about bad language on the internet. *In:* CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES. **Proceedings [...].** Atlanta, Georgia: Association for Computational Linguistics, jun. 2013. P. 359–369. Disponível em: https://aclanthology.org/N13-1037/.

EPURE, E.; HENNEQUIN, R. A Human Subject Study of Named Entity Recognition in Conversational Music Recommendation Queries. *In:* CONFERENCE OF THE EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 17. **Proceedings[...].** Dubrovnik, Croatia: Association for Computational Linguistics, mai. 2023. P. 1281–1296. DOI: 10.18653/v1/2023.eacl-main.92. Disponível em: https://aclanthology.org/2023.eacl-main.92.

ESMAAIL, N. et al. Named Entity Recognition in User-Generated Text: A Systematic Literature Review. **IEEE Access**, v. 12, p. 136330–136353, 2024. DOI: 10.1109/ACCESS.2024.3427714.

ESPINOSA, K. J.; BATISTA-NAVARRO, R. T.; ANANIADOU, S. Learning to recognise named entities in tweets by exploiting weakly labelled data. *In:* WORKSHOP ON NOISY USER-GENERATED TEXT (WNUT), 2. **Proceedings [...].** Osaka, Japan: The COLING 2016 Organizing Committee, dez. 2016. P. 153–163. Disponível em: https://aclanthology.org/W16-3921.

DI-FELIPPO, A.; POSTALI, C.; CEREGATTO, G.; GAZANA, L. S.; ROMAN, N. T. **Diretrizes de Anotação de PoS Tags em Tweets do Mercado Financeiro: Orientações para Anotação em Língua Portuguesa segundo a Abordagem Universal Dependencies**. São Carlos-SP, 2022. P. 24. Relatório Técnico n. 438 – ICMC, USP.

DI-FELIPPO, A.; NUNES, M. d. G. V.; BARBOSA, B. K. d. S. A Dependency Treebank of Tweets in Brazilian Portuguese: Syntactic Annotation Issues and Approach. *In:* SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA, 15. **Anais [...].** Belém/PA: SBC, 2024. P. 192–201. DOI: 10.5753/stil.2024.245383. Disponível em: https://sol.sbc.org.br/index.php/stil/article/view/31131.

DI-FELIPPO, A.; NUNES, M. d. G. V.; BARBOSA, B. K. d. S. **Diretrizes de anotação de relações de dependência em tweets do mercado financeiro**. São Carlos, 2024. P. 70. Relatório Técnico n. 446 – ICMC, USP.

DI-FELIPPO, A.; POSTALI, C.; CEREGATTO, G.; GAZANA, L. S.; SILVA, E. H. d. et al. Descrição preliminar do corpus DANTEStocks: diretrizes de segmentação para anotação segundo Universal Dependencies. *In:* SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (STIL). **Anais [...].** Porto Alegre: [s.n.], 2021. SBC. Disponível em: https://sol.sbc.org.br/index.php/stil/article/view/17813.

DI-FELIPPO, A.; ROMAN, N. et al. Genipapo - A Multigenre Dependency Parser for Brazilian Portuguese. *In:* SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA, 15. **Anais [...].** Belém/PA: SBC, 2024. P. 257–266. DOI: 10.5753/stil.2024.245415. Disponível em: https://sol.sbc.org.br/index.php/stil/article/view/31138.

FININ, T. et al. Annotating Named Entities in Twitter Data with Crowdsourcing. *In:* NAACL HLT 2010 WORKSHOP ON CREATING SPEECH AND LANGUAGE DATA WITH AMAZON'S MECHANICAL TURK. **Proceedings [...].** [S.l.: s.n.], 2010. P. 80–88.

FINKEL, J. R.; GRENAGER, T.; MANNING, C. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *In:* ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL), 43. **Proceedings [...].** Ann Arbor, Michigan: Association for Computational Linguistics, jun. 2005. P. 363–370. DOI: 10.3115/1219840.1219885. Disponível em: https://aclanthology.org/P05-1045/.

FINKEL, J. R.; GRENAGER, T.; MANNING, C. D. Incorporating non-local information into information extraction systems by Gibbs sampling. *In:* 43RD ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL'05). **Proceedings [...].** [S.l.: s.n.], 2005. Association for Computational Linguistics, p. 363–370.

FREITAS, C.; CARVALHO, P.; OLIVEIRA, H. G. et al. Second HAREM: Advancing the State of the Art of Named Entity Recognition in Portuguese. *In:* INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC), 7. **Proceedings [...].** Valetta, Malta: [s.n.], mai. 2010. P. 3630–3637.

FREITAS, C. Dataset e corpus. *In:* CASELI, H. M.; NUNES, M. G. V. (Ed.). **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. 2. ed. [S.l.]: BPLN, 2024. cap. 13. ISBN 978-65-00-95750-1. Disponível em: https://brasileiraspln.com/livro-pln/2a-edicao/parte-dados-avaliacao/cap-dataset-corpus/cap-dataset-corpus.html.

FREITAS, C. **Linguística Computacional**. [S.l.]: Parábola Editorial, 2022. ISBN 978-85-7934-278-3.

FREITAS, C.; SOUZA, E. et al. Recursos linguísticos para o PLN específico de domínio: o Petrolês. **Linguamática**, v. 15, n. 2, p. 51–68, dez. 2023. DOI: 10.21814/lm.15.2.412. Disponível em: https://linguamatica.com/index.php/linguamatica/article/view/412.

FREITAS, E.; BARTH, P. Gênero ou suporte? O entrelaçamento de gêneros no Twitter. **Revista (Con) Textos Linguísticos**, v. 9, n. 12, p. 8–26, 2015.

GERGUIS, M. N.; SALAMA, C.; EL-KHARASHI, M. W. ASU: An Experimental Study on Applying Deep Learning in Twitter Named Entity Recognition. *In:* WORKSHOP ON NOISY USER-GENERATED TEXT (WNUT), 2. **Proceedings [...].** Osaka, Japan: The COLING 2016 Organizing Committee, dez. 2016. P. 188–196. Disponível em: https://aclanthology.org/W16-3925.

GHADDAR, A.; LANGLAIS, P. WiNER: A Wikipedia Annotated Corpus for Named Entity Recognition. *In:* INTERNATIONAL JOINT CONFERENCE ON NATURAL LANGUAGE PROCESSING, 8. **Proceedings [...].** Taipei, Taiwan: Asian Federation of Natural Language Processing, nov. 2017. P. 413–422. Disponível em: https://aclanthology.org/I17-1042.

