UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

ISSN 0103-2569

GUIA DE ANOTAÇÃO DE ENTIDADES NOMEADAS EM TWEETS DO MERCADO FINANCEIRO: ADAPTAÇÃO DA TAXONOMIA HIERÁRQUICA DO SEGUNDO HAREM

LAÍS PIAI
ARIANI DI-FELIPPO
NORTON TREVISAN ROMAN

Nº 452

RELATÓRIOS TÉCNICOS

São Carlos – SP
Jul./2025

UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Guia de Anotação de Entidades Nomeadas em *Tweets* do Mercado Financeiro: Adaptação da Taxonomia Hierárquica do Segundo HAREM

Laís Piai, Ariani Di-Felippo, Norton Trevisan Roman

São Carlos
2025

Sumário

1 Introdução

Apresenta-se o manual de anotação de **entidade nomeada** (EN) no projeto DANTE[1] (Dependency-ANalised corpora of TwEets), que, ao integrar o POeTiSA[2], visa à construção de *corpora* de "conteúdo gerado por usuário" (CGU) com várias camadas de anotação linguística para o processamento automático do português.

A falta de consenso sobre a definição de EN é um tema recorrente tanto na Linguística quanto no Processamento de Língua Natural (PLN) (Marrero *et al.*, 2013). Tradicionalmente, EN é definida como nomes próprios de pessoas, locais e organizações (Grishman; Sundheim, 1996). Ao se adotar a taxonomia hierárquica do Segundo HAREM (Mota; Santos, 2008), como descrito na sequência, concebe-se EN de uma forma mais ampla em concordância com a literatura recente (Jurafsky; Martin, 2025), incluindo elementos que não são entidades *per se*, como datas, horários, outras expressões temporais, expressões numéricas e conceitos nominais específicos de domínio. Essa concepção, como evidenciado mais adiante, levou à criação de uma tipologia que busca atender às necessidades linguísticas e informacionais do domínio financeiro.

Especificamente, este documento contém diretrizes referentes ao primeiro *corpus* construído no projeto DANTE. Trata-se do **DANTEStocks** (Di-Felippo; Roman, 2025), *corpus* de *tweets* (*tweebank*) pioneiro do mercado financeiro em português com diversas anotações linguísticas de referência ou padrão-ouro. Os *posts* foram compilados em 2014 com base na ocorrência de ao menos um ***ticker*** de uma das 73 ações que compunham o Ibovespa[3] à época da sua construção. Um *ticker* é um código que identifica os ativos financeiros negociados, como "Petr4", que representa "ação preferencial da Petrobras".

Por consequência da data de compilação, os *posts* do DANTEStocks possuem limite de 140 caracteres imposto pela plataforma. Além disso, os *tweets* estão em sua forma original, isto é, sem segmentação em unidades estruturais menores (sentenças ou sintagmas) e normalização lexical. Assim, o *corpus* possui uma mistura de linguagem padrão e não-padrão (Scandarolli *et al.*, 2023). A não-canonicidade do DANTEStocks é marcada por ortografia e sintaxe irregulares, pontuação assistemática, truncamento, uso de elementos próprios da plataforma (como *hashtags*, menções, etc.), estilo econômico, fragmentado e coloquial, além de elementos típicos do mercado financeiro como *cashtags* (p.ex.: $BBDC3) e *tickers*. Todos esses fenômenos fazem com que qualquer anotação do DANTEStocks seja uma tarefa bastante complexa.

Atualmente, o recurso conta com anotações morfossintáticas (*part-of-speech* ou PoS *tags*) e de dependências segundo o modelo *Universal Dependencies* (UD) (Nivre *et al.*, 2020; Marneffe *et al.*, 2021), além de emoções baseadas nos eixos de oposição de (Plutchik; Kellerman, 1986).

A anotação gramatical UD, em particular, especifica dois níveis de descrição: morfológico e sintático. No primeiro, cada *token* do enunciado (no caso, *tweet*) recebe uma etiqueta de classe gramatical (*part-of-speech* ou PoS), um lema e um conjunto de traços morfológicos. No nível sintático, tem-se relações de dependência (*deprels)* entre os *tokens*, as quais partem do núcleo (*head*) para os dependentes. A representação UD básica é no formato de árvore, em que uma palavra específica é a raiz (*root*). Na Figura 1, tem-se a anotação-UD básica do *tweet* em (1). Nela, as etiquetas de classe gramatical são exibidas acima do texto original (em caixa alta). Os lemas e os traços morfológicos não estão incluídos na figura, mas o verbo "indicado", por exemplo, tem "indicar" como lema e os traços Gender=Masc|Number=Sing|VerbForm=Part. A anotação-UD de um enunciado é codificada em um arquivo no formato CoNLL-U, ilustrado na Figura 2. A anotação de ENs foi feita na coluna 5 dos arquivos CoNLL-U, como é descrito mais adiante.

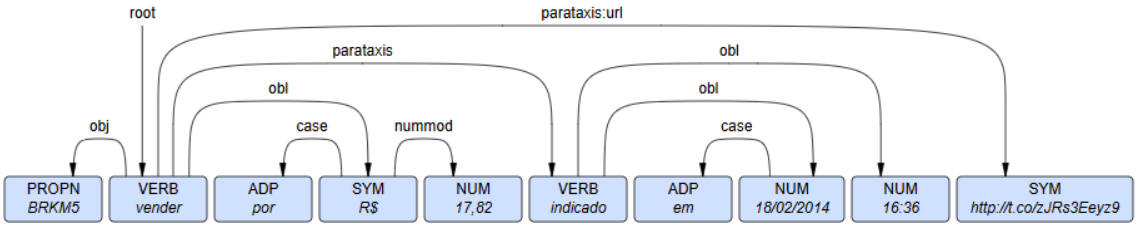
(1) BRKM5 vender por R$ 17,82 indicado em 18/02/2014 16:36 http://t.co/zJRs3Eeyz9.

[1] https://sites.google.com/icmc.usp.br/poetisa/resources-and-tools?authuser=0

[2] https://sites.google.com/icmc.usp.br/poetisa

[3] Principal índice de desempenho das ações negociadas na B3 (Bolsa, Balcão, Brasil), que é a bolsa de valores oficial do Brasil.

Figura 1 – Árvore de dependência sintática segundo o modelo UD.



Fonte: A autora, 2025.

Figura 2 – Arquivo no formato CoNLL-U referente à anotação UD.