GODIN, F. et al. Multimedia Lab @ ACL WNUT NER Shared Task: Named Entity Recognition for Twitter Microposts using Distributed Word Representations. *In:* WORKSHOP ON NOISY USER-GENERATED TEXT. **Proceedings [...].** Beijing, China: Association for Computational Linguistics, jul. 2015. P. 146–153. DOI: 10.18653/v1/W15-4322. Disponível em: https://aclanthology.org/W15-4322.

GOH, X. P. et al. FinTweet: A New Financial Twitter Corpus for Sentiment Analysis. *In:* 2021 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (EMNLP 2021). **Proceedings [...].** Online: Association for Computational Linguistics, 2021. P. 2997–3007.

GRISHMAN, R.; SUNDHEIM, B. Message Understanding Conference–6: A Brief History. **COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics**, 1996. Disponível em: https://aclanthology.org/C96-1079.

JACCARD, P. Distribution de la Flore Alpine dans le Bassin des Dranses et dans quelques régions voisines. **Bulletin de la Societe Vaudoise des Sciences Naturelles**, v. 37, p. 241–72, jan. 1901. DOI: 10.5169/seals-266440.

JANSSON, P.; LIU, S. Distributed Representation, LDA Topic Modelling and Deep Learning for Emerging Named Entity Recognition from Social Media. *In:* WORKSHOP ON NOISY USER-GENERATED TEXT, 3, p. 154–159. DOI: 10.18653/v1/W17-4420. Disponível em: https://aclanthology.org/W17-4420.

JIANG, H. et al. Annotating the Tweebank Corpus on Named Entity Recognition and Building NLP Models for Social Media Analysis. **arXiv preprint arXiv:2201.07281**, 2022.

JUNIOR, E. C. et al. PELESent: Cross-domain polarity classification using distant supervision. *In:* BRAZILIAN CONFERENCE ON INTELLIGENT SYSTEMS (BRACIS),6. **Proceedings [...].** Uberlândia, Brazil: [s.n.], 2017. P. 49–54.

JÚNIOR, C. M. et al. Paramopama: a Brazilian Portuguese corpus for named entity recognition. *In:* ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL E COMPUTACIONAL (ENIAC). **Proceedings [...].** [S.l.]: SBC, 2015.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition**. 3rd (draft). [S.l.: s.n.], 2025. Acesso em: 20 jan. 2025. Disponível em: https://web.stanford.edu/~jurafsky/slp3/.

KAPLAN, M. May I Ask Who's Calling? Named Entity Recognition on Call Center Transcripts for Privacy Law Compliance. *In:* WORKSHOP ON NOISY USER-GENERATED TEXT, 6. **Proceedings [...].** Online: Association for Computational Linguistics, nov. 2020. P. 1–6. DOI: 10.18653/v1/2020.wnut-1.1. Disponível em: https://aclanthology.org/2020.wnut-1.1.

KIM, J.-D. et al. GENIA corpus–semantically annotated corpus for biotextmining. **Bioinformatics**, Oxford University Press, v. 19, n. 1, p. i182–i182, 2003. DOI: 10.1093/bioinformatics/btg1023.

KITCHENHAM, B. Procedures for Performing Systematic Reviews. **Keele, UK, Keele Univ.**, v. 33, ago. 2004.

KLIE, J.-C. et al. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. *In:* INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS: SYSTEM DEMONSTRATIONS, 27. **Proceedings [...].** [S.l.: s.n.], 2018. Association for Computational Linguistics, p. 5–9.

KRIPKE, S. A. **Naming and Necessity**. Cambridge, MA: Harvard University Press, 1980.

KRISHNAN, A. et al. Employing Wikipedia as a resource for Named Entity Recognition in Morphologically complex under-resourced languages. *In:* WORKSHOP ON BUILDING AND USING COMPARABLE CORPORA, 14. **Proceedings[...].** Online (Virtual Mode): INCOMA Ltd., set. 2021. P. 28–39. Disponível em: https://aclanthology.org/2021.bucc-1.5.

KRUMM, J.; DAVIES, N.; NARAYANASWAMI, C. User-Generated Content. **IEEE Pervasive Computing**, v. 7, n. 4, p. 10–11, 2008. DOI: 10.1109/MPRV.2008.85.

KURNIAWAN, K.; LOUVAN, S. Empirical Evaluation of Character-Based Model on Neural Named-Entity Recognition in Indonesian Conversational Texts. *In:* WORKSHOP ON NOISY USER-GENERATED TEXT, 4. **Proceedings [...].** Brussels, Belgium: Association for Computational Linguistics, nov. 2018. P. 85–92. DOI: 10.18653/v1/W18-6112. Disponível em: https://aclanthology.org/W18-6112.

LAFFERTY, J. D.; MCCALLUM, A.; PEREIRA, F. C. N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *In:* INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 18. **Proceedings [...].** San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001. (ICML '01), p. 282–289. ISBN 1558607781.

LARGUESA, R. P. **Engenharia de Prompt para Devs: Um guia para aprender a usar a IA antes que ela te use**. São Paulo: Casa do Código, 2024. Acessado em: 27 fev. 2025. Disponível em: https://www.casadocodigo.com.br/products/livro-engenharia-de-prompt.

LE, N. T.; MALLEK, F.; SADAT, F. UQAM-NTL: Named entity recognition in Twitter messages. *In:* WORKSHOP ON NOISY USER-GENERATED TEXT (WNUT), 2. **Proceedings [...].** Osaka, Japan: The COLING 2016 Organizing Committee, dez. 2016. P. 197–202. Disponível em: https://aclanthology.org/W16-3926.

LEECH, G. Adding linguistic annotation. *In:* **Developing linguistic corpora : a guide to good practice**. Edição: M. Wynne. [S.l.]: Oxbow Books, 2005. P. 17–29.

LIMSOPATHAM, N.; COLLIER, N. Bidirectional LSTM for Named Entity Recognition in Twitter Messages. *In:* WORKSHOP ON NOISY USER-GENERATED TEXT (WNUT), 2. **Proceedings [...].** Osaka, Japan: The COLING 2016 Organizing Committee, dez. 2016. P. 145–152. Disponível em: https://aclanthology.org/W16-3920.

LIN, B. Y. et al. Multi-channel BiLSTM-CRF Model for Emerging Named Entity Recognition in Social Media. *In:* WORKSHOP ON NOISY USER-GENERATED TEXT, 3. **Proceedings [...].** Copenhagen, Denmark: Association for Computational Linguistics, set. 2017. P. 160–165. DOI: 10.18653/v1/W17-4421. Disponível em: https://aclanthology.org/W17-4421.

LIU, X.; ZHANG, S. et al. Recognizing Named Entities in Tweets. *In:* 49TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES. **Proceedings [...].** Edição: Dekang Lin, Yuji Matsumoto e Rada Mihalcea. Portland, Oregon, USA: Association for Computational Linguistics, jun. 2011. P. 359–367. Disponível em: https://aclanthology.org/P11-1037/.

LIU, Y.; ZHU, Y. et al. Parsing Tweets into Universal Dependencies. *In:* 2018 CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES, VOLUME 1 (LONG PAPERS). **Proceedings [...].** [S.l.]: Association for Computational Linguistics, 2018. P. 965–975. DOI: 10.18653/v1/N18-1088. Disponível em: https://aclanthology.org/N18-1088.

LIU, Y.; DENG, G. et al. Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study. **arXiv**, v. 2305, n. 13860, 2023. Disponível em: https://arxiv.org/abs/2305.13860.

LOPES, F.; TEIXEIRA, C.; OLIVEIRA, H. G. Named entity recognition in portuguese neurology text using CRF. *In:* EPIA CONFERENCE ON ARTIFICIAL INTELLIGENCE. **Proceedings [...].** [S.l.: s.n.], 2019. P. 336–348.

LOPES, L.; DURAN, M. S. et al. PortiLexicon-UD: a Portuguese lexical resource according to Universal Dependencies model. *In:* LANGUAGE RESOURCES AND EVALUATION CONFERENCE,13. **Proceedings[...].** [S.l.: s.n.], 2022.

MADUREIRA, B. Avaliação de tecnologias de linguagem. *In:* CASELI, H. M.; NUNES, M. G. V. (Ed.). **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. 3. ed. [S.l.]: BPLN, 2024. cap. 14. ISBN 978-65-01-20581-6. Disponível em: https://brasileiraspln.com/livro-pln/3a-edicao/parte-dados-avaliacao/cap-avaliacao/cap-avaliacao.html.

MARRERO, M. et al. Named Entity Recognition: Fallacies, challenges and opportunities. **Computer Standards & Interfaces**, v. 35, n. 5, p. 482–489, 2013. ISSN 0920-5489. DOI: https://doi.org/10.1016/j.csi.2012.09.004. Disponível em: https://www.sciencedirect.com/science/article/pii/S0920548912001080.

MARTINELLI, G. et al. CNER: Concept and Named Entity Recognition. *In:* CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES. **Proceedings[...].** Mexico City, Mexico: Association for Computational Linguistics, jun. 2024. P. 8336–8351. DOI: 10.18653/v1/2024.naacl-long.461. Disponível em: https://aclanthology.org/2024.naacl-long.461.

MCGREGOR, S. C.; MOLYNEUX, L. Twitter's influence on news judgment: An experiment among journalists. **Journalism**, v. 21, n. 5, p. 597–613, 2020.

MILIDIÚ, R. L.; DUARTE, J. C.; CAVALCANTE, R. Machine Learning Algorithms for Portuguese Named Entity Recognition. **Inteligência Artificial**, v. 11, n. 36, p. 67–75, 2007.

MISHRA, S.; DIESNER, J. Semi-supervised Named Entity Recognition in noisy-text. *In:* WORKSHOP ON NOISY USER-GENERATED TEXT (WNUT), 2. **Proceedings [...].** Osaka, Japan: The COLING 2016 Organizing Committee, dez. 2016. P. 203–212. Disponível em: https://aclanthology.org/W16-3927.

MORAES, S. M. W.; MANSSOUR, I. H.; SILVEIRA, M. S. 7x1-PT: um corpus extraído do Twitter para análise de sentimentos em língua portuguesa. *In:* 10TH SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY (STIL). **Proceedings [...].** Natal, Brazil: Sociedade Brasileira de Computação, 2015. P. 21–25.

MOTA, C.; SANTOS, D. **Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM**. [S.l.]: Linguateca, 2008.

NADEAU, D.; SEKINE, S. A survey of named entity recognition and classification. **Lingvisticae Investigationes**, v. 30, n. 1, p. 3–26, jan. 2007. DOI: 10.1075/li.30.1.03nad. Disponível em: http://www.ingentaconnect.com/content/jbp/li/2007/00000030/00000001/art00002.

NANDIGAM, P.; APPIDI, A.; SHRIVASTAVA, M. Named Entity Recognition for Code-Mixed Kannada-English Social Media Data. *In:* INTERNATIONAL CONFERENCE ON NATURAL LANGUAGE PROCESSING (ICON), 19. **Proceedings[...].** New Delhi, India: Association for Computational Linguistics, dez. 2022. P. 43–49. Disponível em: https://aclanthology.org/2022.icon-main.5.

NGUYEN, D. Q.; VU, T.; NGUYEN, A. T. BERTweet: A pre-trained language model for English tweets. *In:* LIU, Q.; SCHLANGEN, D. (Ed.). **2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**. [S.l.: s.n.], 2020. P. 9–14.

NIVRE, J. et al. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. English. *In:* LANGUAGE RESOURCES AND EVALUATION CONFERENCE, 12. **Proceedings [...].** Marseille, France: European Language Resources Association, mai. 2020. P. 4034–4043. ISBN 979-10-95546-34-4. Disponível em: https://aclanthology.org/2020.lrec-1.497.

OLIVEIRA, L. et al. SemClinBr: a multi-institutional and multi-specialty semantically annotated corpus for Portuguese clinical NLP tasks. **Journal of Biomedical Semantics**, v. 13, n. 13, 2022. DOI: 10.1186/s13326-022-00276-9.

PARDO, T. A. S. et al. Porttinari - a large multi-genre treebank for Brazilian Portuguese. *In:* 14TH SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE. **Proceedings [...].** [S.l.: s.n.], 2021. P. 1–10.

PARTALAS, I. et al. Learning to Search for Recognizing Named Entities in Twitter. *In:* WORKSHOP ON NOISY USER-GENERATED TEXT (WNUT), 2. **Proceedings [...].** Osaka, Japan: The COLING 2016 Organizing Committee, dez. 2016. P. 171–177. Disponível em: https://aclanthology.org/W16-3923.

PENG, N.; DREDZE, M. Named Entity Recognition for Chinese Social Media with Jointly Trained Embeddings. *In:* CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING. **Proceedings [...].** Lisbon, Portugal: Association for Computational Linguistics, set. 2015. P. 548–554. DOI: 10.18653/v1/D15-1064. Disponível em: https://aclanthology.org/D15-1064.