Id	Form	Lemma	Upos Tag	Xpos Tag	Feats	Head	DepRel	Deps	Misc
1	BRKM5	BRKM5	PROPN	ENTIDADE=S-COISA-TICKER	_	_	_	_	_
2	vender	vender	VERB	_	VerbForm=Inf	_	_	_	_
3	por	por	ADP	_	_	_	_	_	_
4	R$	R$	SYM	ENTIDADE=B-VALOR-MOEDA	_	_	_	_	_
5	17,82	17,82	NUM	ENTIDADE=E-VALOR-MOEDA	NumType=Card	_	_	_	_
6	indicado	indicar	VERB	_	Gender=Masc|Number=Sing|VerbForm=Part	_	_	_	_
7	em	em	ADP	_	_	_	_	_	_
8	18/02/2014	18/02/2014	NUM	ENTIDADE=B-TEMPO-TEMPO_CALEND	NumType=Card	_	_	_	_
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	_	_	_	_	SpacesAfter=\n

Fonte: A autora, 2025.

O DANTEStocks já foi enriquecido com uma camada *stand-off* de EN (Zerbinati; Roman, 2023; Zerbinati; Roman; Di-Felippo, 2024), tornando-se um recurso apropriado para apoiar o desenvolvimento e a exploração de métodos existentes de reconhecimento de entidades nomeadas (REN). Nessa anotação, as entidades foram anotadas manualmente de acordo com as categorias mais genéricas da taxonomia do Segundo HAREM. Por mais valiosa que seja essa anotação, o emprego das categorias de topo dificulta uma análise mais aprofundada das entidades existentes, o que pode ser de grande interesse em estudos práticos relacionados ao mercado de ações.

Este manual busca ir além das categorias ao fornecer diretrizes para a associação de tipos às entidades, permitindo distinções mais refinadas dentro de cada categoria. Além disso, nossa classificação independente desses *tweets* busca esclarecer alguns trechos e identificar decisões gramaticais imprecisas feitas nas diretrizes originais de anotação (Zerbinati; Roman, 2023), assim como trechos nos quais a anotação de *tweets* deveria se desviar das diretrizes do Segundo HAREM. Nossa contribuição, portanto, não é apenas oferecer uma classificação mais detalhada para a anotação de ENs existente no DANTEStocks, mas também estabelecer diretrizes para lidar com CGU e fenômenos específicos do domínio em *tweets* sobre o mercado de ações.

As diretrizes de anotação de ENs foram delineadas com base na versão 1.0 do *corpus* DANTEStocks, publicada em 15/12/2022[4], a qual possuía apenas anotação PoS à época. Especificamente, o *corpus* possui 4.048 *tweets* e um total de 84.396 *tokens*. Outras versões do *corpus* estão descritas e disponíveis na página do projeto Porttinari 2.0[5].

O restante do relatório está organizado da seguinte forma. Na Seção 2, apresenta-se a adaptação da taxonomia de ENs original do Segundo HAREM ao DANTEStocks. Na Seção 3, detalha-se o esquema de marcação utilizado. Por fim, a Seção 4 reúne o conjunto de diretrizes efetivamente propostas e empregadas no referido *corpus*.

[4] https://drive.google.com/file/d/1ioguMX7dsPsPGrXpHdMac8y7A2wKxhoO/view
[5] https://sites.google.com/icmc.usp.br/poetisa/porttinari-2-0

2 Taxonomia do Segundo HAREM: Adaptação ao DANTEStocks

Para a anotação de ENs no *corpus* DANTEStocks, partiu-se da taxonomia hierárquica empregada na Coleção Dourada do Segundo HAREM (Mota; Santos, 2008), isto é, nos textos de referência utilizados na segunda edição da avaliação conjunta de ferramentas de PLN dedicadas à tarefa de reconhecimento de entidades nomeadas (REN) em português, organizada pela Linguateca[6]. Trata-se de uma taxonomia de dois níveis, composta por 10 categorias genéricas e 43 tipos de ENs (Figura 3).

Diferentemente de Zerbinati, Roman e Di-Felippo (2024), este manual contempla o emprego não só das categorias, mas também dos tipos de ENs. Tal decisão se deve à necessidade do emprego de um *tagset* que fosse abrangente e informativo (Freitas, 2024), pois, embora as 10 categorias fossem importantes para contemplar a abrangência, careciam de informatividade para uma caracterização refinada do domínio.

Objetivando a referida informatividade, a coleção original de tipos foi estendida pela adição de 4 novos tipos específicos para o domínio do mercado financeiro e gênero *tweet*. Especificamente, os tipos *certificado*, *indicador* e *ticker* foram inseridos à categoria COISA, enquanto o tipo *usuário* foi incluído para especificar entidades da categoria PESSOA associadas a perfis na rede social, totalizando 47 tipos. Da taxonomia original, a categoria OUTRO e o tipo *outro* de todas as demais categorias não foram necessários. Além disso, os tipos específicos COISA-*substância*[7] e OBRA-*arte*[8] não ocorreram no *corpus*. Assim, ao final, tem-se a taxonomia descrita na Figura 4, com 9 classes e 36 tipos, sendo os novos tipos destacados em negrito e itálico. As diretrizes e exemplos aqui apresentados são referentes apenas às categorias e tipos da Figura 4. Para mais ilustrações sobre a taxonomia original do Segundo HAREM, recomenda-se a consulta ao seu Exemplário[9].

Figura 3 – Categorias e tipos originais do Segundo HAREM.

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

Fonte: Baseada em Mota e Santos (2008).

Figura 4 – Adaptação da taxonomia do Segundo HAREM ao DANTEStocks.

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

Fonte: A autora, 2025.

[6] Centro de recursos distribuídos para o processamento do português (https://www.linguateca.pt)
[7] Abrange entidades concretas massivas, como vitaminas e elementos químicos.
[8] Engloba obras únicas, como pinturas, monumentos ou esculturas.
[9] https://www.linguateca.pt/aval_conjunta/HAREM/ExemplarioSegundoHAREM.pdf

3 Formato de Anotação

A anotação de ENs, assim como outras tarefas de classificação, envolve duas etapas fundamentais (Jurafsky; Martin, 2025): (i) identificação/delimitação da entidade, isto é, determinar o início e fim do *span* (trecho contínuo de texto) que corresponde à EN, e (ii) classificação da entidade delimitada.

Para explicitar as informações resultantes das duas etapas, optou-se pelo formato de marcação ou anotação denominado BIOES, o mesmo empregado por Zerbinati, Roman e Di-Felippo (2024). Nesse formato, que é uma extensão do esquema BIO, o prefixo "S-" (*Single*) é utilizado para identificar entidades compostas por um único *token*. As ENs multipalavras são delimitadas da seguinte forma: os *tokens* iniciais são marcados como "B-" (*Begin*), os intermediários como "I-" (*Inside*) e os finais recebem o prefixo "E-" (*End*). *Tokens* que não correspondem a entidades não são anotados, sendo representados implicitamente por "O" (*Outside*). Considerando o exemplo em (1), ilustra-se a etapa de identificação/delimitação:

i. Identificação das ENs:

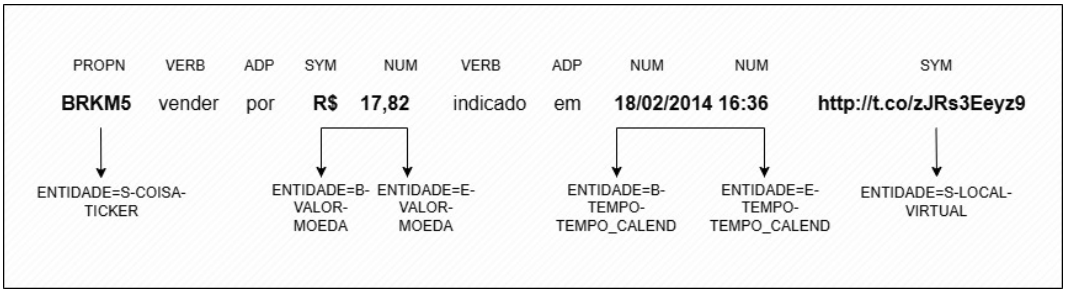
- **BRKM5**: *ticker* de uma ação
- **R$ 17,82**: valor monetário
- **18/02/2014 16:36**: expressão temporal
- **http://t.co/zJRs3Eeyz9**: URL

ii. Delimitação dos *spans*:

- O *ticker* ("BRKM5") e a URL ("http://t.co/zJRs3Eeyz9") são entidades expressas por *tokens* únicos e, portanto, anotadas com a etiqueta **S-**.
- O valor monetário "R$ 17,82" contém dois *tokens*. O símbolo "R$" é marcado como **B-**, pois inicia a entidade, enquanto "17,82" recebe **E-** para sinalizar o final.
- A entidade temporal "18/02/2014 16:36" também é composta por dois *tokens* (conforme diretriz apresentada a seguir), em que a data "18/02/2014" recebe **B-** e o horário "16:36" recebe **E-**.

Na Figura 5, ilustra-se graficamente a identificação/delimitação das ENs no formato BIOES:

Figura 5 – Exemplos de delimitação de ENs simples e compostas em BIOES.



Fonte: A autora, 2025.

Quando uma EN é composta por mais de dois *tokens*, como "Banco do Brasil" (tokenizado para anotação-UD em "Banco de o Brasil"), os *tokens* intermediários são anotados com o rótulo **I-** (*Inside*), como exemplificado na Figura 6.

Figura 6 – Exemplo de EN composta por mais de dois *tokens* delimitada em BIOES.

Fonte: A autora, 2025.

A anotação final de cada EN segue o formato **ENTIDADE=BIOES-CATEGORIA-TIPO**, em que todas as informações são representadas em **caixa alta e sem acentuação**. A anotação de "BRKM5" em (1), por exemplo, é ENTIDADE=S-COISA-TICKER, indicando que se trata de um *token* único (`S-`) da categoria COISA e tipo *ticker*.

A anotação de EN foi incorporada à coluna 5 do arquivo CoNLL-U, originalmente destinada às XPoS (isto é, categorias morfossintáticas específicas de uma língua). Como essa coluna não é utilizada nos projetos DANTE e POeTiSA, optou-se por registrar nela as anotações de ENs, preservando a estrutura padrão do formato CoNLL-U. Um exemplo dessa incorporação é fornecido na Figura 7, que exibe especificamente a anotação de ENs na coluna 5 do CoNLL-U referente ao *tweet* (1) do DANTEStocks.

Figura 7 – Anotação BIOES adicionada ao arquivo CoNLL-U - coluna XPoS

Id	Form	Lemma	Upos Tag	Xpos Tag
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.

4 Diretrizes

Nesta seção, apresentam-se as diretrizes que norteiam a identificação/delimitação e classificação das ENs do DANTEStocks. Esta última em função das categorias e tipos da Figura 4. Por se tratar de uma adaptação da taxonomia e diretrizes do Segundo HAREM, recomenda-se a leitura prévia das diretrizes de anotação de *corpus* tanto do Primeiro (Cardoso; Santos, 2007) quanto do Segundo HAREM (Mota; Santos, 2008), uma vez que o guia de anotação da segunda edição da avaliação conjunta apresenta apenas as alterações em relação à classificação da edição anterior. Recomenda-se também a leitura de Hagège, Baptista e Mamede (2008), pois os autores fornecem as diretrizes específicas para a anotação das entidades temporais do Segundo HAREM. Essas publicações podem ser encontradas na *webpage* da Linguateca, indicada anteriormente na nota 6.

Os exemplos que compõem esta seção foram extraídos do DANTEStocks e as entidades sob descrição estão destacadas em negrito.