PERES, R. d. S.; ESTEVES, D.; MAHESHWARI, G. Bidirectional LSTM with a Context Input Window for Named Entity Recognition in Tweets. *In:* 9TH KNOWLEDGE CAPTURE CONFERENCE. **Proceedings [...].** [S.l.: s.n.], dez. 2017. P. 1–4. DOI: 10.1145/3148011.3154478.

PIAI, L.; DI-FELIPPO, A.; ROMAN, N. T. **Guia de anotação de entidades nomeadas em tweets do mercado financeiro: adaptação da taxonomia hierárquica do segundo HAREM**. São Carlos, 2025. Disponível em: https://repositorio.usp.br/item/003258357. Acesso em: 19 de ago. de 2025.

PIRES, A. R. O. **Named Entity Extraction from Portuguese Web Text**. 2017. Diss. (Mestrado) – Faculdade de Engenharia da Universidade do Porto. AAI30206191. ISBN 9798358424623.

PIROVANI, J.; OLIVEIRA, E. Portuguese named entity recognition using conditional random fields and local grammars. *In:* INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC), 11. **Proceedings [...].** [S.l.: s.n.], 2018. P. 4452–4456.

PLUTCHIK, R.; KELLERMAN, H. (Ed.). **Emotion: theory, research and experience**. New York: Academic Press, 1986.

QI, P. et al. Stanza: A Python natural language processing toolkit for many human languages. **arXiv preprint arXiv:2003.07082**, 2020.

RADEMAKER, A. et al. Universal dependencies for Portuguese. *In:* FOURTH INTERNATIONAL CONFERENCE ON DEPENDENCY LINGUISTICS (DEPLING 2017). **Proceedings [...].** [S.l.: s.n.], 2017. P. 197–206.

RAMSHAW, L. A.; MARCUS, M. P. Text chunking using transformation-based learning. *In:* ARMSTRONG, S. et al. (Ed.). **Natural Language Processing Using Very Large Corpora**. [S.l.]: Springer, 1999. P. 157–176.

RATINOV, L.; ROTH, D. Design Challenges and Misconceptions in Named Entity Recognition. *In:* CONFERENCE on Computational Natural Language Learning (CoNLL), 13. Boulder, Colorado: Association for Computational Linguistics, jun. 2009. P. 147–155. Disponível em: https://aclanthology.org/W09-1119/.

RIJHWANI, S.; PREOTIUC-PIETRO, D. Temporally-Informed Analysis of Named Entity Recognition. *In:* ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 58. **Proceedings[...].** Online: Association for Computational Linguistics, jul. 2020. P. 7605–7617. DOI: 10.18653/v1/2020.acl-main.680. Disponível em: https://aclanthology.org/2020.acl-main.680.

RITTER, A. et al. Named Entity Recognition in Tweets: An Experimental Study. *In:* CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING. **Proceedings [...].** Edinburgh, Scotland, UK.: Association for Computational Linguistics, jul. 2011. P. 1524–1534. Disponível em: https://aclanthology.org/D11-1141/.

ROCHA, C. et al. PAMPO: using pattern matching and pos-tagging for effective named entities recognition in portuguese. **Conference on Computational Processing of the Portuguese Language (PROPOR)**, 2016.

RUSSELL, S. J.; NORVIG, P. **Artificial Intelligence: A Modern Approach**. [S.l.]: Pearson Education, 2016.

SANGUINETTI, M.; BOSCO, C.; CASSIDY, L. et al. Treebanking user-generated content: a UD based overview of guidelines, corpora and unified recommendations. **Language Resources & Evaluation**, v. 57, p. 493–544, 2023. DOI: https://doi.org/10.1007/s10579-022-09581-9.

SANGUINETTI, M.; BOSCO, C.; SARTI, L. The Tweebank: a syntactically annotated corpus of English tweets. *In:* INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS (COLING), 27. Santa Fe, New Mexico, USA. **Proceedings [...].** [S.l.]: Association for Computational Linguistics, 2018. P. 3493–3503. Disponível em: https://aclanthology.org/C18-1334/.

SANTOS, C. N. dos; GUIMARÃES, V. Boosting Named Entity Recognition with Neural Character Embeddings. *In:* NAMED ENTITY WORKSHOP, 5. **Proceedings [...].** [S.l.: s.n.], 2015. P. 25–33.

SANTOS, D.; CARDOSO, N. **Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área**. [S.l.]: Linguateca, 2007. DOI: http://hdl.handle.net/10400.26/380.

SANTOS, D.; FREITAS, C. Avaliação conjunta em português. *In:* CASELI, H. M.; NUNES, M. G. V. (Ed.). **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. 3. ed. [S.l.]: BPLN, 2024. cap. 15. ISBN 978-65-01-20581-6. Disponível em: https://brasileiraspln.com/livro-pln/3a-edicao/parte-dados-avaliacao/cap-avaliacao-conjunta/cap-avaliacao-conjunta.html.

SANTOS, D.; FREITAS, C. et al. Segundo HAREM: Balanço e perspectivas de futuro. *In:* DESAFIOS na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. [S.l.]: Linguateca, jan. 2008. P. 131–146.

SARMENTO, L. SIEMÊS –a named-entity recognizer for portuguese relying on similarity rules. *In:* INTERNATIONAL WORKSHOP ON COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE (PROPOR). **Proceedings [...].** [S.l.: s.n.], 2006. P. 90–99.

SARMENTO, L.; PINTO, A. S.; CABRAL, L. REPENTINO – A Wide-Scope Gazetteer for Entity Recognition in Portuguese. *In:* COMPUTATIONAL PROCESSING OF THE

PORTUGUESE LANGUAGE. **Proceedings [...].** Berlin, Heidelberg: Springer Berlin Heidelberg, 2006. P. 31–40. ISBN 978-3-540-34046-1.

SCANDAROLLI, C. L. et al. Tipologia de fenômenos ortográficos e lexicais em CGU: o caso dos tweets do mercado financeiro. *In:* SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (STIL), 14. **Anais [...].** [S.l.]: Sociedade Brasileira de Computação, 2023. (STIL 2023). DOI: 10.5753/stil.2023.233948. Disponível em: http://dx.doi.org/10.5753/stil.2023.233948.

SCANNAVINO, K. R. F. et al. **Revisão Sistemática da Literatura em Engenharia de Software: teoria e prática**. [S.l.]: Elsevier, 2017.