4.1 Diretrizes gerais de delimitação e classificação

1. Não identificar as ENs exclusivamente com base na ocorrência de letra maiúscula ou algarismo como no HAREM (Santos; Cardoso, 2007; Mota; Santos, 2008). Isso se deve à natureza da linguagem não-canônica do tipo de CGU coberto pelo DANTEStocks, o (*tweet*), que nem sempre segue as convenções ortográficas da norma padrão. Dessa forma, a maiúscula funciona apenas como uma pista.
2. Utilizar algumas etiquetas PoS do modelo UD como pistas para a identificação das entidades, como PROPN (nome próprio), NUM (numeral), X e NOUN. A etiqueta X, por exemplo, tem entre seus vários usos a categorização dos *tickers* quando, precedidos pelos sinais de *hashtag* (#) ou *cashtag* ($) (como "#Petr4" ou "$Petr4"), funcionam como indexadores[10] de assuntos na plataforma Twitter. Os *tickers*, por codificarem conceitos importantes no mercado de ações, são considerados ENs e a etiqueta X o seu indício.
3. Identificar as ENs com base em certos fenômenos CGU como menções e URLs (além das *hashtags* e *cashtags*), uma vez que eles são importantes na caracterização do gênero e do domínio e classificá-los em decorrência de sua interpretação em contexto.
4. Empregar uma única categoria e tipo, ou seja, aplicar o esquema *single-label*. Isso significa que, diante de múltiplas classificações possíveis para uma mesma EN no *tweet*, deve-se escolher aquela que melhor se encaixa no contexto.
5. Anotar as metáforas conforme a entidade à qual se referem. Quando uma expressão metafórica é usada para representar um ativo (ação da bolsa) ou outro elemento, ela recebe a mesma categoria e tipo da entidade referida, garantindo consistência na anotação. No *tweet* (2), por exemplo, "King Kong" funciona como uma metáfora pejorativa para as ações da Petrobras, sendo interpretada como o "maior mico". Assim, anota-se a expressão com a categoria COISA e o tipo *certificado*, que são os mesmos usados para os ativos financeiros, preservando a coerência semântica da análise.

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

6. Delimitar as entidades pela menor expressão capaz de representar adequadamente seu significado. Para as categorias VALOR e TEMPO, isso normalmente implica a exclusão de modificadores periféricos, como preposições, determinantes e quantificadores. No *tweet* (3), a entidade de valor é expressa unicamente como "R$ 1 bilhão", sem o modificador "mais de". O mesmo ocorre com a entidade temporal "ontem" em (4), delimitada sem incluir o especificador "dia de".

   (3) 13 empresas perdem mais de **R$ 1 bilhão** na Bolsa em fevereiro: Petrobras lidera lista com perdas significativas... http://t.co/t3uG25y3gj #infomoney #vale5

[10] Nesses casos, os *tickers* recebem a etiqueta X do modelo UD por não comporem a estrutura sintática dos *tweets*.

(4) 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

7. Não anotar entidade encaixada, isto é, aquela contida dentro de outra. Assim, em (5), por exemplo, anota-se apenas a entidade maior "Ex-presidente da Petrobras" e não "Petrobras". Essa escolha pauta-se em razões práticas e metodológicas: (i) simplificação do processo de anotação, reduzindo a ambiguidade e aumentando a consistência entre anotadores, (ii) captura da informação mais relevante, evitando redundâncias e favorecendo a padronização, e (iii) alinhamento com convenções estabelecidas em *corpora* de referência, como os do HAREM, o que facilita a reprodutibilidade e a comparação entre anotações e sistemas.

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

4.2 Diretrizes específicas de delimitação

4.2.1 Variações de valor

Com base nas diretrizes de *tokenização* para a anotação UD do *corpus*, uma variação de valor (de um ativo), como "+ 2,09%" em (6), foi segmentada em três *tokens* com as seguintes PoS: símbolo (SYM), valor numérico (NUM) e sinal de porcentagem (SYM). Tratada como uma entidade multipalavra, tal variação é rotulada segundo o esquema BIOES da seguinte forma: o *token* "+" recebe o rótulo "B-", "2,09" recebe "I-", e "%" recebe "E-". A Figura 8 ilustra a proposta de anotação das variações de valor expressas por completo. A diretriz para variações truncadas como "+2," em (6) é definida na Seção a seguir.

(6) 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 8 – Exemplo de anotação de "variação de valor".

Id	Form	Lemma	Upos Tag	Xpos Tag
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.

4.2.2 Truncamentos lexicais

Esse fenômeno ocorre quando uma palavra (ou *token*) é truncada devido ao limite de caracteres imposto pela plataforma do Twitter, sendo frequentemente indicadas por reticências ao final.

As ENs truncadas devem ser anotadas quando sua categoria-tipo puder ser determinada com base no contexto, em outro *tweet*, em julgamento especializado ou em fontes externas. Em todos os casos, as reticências não são consideradas parte integrante da entidade. Em (7), por exemplo, o contexto indica que "Bove" se refere a "Bovespa", sendo então anotada como ENTIDADE=S-ORGANIZACAO-EMPRESA.

(7) RT @Rede45: Graça Foster segue tentando explicar prejuízo da Petrobras em Pasadena e ação PETR4 tem queda de quase 5% neste momento na **Bove...**

As URLs truncadas, desde que sejam reconhecíveis como tais, recebem o rótulo ENTIDADE=S-LOCAL-VIRTUAL, como ocorre em (8).

(8) RT @daltonvieira : Ações ex-dividendos hoje : ARZZ3 , ELPL , EQTL3 , GETI , GFSA3 , IMCH3 , JSLG3 , KEPL3 , MAGG3 e MILS3 . Cotações ajustadas ! **http:/...**

As *hashtags* e menções truncadas que contenham apenas o prefixo visível ("#" ou "@") e cuja identidade não possa ser determinada com segurança, como em (9), não são anotadas.

(9) RT @Smarttrade10: Ontem o dia foi de tomarem #BBDC3 e #BBDC4 Outros destaques do BTC: #PETR3 #CMIG4 #SUZB5 R$ em tx: #BBDC4 #RENT3 #OIBR4 @...

As variações de valor em que apenas o símbolo inicial está presente, sem o número correspondente, como "+..." em (10), também não são anotadas.

(10) RT @garimpodeacoes: Conforme esperado ações de estatais sobem qdo Dilma cai em as pesquisas: BBAS3 +3,07%, PETR3, +2,30%, PETR4, +2,29%, ELET3, **+**...

As variações truncadas que não apresentam a casa decimal do número ou o sinal de porcentagem, como "+2," em (6) são anotadas da seguinte forma segundo o esquema BIOES, com *token* "+" recebendo o rótulo "B-" e "2," recebendo "E-", como também ilustrado pela Figura 8.

4.2.3 Contrações

Em decorrência das normas de *tokenização* para a anotação-UD do *corpus*, as contrações formais foram decompostas. No *tweet* (11), por exemplo, ocorre a contração "neste", decomposta na preposição "em" (ADP) e no determinante "este" (DET), como ilustrado no arquivo CoNLL-U da Figura 9. Nele, observa-se que a contração é mantida em uma linha própria, com indicação dos *tokens* que a constituem (12-13).

Nos casos em que a contração ocorre no início de uma EN temporal, como "neste mês" em (11), ela em si (linha 12-13) permanece sem anotação, assim como a preposição (ADP).No caso, a EN temporal é composta pelo determinante (DET), que recebe o rótulo "B-", e pelo nome (NOUN) "mês", anotado com "E-".

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

Figura 9 – Anotação de EN com contração em posição inicial.

Id	Form	Lemma	UPoS	XPoS
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.

Se a contração ocorrer no interior de uma EN, como em "Bando do (de o) Brasil" (12), tanto a contração (linha 2-3) quanto seus elementos constitutivos (ADP e DET) recebem o rótulo "I-".

(12) **Banco do Brasil On** (Bbas3) , Gráfico Diário. Ativo... http://t.co/b4pvQInln7

Figura 10 – Anotação de EN composta por contração em posição intermediária.

Id	Form	Lemma	UPoS	XPoS
1	Banco	Banco	PROPN	ENTIDADE=B-ORGANIZACAO-CERTIFICADO
2-3	do	_	_	ENTIDADE=I-ORGANIZACAO-CERTIFICADO
2	de	de	ADP	ENTIDADE=I-ORGANIZACAO-CERTIFICADO
3	o	o	DET	ENTIDADE=I-ORGANIZACAO-CERTIFICADO
4	Brasil	Brasil	PROPN	ENTIDADE=I-ORGANIZACAO-CERTIFICADO
5	On	On	PROPN	ENTIDADE=E-ORGANIZACAO-CERTIFICADO

Fonte: A autora, 2025.

4.3 Diretrizes específicas de classificação

Nesta seção, apresentam-se as diretrizes de classificação das ENs organizadas por categoria e tipo. São abordados exclusivamente os tipos observados no *corpus*.

4.3.1 ABSTRAÇÃO

- **DISCIPLINA**

  Abrange, no geral, disciplinas científicas, teorias e práticas consolidadas em diferentes áreas do conhecimento. No *corpus* DANTEStocks, o tipo abarca estratégias de compra e venda de ações, métodos de *trade*, estratégias de investimento e métodos de análise gráfica aplicados ao mercado financeiro.

  (13) @Live_Trade Marcos, qual o objetivo **daytrade** de OIBR4? Abs

  (14) Infelizmente não conseguir montar a **Streaddle** com opções da #PETR4 queria ter montado entre sexta e hj, agora fica pra próxima..

  Quando frequentes, os modificadores pré-nominais que contribuem para identificar ou especificar a natureza da disciplina devem ser incluídos na delimitação da entidade. Termos como "*análise*", "*estratégia*", "*guia*", entre outros, devem ser considerados parte integrante da expressão:

  (15) #VALE5 - **Análise #Ichimoku** - pregão de quarta-feira, 21 de maio. http://t.co/hWULGbJ1ps

- **ESTADO**

  Representa estados físicos, condições ou funções, tais como doenças. Na delimitação desse tipo devem ser incluídos termos que expressam a noção de estado ou condição, como *"doença"*, *"mal"*, *"síndrome"* e *"estado"*.

  (16) #JBSS3 ignorando **VACA LOUCA** , diário de a JBSS3 é bem interessante pois consolida com possível fundo triplo entre médias de 144 e 200.

  (17) #jbss3 vai p/o brejo amanhã RT @EstadaoEconomia: Governo confirma caso 'atípico' do **mal da vaca louca** em Mato Grosso http://t.co/bccczFxEUn

- **IDEIA**

  Abrange entidades que representam conceitos abstratos. Nos exemplos (18) e (19), as expressões "Mercado" e "mão invisível" ilustram esse tipo, pois remetem a ideias e princípios relevantes para o domínio analisado.

  (18) BBAS3 , ITUB4 e BBDC4 sofrem valorização até o momento . O **Mercado** não está se deixando influenciar por as ag. de ratings como antigamente

  (19) #petr4 abriu as porteiras, agora passa uma boiada! O elástico começou a esticar, a **mão invisível** logo logo vai entrar pesado em esse papel!

- **NOME**

  Esse tipo é utilizado quando o foco do enunciado recai sobre o objeto linguístico em si, isto é, o nome enquanto forma lexical ou referência discursiva, e não sobre a entidade real a que ele se refere. No exemplo (20), "Graça" aparece entre aspas (simples) e é usado de forma irônica, como recurso estilístico, sem remeter diretamente à pessoa Graça Foster.

  (20) Concluir q as coisas n #Petr4 estão feias não é difícil é só ver a **'Graça'** do Sr. Goonies Sloth Cerveró @clubedopairico http://t.co/Ns5fdPB4aR

4.3.2 ACONTECIMENTO

- **EFEMÉRIDE**

  Acontecimento ocorrido no passado e não repetível, ou seja, é único em seu contexto histórico.

  (21) **CASO LAVA JATO** - Youssef pode ter atuado em refinaria da Petrobras no Paraná http://t.co/6upiN3J6nR #PETR3 #Petr4 #**LavaJATO**

- **ORGANIZADO**

  Acontecimento multifacetado, que poderá durar vários dias, e geralmente conter vários eventos.

  (22) #ABEV3 - Ambev vai manter preço da cerveja sem aumento até o fim da **Copa** - http://t.co/uQvcMid5PG

  (23) 18 ações passam por 'ressaca' após **carnaval** e fecham em queda, Vale recua 3 %: Mesmo com proj... http://t.co/GNveU1iOgh #infomoney#vale5

- **EVENTO**

  Trata-se de um acontecimento pontual, organizado ou não. No DANTEStocks, esse tipo inclui eventos como assembleias gerais ou extraordinárias, reuniões e demais eventos relevantes para o mercado financeiro.

  No exemplo (26), o *token* "**Ago**" se refere à "Assembleia Geral Ordinária da Eletrobras" e, por esse motivo, anotado com a etiqueta ENTIDADE=S-ACONTECIMENTO–EVENTO. Quando o evento vier acompanhado de expressões temporais (como datas e horas), essas não devem ser incluídas na delimitação da entidade ACONTECIMENTO. Nesses casos, as informações temporais devem ser anotadas separadamente, conforme as diretrizes específicas da categoria TEMPO.