SEKINE, S.; NOBATA, C. Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy. *In:* THE INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC), 4. **Proceedings[...].** Lisbon, Portugal: European Language Resources Association (ELRA), mai. 2004. Disponível em: https://aclanthology.org/L04-1051/.

SIKDAR, U. K.; GAMBÄCK, B. A Feature-based Ensemble Approach to Recognition of Emerging and Rare Named Entities. *In:* WORKSHOP ON NOISY USER-GENERATED TEXT, 3. **Proceedings [...].** Copenhagen, Denmark: Association for Computational Linguistics, set. 2017. P. 177–181. DOI: 10.18653/v1/W17-4424. Disponível em: https://aclanthology.org/W17-4424.

SIKDAR, U. K.; GAMBÄCK, B. Feature-Rich Twitter Named Entity Recognition and Classification. *In:* WORKSHOP ON NOISY USER-GENERATED TEXT (WNUT), 2. **Proceedings [...].** Osaka, Japan: The COLING 2016 Organizing Committee, dez. 2016. P. 164–170. Disponível em: https://aclanthology.org/W16-3922.

SILVA, A. V. e. Uma revisão para o Reconhecimento de Entidades Nomeadas aplicado à língua portuguesa. v. 15, p. 69–85, dez. 2023. DOI: 10.21814/lm.15.2.396. Disponível em: https://linguamatica.com/index.php/linguamatica/article/view/396.

SILVA, E. H.; PARDO, T. A. S. et al. Universal Dependencies for tweets in Brazilian Portuguese: tokenization and part of speech tagging. *In:* NATIONAL MEETING ON ARTIFICIAL AND COMPUTATIONAL INTELLIGENCE, 18. **Proceedings [...].** [S.l.: s.n.], 2021. P. 1–12.

SILVA, E. et al. Etiquetagem morfossintática multigênero para o português do Brasil segundo o modelo "Universal Dependencies". *In:* PROCEEDINGS [...]. Belo Horizonte, Brazil: SBC, 2023. P. 63–73.

SILVA, F. J. V. da; ROMAN, N. T.; CARVALHO, A. Stock market tweets annotated with emotions. **Corpora**, v. 15, p. 343–354, 2020. Disponível em: https://api.semanticscholar.org/CorpusID:234526636.

SILVA, I. S.; GOMIDE, J. et al. Effective sentiment stream analysis with self-augmenting training and demand-driven projection. *In:* 34TH INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL. **Proceedings [...].** Beijing, China: [s.n.], 2011. P. 475–484.

SILVA, R. R.; PARDO, T. A. S. Córpus 4P: um córpus anotado de opiniões em português sobre produtos eletrônicos para fins de sumarização contrastiva de opinião. *In:* JORNADA DE

DESCRIÇÃO DO PORTUGUÊS (JDP), 6. **Anais [...].** Salvador, Bahia, Brazil: [s.n.], out. 2019. P. 330–338.

SINCLAIR, J. Corpus and Text - Basic Principles. *In:* WYNNE, M. (Ed.). **Developing Linguistic Corpora: a Guide to Good Practice**. Oxford: Oxbow Books, 2005. Acesso em: 30/10/2006. P. 1–16. Disponível em: http://ahds.ac.uk/linguistic-corpora/.

SINGH, K.; SEN, I.; KUMARAGURU, P. Language Identification and Named Entity Recognition in Hinglish Code Mixed Tweets. *In:* STUDENT RESEARCH WORKSHOP. **Proceedings [...].** Melbourne, Australia: Association for Computational Linguistics, jul. 2018. P. 52–58. DOI: 10.18653/v1/P18-3008. Disponível em: https://aclanthology.org/P18-3008.

SINGH, V.; VIJAY, D. et al. Named Entity Recognition for Hindi-English Code-Mixed Social Media Text. *In:* NAMED ENTITIES WORKSHOP, 7. **Proceedings [...].** Melbourne, Australia: Association for Computational Linguistics, jul. 2018. P. 27–35. DOI: 10.18653/v1/W18-2405. Disponível em: https://aclanthology.org/W18-2405.

SIRTS, K. Estonian Named Entity Recognition: New Datasets and Models. *In:* NORDIC CONFERENCE ON COMPUTATIONAL LINGUISTICS (NODALIDA), 24. **Proceedings [...].** Tórshavn, Faroe Islands: University of Tartu Library, mai. 2023. P. 752–761. Disponível em: https://aclanthology.org/2023.nodalida-1.76.

SOLORIO, T. MALINCHE: a NER system for portuguese that reuses knowledge from Spanish. *In:* RECONHECIMENTO DE ENTIDADES MENCIONADAS EM PORTUGUÊS: DOCUMENTAÇÃO E ACTAS DO HAREM A PRIMEIRA AVALIAÇÃO CONJUNTA NA ÁREA. **Proceedings [...].** [S.l.]: Linguateca, 2007. P. 123–136.

SOUZA, E.; ALBUQUERQUE, H. O. et al. PLN no Direito - REN: Reconhecimento de Entidades Nomeadas no Domínio Legal: um Panorama para a Língua Portuguesa. *In:* CASELI, H. M.; NUNES, M. G. V. (Ed.). **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. 3. ed. [S.l.]: BPLN, 2024. cap. 31. ISBN 978-65-01-20581-6. Disponível em: https://brasileiraspln.com/livro-pln/3a-edicao/parte-dominios/cap-direito-ren/cap-direito-ren.html.

SOUZA, E. de; FREITAS, C. ET: A Workstation for Querying, Editing and Evaluating Annotated Corpora. *In:* CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING: SYSTEM DEMONSTRATIONS. **Proceedings [...].** Online e Punta Cana, Dominican Republic: Association for Computational Linguistics, nov. 2021. P. 35–41. Disponível em: https://aclanthology.org/2021.emnlp-demo.5.

SOUZA, F.; NOGUEIRA, R. F.; LOTUFO, R. A. Portuguese Named Entity Recognition using BERT-CRF. **CoRR**, abs/1909.10649, 2019. arXiv: 1909.10649. Disponível em: http://arxiv.org/abs/1909.10649.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. *In:* 9TH BRAZILIAN CONFERENCE ON INTELLIGENT SYSTEMS (BRACIS). **Proceedings [...].** [S.l.: s.n.], 2020. P. 403–417.

SOUZA, W.; FERNANDES, D.; FERNANDES, M. Ontologia aplicada à redução de ruído em base de dados de tweets sobre mercado financeiro. *In:* ESCOLA REGIONAL DE INFORMÁTICA DE GOIÁS, 9. **Anais [...].** Evento Online: SBC, 2021. P. 26–39. DOI:

10.5753/erigo.2021.18431. Disponível em: https://sol.sbc.org.br/index.php/erigo/article/view/18431.