  (24) Antecipação de o debate político gera clima de **'Fla-Flu'** no mercado, diz CEO da Vale: Murilo F... http://t.co/Kjjf3dE5HM #infomoney #vale5

  (25) **Mega-Sena**: o número do recibo da sua declaração de Imposto de renda tem 6 dígitos joga na **Mega Sena**

  (26) $ELET3 - Eletrobras (elet-n1) - **Ago** - 30/04/2014 - 14h00 - Jcp http://t.co/ErDMSB1oGg

4.3.3 COISA

- **OBJETO**

  Abrange entidades que podem ser individualizadas e referenciadas como itens únicos. No DANTEStocks, esse tipo inclui, por exemplo, padrões gráficos específicos como "OCO" (isto é, "Ombro-Cabeça-Ombro") e "OCOI" (ou seja, "Ombro-Cabeça-Ombro Invertido"), utilizadas na análise técnica de ativos financeiros.

  (27) #BBAS3 **OCO** no intra, quem gosta das vendinhas... http://t.co/lEZSSC2mb5

  (28) fundo duplo ALSC3, **OCOI** DAYC4, LLIS3, EMBR3

- **CLASSE**

  Abrange uma população de objetos, tais como marcas ou modelos, assim como raças de animais ou programas de computador. No contexto do *corpus* DANTEStocks, esse tipo também é utilizado para categorias de representação gráfica recorrentes em análises técnicas do mercado financeiro, como é o caso dos gráficos "*candlestick*" (29), que se refere a um tipo ou forma de representação/visualização gráfica.

  (29) SANB11 gerou um engolfo de baixa em o dia 28. O que significa e como identificar esse padrão dos **candlesticks**. Assista ! http://t.co/E9Jvr5nThH

(30) Vale diz que acesso de **Valemax** a a China é desejável , mas não necessário : Os navios , com capac ... http://t.co/DUDYluUL6A #infomoney #vale5

(31) **Boeing** novinho em a frota de a gol ! ! ! Está valendo a pena investir em goll4 ! Rsrs parabéns @VoeGOLoficial http://t.co/MXGHUqK3db

- **MEMBROCLASSE**

  Refere-se a elementos que não possuem nome individual ou identidade única, mas são reconhecidos por pertencerem a uma classe ou categoria mais ampla. Esses elementos compartilham propriedades com outros membros da mesma classe e são identificados justamente por essa relação de pertencimento.

(32) @ppaulovagner @Fontes_ Hj a tarde recebi relatório em **PDF** de a corretora @Citi recomendando compra de #USIM5, fica de olho Paulo .

- **CERTIFICADO**

  Esse tipo reúne 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 (33). Inclui os papéis negociados no mercado financeiro, considerando suas classificações em ações ordinárias (ON), preferenciais (PN) e suas subclasses (como PNA, PNB) (34-35). Inclui também o segmento da B3 voltado a empresas com elevado padrão de governança corporativa, que é o Novo Mercado (-nm) (35), assim como os níveis diferenciados de governança corporativa, como Nível 1 (-n1) (36).

(33) **ADR** da PETR4 , $PBR testando forte resistência agora em $ 12,55 . Topo triplo !

(34) Fechamento das estatais na Bovespa: PETR3, +7,54%, PETR4, +8,12%, Banco do Brasil, +6,63%, **Eletrobrás ON**, +9,84% e **Eletrobrás PNB**, +3,51%

(35) $BBAS3 - Brasil (**bbas-nm**) - Demonstracoes Financeiras De 31/12/2013 (individual) http://t.co/N0DE4XbwLX

(36) $VALE3 - Vale (**vale-n1**) Ago/e 17/04/2014 11h00 http://t.co/yYZR3MNNu2

- **INDICADOR**

  No *corpus* DANTEStocks, esse tipo abrange índices, indicadores, taxas, métricas financeiras e câmbio de moedas, utilizados para descrever, medir ou avaliar o desempenho de ativos, mercados ou operações econômicas. Destaca-se que os **índices** são medidas agregadas que representam uma média ponderada de ativos ou dados, refletindo o desempenho geral de um setor, mercado ou economia. Exemplos são “IBOV” em (37), que mede o comportamento das ações mais negociadas na B3, e “DJI” em (39), que é o índice de ações dos Estados Unidos. Os **indicadores**, por sua vez, correspondem a métricas específicas utilizadas para avaliar aspectos técnicos ou financeiros de empresas ou ativos individuais, como “P/L” (“Preço sobre Lucro”) e “P/VPa” (“Preço sobre Valor Patrimonial por Ação”).

(37) ELET3 em alta mesmo com o **IBOV** em leve queda. http://t.co/YTsSOjXwKx

(38) Uma ação com **P/L** de 1,80 e um **P/VPa** de 0,5 deveria despertar interesse em qq investidor, o problema é q é: #GFSA3

(39) Curso de aplicação em a bolsa de valores Probabilidades de a PETR4 VALE5 **DJI** e **EURUSD** 16 3 14 : http://t.co/6FJwSFd7d3 via @YouTube

- **TICKER**

  O tipo TICKER abrange os códigos alfanumérico cujas letras indicam a empresa e o número é o tipo da ação, como nos exemplos (40-42). Eles funcionam como rótulos padronizados para facilitar a identificação dos papéis no mercado. No *corpus* DANTEStocks, os *tickers* podem aparecer sozinhos (40) ou precedidos dos sinais de *hashtags* (41) ou *cashtags* (42).

  (40) Confira as análises atualizadas de **BVMF3**, **ITUB4**, **PDGR3**, **PETR4**, **USIM5** e **VALE5** no INVISTA EM AÇÕES http://t.co/YWnlY5cM6J

  (41) **#cmig4** Pelo o que vi segunda e terça cmig está inclinada para alta.

  (42) **$LIGT3** - Light S/a (**ligt-nm**) - Alteracao Do Calendario De Eventos Corporativos http://t.co/XjNJgpnqcw

4.3.4 LOCAL

- **FÍSICO**

  Esse tipo inclui localizações de geografia física que apenas foram batizadas pelo Homem e não construídas.

  (43) RT @mfukayama: Em 5 anos, as ações de a SABESP (SBSP3) subiram 109,40%. Em esse mesmo período, a**Reserva da Cantareira** desabou d... http://t.c…

  (44) #OIL #BR #PETR4 Petrobras põe plataforma P-58 em operação na **Bacia de Campos**: http://t.co/VRlzpXUGF7

- **HUMANO**

  Designa localizações criadas e/ou delimitadas pela ação humana, como países, cidades, regiões e construções, tais como edifícios, portos, barragens, entre outras.