SRIRANGAM, V. K. et al. Corpus Creation and Analysis for Named Entity Recognition in Telugu-English Code-Mixed Social Media Data. *In:* ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: STUDENT RESEARCH WORKSHOP, 57. **Proceedings [...].** Florence, Italy: Association for Computational Linguistics, jul. 2019. P. 183–189. DOI: 10.18653/v1/P19-2025. Disponível em: https://aclanthology.org/P19-2025.

STEFANOWITSCH, A. **Corpus Linguistics: A Guide to the Methodology**. [S.l.: s.n.], mai. 2020. ISBN 978-3-96110-224-2. DOI: 10.5281/zenodo.3735822.

STENETORP, P. et al. brat: a web-based tool for NLP-assisted text annotation. *In:* DEMONSTRATIONS AT THE 13TH CONFERENCE OF THE EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings [...].** [S.l.: s.n.], 2012. P. 102–107.

STRAKA, M. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. *In:* CONLL 2018 SHARED TASK: MULTILINGUAL PARSING FROM RAW TEXT TO UNIVERSAL DEPENDENCIES. **Proceedings [...].** [S.l.: s.n.], 2018. P. 197–207.

STRAUSS, B. et al. Results of the WNUT16 Named Entity Recognition Shared Task. *In:* WORKSHOP ON NOISY USER-GENERATED TEXT (WNUT), 2. **Proceedings [...].** Osaka, Japan: The COLING 2016 Organizing Committee, dez. 2016. P. 138–144. Disponível em: https://aclanthology.org/W16-3919.

SUMUKH, S.; SHRIVASTAVA, M. "Kanglish alli names!" Named Entity Recognition for Kannada-English Code-Mixed Social Media Data. *In:* WORKSHOP ON NOISY USER-GENERATED TEXT, 8. **Proceedings [...].** Gyeongju, Republic of Korea: Association for Computational Linguistics, out. 2022. P. 154–161. Disponível em: https://aclanthology.org/2022.wnut-1.17.

TEDESCHI, S.; NAVIGLI, R. MultiNERD: A Multilingual, Multi-Genre and Fine-Grained Dataset for Named Entity Recognition (and Disambiguation). *In:* FINDINGS OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings[...].** Seattle, United States: Association for Computational Linguistics, jul. 2022. P. 801–812. DOI: 10.18653/v1/2022.findings-naacl.60. Disponível em: https://aclanthology.org/2022.findings-naacl.60.

TETEO, L. et al. Um Framework de Extração e Etiquetamento de Informações de Trânsito. *In:* WORKSHOP EM DESEMPENHO DE SISTEMAS COMPUTACIONAIS E DE COMUNICAÇÃO, 8. **Anais [...].** Belém: SBC, 2019. DOI: 10.5753/wperformance.2019.6472. Disponível em: https://sol.sbc.org.br/index.php/wperformance/article/view/6472.

THOMPSON, P.; ANANIADOU, S.; TSUJII, J. The GENIA corpus: Annotation levels and applications. *In:* IDE, N.; PUSTEJOVSKY, J. (Ed.). **Handbook of Linguistic Annotation**. [S.l.]: Springer, 2017. P. 1395–1432. DOI: 10.1007/978-94-024-0881-2_54.

TIAN, T.; DINARELLI, M.; TELLIER, I. Data Adaptation for Named Entity Recognition on Tweets with Features-Rich CRF. *In:* WORKSHOP ON NOISY USER-GENERATED TEXT.

Proceedings[...]. Beijing, China: Association for Computational Linguistics, jul. 2015. P. 68–71. DOI: 10.18653/v1/W15-4309. Disponível em: https://aclanthology.org/W15-4309.

TJONG KIM SANG, E. F. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. *In:* CONFERENCE ON NATURAL LANGUAGE LEARNING 2002 (CONLL), 6. **Proceedings [...].** [S.l.: s.n.], 2002. Disponível em: https://aclanthology.org/W02-2024.

TJONG KIM SANG, E. F.; DE MEULDER, F. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. *In:* CONFERENCE ON NATURAL LANGUAGE LEARNING AT HLT, 7. **Proceedings [...].** [S.l.: s.n.], 2003. P. 142–147. Disponível em: https://aclanthology.org/W03-0419.

TOH, Z.; CHEN, B.; SU, J. Improving Twitter Named Entity Recognition using Word Representations. *In:* WORKSHOP ON NOISY USER-GENERATED TEXT. **Proceedings [...].** Beijing, China: Association for Computational Linguistics, jul. 2015. P. 141–145. DOI: 10.18653/v1/W15-4321. Disponível em: https://aclanthology.org/W15-4321.

TOPÇU, B.; DURGAR EL-KAHLOUT, İ. TR-SEQ: Named Entity Recognition Dataset for Turkish Search Engine Queries. *In:* INTERNATIONAL CONFERENCE ON RECENT ADVANCES IN NATURAL LANGUAGE PROCESSING. **Proceedings[...].** Held Online: INCOMA Ltd., set. 2021. P. 1417–1422. Disponível em: https://aclanthology.org/2021.ranlp-1.158.

USHIO, A. et al. Named Entity Recognition in Twitter: A Dataset and Analysis on Short-Term Temporal Shifts. **arXiv preprint arXiv:2210.03797**, 2022.

VASWANI, A. et al. Attention is all you need. *In:* ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS. **Proceedings [...].** [S.l.: s.n.], 2017. P. 5998–6008. Disponível em: http://arxiv.org/abs/1706.03762.

VIEIRA DA SILVA, F. J.; ROMAN, N. T.; CARVALHO, A. M. B. R. **Building An Emotionally Annotated Corpus of Investor Tweets**. [S.l.], mar. 2018. In English, 10 pages. **Abstract** Emotionally annotated corpora are specially important for training machine learning models for automatic emotion identification, among other applications. However, the task of manually assigning emotions to a corpus carries a high level of subjectivity. In this technical report, we describe the annotation tools and methodology we used for dealing with this challenge when building an emotionally annotated corpus of investor tweets.

YAMADA, I.; TAKEDA, H.; TAKEFUJI, Y. Enhancing Named Entity Recognition in Twitter Messages Using Entity Linking. *In:* WORKSHOP ON NOISY USER-GENERATED TEXT. **Proceedings [...].** Beijing, China: Association for Computational Linguistics, jul. 2015. P. 136–140. DOI: 10.18653/v1/W15-4320. Disponível em: https://aclanthology.org/W15-4320.