  (45) $PETR3 - Petrobras (petr) - Prod De 40 Mil Barris Dia **Cascade E Chinook** - Reapresentacao http://t.co/RBAlmZx1OK

  (46) Foco de empresas do **Brasil** em eficiência passa por menos investimento em 2014: Até agora ao... http://t.co/GziH575KhU #infomoney #vale5

- **VIRTUAL**

  Uma entidade é anotada com o tipo VIRTUAL quando não representa um local físico, ou quando se refere a um "alojamento". Isso abrange todos os meios de comunicação social, como jornais, televisão, rádio e referências a obras impressas. Diferentemente do Segundo HAREM e (Zerbinati; Roman, 2023; Zerbinati; Roman; Di-Felippo, 2024), que não consideraram as URLs, o tipo VIRTUAL deste manual inclui todos os locais virtuais no contexto eletrônico. Isso se deve ao fato de que as URLs são muito frequentes nos *tweets* e, por isso, deixar de anotá-las como EN pode resultar na perda de informações referenciais importantes, comprometendo a capacidade dos sistemas de PLN de rastrear fontes de informação ou mesmo desambiguar entidades.

  (47) Hora de comprar mais PETR4 RT **@JornalOGlobo** : Ações de a Petrobras caem mais de 3 % e derrubam Bolsa . **http://t.co/LJluyesRk5**

4.3.5 OBRA

- **REPRODUZIDA**

  Tipo de obra que contém que muitas cópias ou exemplares.

  (48) $GOLL4 - GOL Arquiva Formulário **20-F** de 2013 na SEC http://t.co/rLgwcouE7D

  (49) Se o **Relatório** sumiu ou se encontra adulterado, estaremos claramente diante de uma ação criminosa e mafiosa, corroendo por dentro a #PETR4.

- **PLANO**

  Tipo que abrange medidas políticas, administrativas e/ou financeiras, assim como projetos ou acordos.

  (50) **Plano de demissao voluntária** na Petrobras já atrai 8 mil ... Depois vao reclamar que nao tem especialistas ... petr4 virando OGX

  (51) Governo cedeu e **MP 627** será alterada. A medida bitributava as empresas que abriam filais no exterior. Positivo para a #VALE5 e outras.

4.3.6 ORGANIZAÇÃO

- **ADMINISTRAÇÃO**

  Refere-se a organizações responsáveis pela administração e governança de territórios em níveis nacional, internacional ou supranacional. No DANTEStocks, essa categoria também inclui entidades que exercem funções de gestão e controle em empresas e instituições.

  (52) O **Conselho de Ministros** discute e aprova Propostas de Lei e pedidos de autorização legislativa (autorização para fazer leis) à **Assembleia da República**.

  (53) $BBAS3 - É hoje: **STF** julga bancos nacionais [Newsletter ADVFN] http://t.co/R8T7JNyVG4

- **EMPRESA**

  Abrange organizações com fins lucrativos.

  (54) **Cielo** ultrapassou o **Banco do Brasil** em valor de mercado. Topo histórico de um lado e desvalorização de 10 % em 2013, de o outro. #CIEL3 #BBAS3

  (55) $PETR3 -**Petrobras** (petr) - Comunicado http://t.co/mHuCIyQmFi

- **INSTITUIÇÃO**

  Abrange organizações sem fins lucrativos. Esse tipo inclui partidos políticos.

  (56) O incidente aconteceu na noite de 3 de Janeiro e está a ser investigado pela **Polícia Judiciária**.

  (57) @piacesiramos desvenda o que está por trás de PETR4 a $13,13! Propaganda eleitoral antecipada do **PT**, eh eh eh...

Diretrizes gerais de classificação para a categoria ORGANIZAÇÃO:

- No DANTEStocks, setores internos de uma organização, como comissões, comitês, assembleias gerais, departamentos e seções, não são anotados automaticamente com o mesmo tipo da organização a que pertencem. Ao contrário do que é adotado no Segundo HAREM (Santos *et al.*, 2008), comissões e comitês associados à gestão organizacional são classificados como ADMINISTRAÇÃO, enquanto assembleias, por seu caráter pontual (muitas vezes marcado no tempo), são anotadas como ACONTECIMENTO-EVENTO.

- Embora nomes de países, cidades e afins geralmente designem locais, em determinados contextos a menção a esses nomes implica uma referência ao seu governo. Nesses casos, a entidade deve ser classificada com ORGANIZAÇÃO-ADMINISTRAÇÃO.

  (58) #PETR4 RT @kitowgallo : PGR manda para o **Rio** pedido para apurar suposto suborno em a Petrobras : Procuradoria de a Repúb ... http://t.co/kDSqPTbzcD

4.3.7 PESSOA

- **CARGO**

  Representa um posto que é ocupado por uma pessoa, mas que poderá no futuro ser ocupado por outros indivíduos.

  (59) Antecipação do debate político gera clima de 'Fla-Flu' no mercado, diz **CEO da Vale**: Murilo F... http://t.co/Kjjf3dE5HM #infomoney #vale5

  (60) Por ordem de Dilma, Mantega fará reunião com 'gigantes' do empresariado: **Ministro da fazenda**... http://t.co/wEYR2pNDqY #infomoney #vale5

- **GRUPOCARGO**

  Representa um conjunto de indivíduos, através de um cargo.

  (61) #PETR4 RT @agenciabrasil: Petrobras: **ministros da Justiça** e da CGU vão à Câmara para falar sobre denúncias http://t.co/zhV2Gty8Ny

- **INDIVIDUAL**

  Todo indivíduo, inclusive o título que delimita essa pessoa. Inclui diminutivos, alcunhas, iniciais, nomes mitológicos e entidades religiosas.

  (62) RT @reminiscences: A partir das 14h vamos ver o voto da **Ministra Rosa Weber**... CPI da Petrobras exclusiva ou escancarada... Petr4 sobe com CPI?

  (63) Coitada da **Graça Foster**. Tem o feeling, conhece a @petrobras, mas tem q fazer o q o governo manda. E ele manda ruim. #PETR4

- **GRUPOIND**

  Esta categoria representa grupos de indivíduos que não possuem um nome "estático" ou institucionalizado como grupo. Nos exemplos, "governo Dilma" e "Gov FHC" referem-se a coletivos ligados a figuras políticas específicas. Nesse caso, os grupos são reconhecidos contextualmente pela liderança que os define, e por isso se enquadram no tipo GRUPOIND.