YANG, E.-S.; KIM, Y.-S. Hallym: Named Entity Recognition on Twitter with Word Representation. *In:* WORKSHOP ON NOISY USER-GENERATED TEXT. **Proceedings[...].** Beijing, China: Association for Computational Linguistics, jul. 2015. P. 72–77. DOI: 10.18653/v1/W15-4310. Disponível em: https://aclanthology.org/W15-4310.

YANG, L.; LIN, C. et al. FinNER-Twitter: A Named Entity Recognition Dataset for Financial Tweets. *In:* 2020 IEEE INTERNATIONAL CONFERENCE ON BIG DATA (BIG DATA 2020). **Proceedings [...].** Online: IEEE, 2020. P. 1928–1937.

ZERBINATI, M. M.; ROMAN, N. T. **Manual de Anotação de Entidades Nomeadas do DANTEStocks utilizando categorias do Segundo HAREM**. São Paulo, SP, set. 2023. Disponível em: http://www.each.usp.br/ppgsi.

ZERBINATI, M. M.; ROMAN, N. T.; DI-FELIPPO, A. A Corpus of Stock Market Tweets Annotated with Named Entities. *In:* INTERNATIONAL CONFERENCE ON COMPUTATIONAL PROCESSING OF PORTUGUESE, 16. **Proceedings [...].** Santiago de Compostela, Galicia/Espanha: Association for Computational Linguistics, mar. 2024. P. 276–284. Disponível em: https://aclanthology.org/2024.propor-1.28.

ZIRIKLY, A.; DIAB, M. Named Entity Recognition for Arabic Social Media. *In:* WORKSHOP ON VECTOR SPACE MODELING FOR NATURAL LANGUAGE PROCESSING, 1. **Proceedings [...].** Denver, Colorado: Association for Computational Linguistics, jun. 2015. P. 176–185. DOI: 10.3115/v1/W15-1524. Disponível em: https://aclanthology.org/W15-1524.

Apêndice A

Primeiro Apêndice

A.1 Artigos Selecionados na Revisão Sistemática

Tabela A.1: Listagem de artigos selecionados para a Revisão Sistemática

Fonte	Referência	Título
Acervo ACL	Peng e Dredze (2015)	Named entity recognition for Chinese social media with jointly trained embeddings
Acervo ACL	Zirikly e Diab (2015)	Named entity recognition for Arabic social media
Acervo ACL	Derczynski, Bontcheva e Roberts (2016)	Broad twitter corpus: A diverse named entity recognition resource
Acervo ACL	Ghaddar e Langlais (2017)	Winer: A Wikipedia annotated corpus for named entity recognition
Acervo ACL	Singh, Sen e Kumaraguru (2018)	Language identification and named entity recognition in Hinglish code mixed tweets
Acervo ACL	Singh, Vijay et al. (2018)	Named entity recognition for Hindi-English code-mixed social media text
Acervo ACL	Srirangam et al. (2019)	Corpus creation and analysis for named entity recognition in Telugu-English code-mixed social media data
Acervo ACL	Rijhwani e Preotiuc-Pietro (2020)	Temporally-informed analysis of named entity recognition

Continua na próxima página

Fonte	Referência	Título
Acervo ACL	Topçu e Durgar El-Kahlout (2021)	TR-SEQ: Named Entity Recognition Dataset for Turkish Search Engine Queries
Acervo ACL	Krishnan et al. (2021)	Employing Wikipedia as a resource for named entity recognition in morphologically complex under-resourced languages
Acervo ACL	Nandigam, Appidi e Shrivastava (2022)	Named Entity Recognition for Code-Mixed Kannada-English Social Media Data
Acervo ACL	Tedeschi e Navigli (2022)	MultiNERD: A multilingual, multi-genre and fine-grained dataset for named entity recognition (and disambiguation)
Acervo ACL	Epure e Hennequin (2023)	A Human Subject Study of Named Entity Recognition in Conversational Music Recommendation Queries
Acervo ACL	Sirts (2023)	Estonian Named Entity Recognition: New Datasets and Models
Acervo ACL	Chaplynskyi e Romanyshyn (2024)	Introducing NER-UK 2.0: A Rich Corpus of Named Entities for Ukrainian
Acervo ACL	Martinelli et al. (2024)	CNER: Concept and Named Entity Recognition
W-NUT	Cherry, Guo e Dai (2015)	NRC: Infused Phrase Vectors for Named Entity Recognition in Twitter
W-NUT	Akhtar, Sikdar e Ekbal (2015)	IITP: Multiobjective Differential Evolution based Twitter Named Entity Recognition
W-NUT	Tian, Dinarelli e Tellier (2015)	Lattice: Data Adaptation for Named Entity Recognition on Tweets with Features-Rich CRF
W-NUT	Yang e Kim (2015)	Hallym: Named Entity Recognition on Twitter with Induced Word Representation
W-NUT	Baldwin et al. (2015)	Shared Tasks of the 2015 Workshop on Noisy User-generated Text: Twitter Lexical Normalization and Named Entity Recognition
W-NUT	Yamada, Takeda e Takefuji (2015)	Enhancing Named Entity Recognition in Twitter Messages Using Entity Linking
W-NUT	Toh, Chen e Su (2015)	Improving Twitter Named Entity Recognition using Word Representations

Continua na próxima página

Fonte	Referência	Título
W-NUT	Godin et al. (2015)	Multimedia Lab @ ACL WNUT NER Shared Task: Named Entity Recognition for Twitter Microposts using Distributed Word Representations
W-NUT	Strauss et al. (2016)	Results of the WNUT16 Named Entity Recognition Shared Task
W-NUT	Limsopatham e Collier (2016)	Bidirectional LSTM for Named Entity Recognition in Twitter Messages
W-NUT	Espinosa, Batista-Navarro e Ananiadou (2016)	Learning to recognise named entities in tweets by exploiting weakly labelled data
W-NUT	Sikdar e Gambäck (2016)	Feature-Rich Twitter Named Entity Recognition and Classification
W-NUT	Partalas et al. (2016)	Learning to Search for Recognizing Named Entities in Twitter
W-NUT	Dugas e Nichols (2016)	DeepNNNER: Applying BLSTM-CNNs and Extended Lexicons to Named Entity Recognition in Tweets
W-NUT	Gerguis, Salama e El-Kharashi (2016)	ASU: An Experimental Study on Applying Deep Learning in Twitter Named Entity Recognition
W-NUT	Le, Mallek e Sadat (2016)	UQAM-NTL: Named entity recognition in Twitter messages
W-NUT	Mishra e Diesner (2016)	Semi-supervised Named Entity Recognition in noisy-text
W-NUT	Däniken e Cieliebak (2017)	Transfer Learning and Sentence Level Features for Named Entity Recognition on Tweets
W-NUT	Aguilar et al. (2017)	A Multi-task Approach for Named Entity Recognition in Social Media Data
W-NUT	Jansson e Liu (2017)	Distributed Representation, LDA Topic Modelling and Deep Learning for Emerging Named Entity Recognition from Social Media
W-NUT	Lin et al. (2017)	Multi-channel BiLSTM-CRF Model for Emerging Named Entity Recognition in Social Media
W-NUT	Sikdar e Gambäck (2017)	A Feature-based Ensemble Approach to Recognition of Emerging and Rare Named Entities
W-NUT	Derczynski, Nichols et al. (2017)	Results of the WNUT2017 Shared Task on Novel and Emerging Entity Recognition

Continua na próxima página

Fonte	Referência	Título
W-NUT	Kurniawan e Louvan (2018)	Empirical Evaluation of Character-Based Model on Neural Named-Entity Recognition in Indonesian Conversational Texts
W-NUT	Kaplan (2020)	May I Ask Who's Calling? Named Entity Recognition on Call Center Transcripts for Privacy Law Compliance
W-NUT	Aghajani, Badri e Beigy (2021)	ParsTwiNER: A Corpus for Named Entity Recognition at Informal Persian
W-NUT	Davidson et al. (2021)	Improved Named Entity Recognition for Noisy Call Center Transcripts
W-NUT	Sumukh e Shrivastava (2022)	"Kanglish alli names!" Named Entity Recognition for Kannada-English Code-Mixed Social Media Data
PROPOR	Zerbinati, Roman e Di-Felippo (2024)	A Corpus of Stock Market Tweets Annotated with Named Entities
Adicional	Teteo et al. (2019)	Um Framework de Extração e Etiquetamento de Informações de Trânsito
Adicional	Costa (2023)	Reconhecimento de Entidades Nomeadas em Textos Informais no Domínio Legislativo
Adicional	Silva (2023)	Uma revisão para o Reconhecimento de Entidades Nomeadas aplicada à língua portuguesa

Apêndice B

Segundo Apêndice

B.1 Sistemas de REN para CGU

Tabela B.1: Sistemas de REN para CGU

Referência	Corpora	Método Base	F1
Kaplan (2020)	Call Center transcripts	BiLSTM-CRF	97.50%
Teteo et al. (2019)	Tweets envolvendo localização	CRF	96.49%
Srirangam et al. (2019)	Telugu-Inglês	CRF	96%
Nandigam, Appidi e Shrivastava (2022)	Kannada-Inglês	LSTM e LSTM-CRF	96%
Singh, Vijay et al. (2018)	Hindi-Inglês (b)	CRF e LSTM	95%
Sumukh e Shrivastava (2022)	Kannada-Inglês (b)	BiLSTM + CRF	94%
Chaplynskyi e Romanyshyn (2024)	NER-UK 2.0	RoBerta large	89%
Krishnan et al. (2021)	WikiZu	XLM-RoBERTa	89%
Krishnan et al. (2021)	WikiMl	XLM-RoBERTa	87%
Martinelli et al. (2024)	CNER	DeBERTa-v3	87.20%
Kurniawan e Louvan (2018)	SMALL-TALK	BiLSTM + CRF	84.97%
Tedeschi e Navigli (2022)	MultiNERD	+Attention mBERT + Bi-LSTM + CRF	83.11%
Davidson et al. (2021)	Call Center transcripts (b)	RoBERTa	81%
Kurniawan e Louvan (2018)	TASK-ORIENTED	BiLSTM + CRF	80.22%
Costa (2023)	C-CORPUS	BERT	78.65%
Teteo et al. (2019)	TC-BERT	BERTurk	76.95%
Epure e Hennequin (2023)	MusicRecoNER	MPNet	76%±4%
Kurniawan e Louvan (2018)	TTC	BiLSTM + CRF	75.54%
Ghaddar e Langlais (2017)	Winer	LSTM + CRF	73%
Sirts (2023)	New NER	EstBERT	73.5%±0.6%

Continua na próxima página

Referência	Corpora	Método Base	F1
Singh, Sen e Kumaraguru (2018)	Hindi-Inglês	CRF	72.06%
Zirikly e Diab (2015)	DA-EGY NER	CRF	72.68%
Aghajani, Badri e Beigy (2021)	ParsTwiNER	ParsBert	69.5%
Yamada, Takeda e Takefuji (2015)	TwitterNER	Entity Linking	56.41%
Limsopatham e Collier (2016)	W-NUT2016	BilLSTM	52.41%
Toh, Chen e Su (2015)	TwitterNER	CRF	51.40%
Mishra e Diesner (2016)	W-NUT2016	CRF	47.30%
Partalas et al. (2016)	W-NUT2016	LS2	46.16%
Peng e Dredze (2015)	Golden Horse	CRF	45.82%
Espinosa, Batista-Navarro e Ananiadou (2016)	W-NUT2016	BilLSTM	44.77%
Cherry, Guo e Dai (2015)	TwitterNER	semi-Markov MIRA	44.74%
Godin et al. (2015)	TwitterNER	FFNN	43.45%
Aguilar et al. (2017)	W-NUT2017	CRF	41.86%
Däniken e Cieliebak (2017)	W-NUT2016	BiLSTM + CRF	40.78%
Lin et al. (2017)	W-NUT2017	BiLSTM + CRF	40.42%
Sikdar e Gambäck (2016)	W-NUT2016	CRF	40.06%
Jansson e Liu (2017)	W-NUT2017	LDA + BiLSTM + CRF	39.98%
Akhtar, Sikdar e Ekbal (2015)	TwitterNER	CRF	39.84%
Sikdar e Gambäck (2017)	W-NUT2017	CRF + SVM + LSTM	38.35%
Gerguis, Salama e El-Kharashi (2016)	W-NUT2016	LSTM	39.02%
Dugas e Nichols (2016)	W-NUT2016	BilLSTM + CNN	37.24%
Yang e Kim (2015)	TwitterNER	CRF	37.21%
Le, Mallek e Sadat (2016)	W-NUT2016	BiLSTM + CRF	29.82%
Tian, Dinarelli e Tellier (2015)	TwitterNER	CRF	16.44%