  (64) Ações da Petrobras (PETR4) disparam +4% com queda de popularidade do **governo Dilma**.

  (65) RT @midiacrucis: último dia do **Gov FHC**, a ação Petrobras (PETR3) $>$ R$ 3,30 @geraldoAlckmin_ diz q ações da Petrobras "viraram pó". http://t...

- **MEMBRO**

  Indivíduo é mencionado pela organização que representa.

  (66) **Ex-Itaú** na Marisa, **ex-TAM** na Via Varejo, Petrobras, Vale e imobiliárias agitam a noite: Petr... http://t.co/PoqQgsaVIY #infomoney #vale5

- **GRUPOMEMBRO**

  Abrange ENs que se referem a um conjunto de pessoas como membros de uma organização ou conceito semelhante.

  (67) BBSE3 - 'É muito cedo para basear a recomendação no resultado das eleições', diz **BTG** - InfoMoney Veja mais em: http://t.co/thncBwA5YP

  (68) #PETR4 é do **Povo Brasileiro**, mas está subjugada por rapina @MirandaSa_: A Petrobras é Nossa? http://t.co/VeNKy7BAZV #EuApoioCPIdaPetrobras

- **POVO**

  Refere-se a casos em que um local ou grupo de pessoas é personificado para descrever ações coletivas.

  (69) **China** deve puxar queda de Vale, OGX, Marfrig, TIM e mais 6 estão no radar: Marfrig tem melho... http://t.co/vbpAbwH4Ax #infomoney #vale5.

  (70) É hoje que a #PETR4 dá adeus ao fundo de 2008? Bom, pelo menos já temos a **Rússia** pra culpar né?

- **USUÁRIO**

  Neste tipo, estão os usuários representados pelas @menções (*at mentions*). As menções a perfis devem ser anotadas de acordo com a entidade no mundo extralinguístico que representam. Se uma menção diz respeito a uma organização, ela será classificada como ORGANIZAÇÃO e seu tipo correspondente, como por exemplo @petrobras; caso se refira a um canal de comunicação, será rotulada como LOCAL-VIRTUAL, como @RevistaEpoca. Menções que não se encaixam nessas categorias serão anotadas como PESSOA-USUARIO.

  (71) **@Live_Trade** Marcos bom dia! pode falar sobre oibr4.

  (72) **@PaiRico @frfontanella @eddu56 @TiagoBDS** Acabei de fazer análise de a sua #petr4. Abraços!

  Atenção: o símbolo @ nem sempre indica um usuário, às vezes, é usado apenas para destacar uma palavra ou expressão, como no exemplo (73).

  (73) Elet3 minha linda!!! Hoje vou beber um **@vinho** para comemorar!!!

Diretriz de delimitação para a categoria PESSOA:

- Palavras específicas escritas com letras minúsculas — como aquelas que indicam cargos, funções, títulos ou pronomes de tratamento — são anotadas como parte integrante de uma EN, desde que acompanhadas de um elemento que identifique claramente a entidade. Termos como "presidente", "ministro(a)", "executivo-chefe", "diretor", "senhor", "senhora", "Sr.", "Sra.", entre outros, podem ser considerados parte de uma EN quando aparecem junto a nomes próprios ou outros identificadores, como em "ministros da Justiça" ou "presidente da vale". Além disso, essas palavras podem ser precedidas de modificadores como "ex-", "vice-", "co-" ou "sub-" (p.ex.: "ex-ministro da Fazenda").

4.3.8 TEMPO

- **DURAÇÃO**

  Correspondem a expressões de quantificação temporal. Essas entidades respondem à interrogativa: “quanto tempo?”.

  (74) @andremassaro vc acha que comprar petr4 pra LP (**20-30 anos**) é bater palma pra louco dançar ou dá pra ter esperança na nossa PDVESA...

  (75) @Smarttrade10 @ferrisss @dfittarelli Volume em PETR4 bom hj, 45 % superior ao volume medio dos últimos **7 dias**.

- **FREQUÊNCIA**

  Expressões que indicam repetição de eventos ao longo do tempo e que respondem a interrogativas do tipo: “com que frequência?”.

  (76) **Todos os dias** no fechamento do mercado #ITUB4 mesmo caindo tem uma puxada no pregão. Estranho, dia todo em baixa e pregao puxando pra cima.

  (77) #ELPL4 acho que está armando um COMPRÃO como **poucas vezes** já visto. #deolho

- **GENÉRICO**

  Corresponde ao uso genérico de expressões relacionadas à noção de tempo, mas que não fazem referência a uma data específica. Nessas ocorrências, não é possível ancorar o predicado em uma linha do tempo de forma que responda adequadamente às interrogativas “quando?", “por quanto tempo?", “há quanto tempo?"ou “com que frequência?".

  (78) PETR4 famoso: o que te fazia rir, o **hoje** já nem faz mais.

  (79) @ferrisss lembro que o @Fontes_ duvidou de 9,XX p/ #PETR4. Acho que **hoje**, ninguém mais duvida.

  No exemplos (78) e (79), “hoje” não remete a um ponto específico no tempo, mas funciona como uma noção genérica do momento presente.

- **TEMPO_CALEND**

  As expressões que permitem posicionar ou ancorar o predicado que modificam ao longo de uma linha temporal, atuando como pontos específicos ou intervalos definidos no tempo. Esse tipo inclui **datas** que podem ser absolutas, expressas no formato dia, mês e ano (80), com possibilidade de omissão de até dois desses elementos, como o exemplo (82), ou referenciais, cuja interpretação depende do momento da enunciação ou de um evento temporal previamente mencionado (81-82).

  (80) VALE5 em **28/04/2014**: Ola Olhada rápida em VALE5, gráfico diário. O ativo até estava se saindo bem depois da pern... http://t.co/PbH8HnvcDi

  (81) OPORTUNIDADE #PETROBRAS: O MENOR PREÇO EM **20 ANOS**? #PETR4 Na **semana passada**, vários indicadores de péssima... http://t.co/o57yNlV4WC

  (82) @fabiokraemer @jprcampos Comprem ações #PETR4 **hoje** a R$ 13 vendida para o pequeno acionista por R$ 26,30 em **2010**. Antes comprou a R$ 29,31.

  Também fazem parte desse tipo as **horas** e os **intervalos**, que representam durações temporais com limites claramente definidos, ilustrados respectivamente em (83) e (84).

(83) Rastreamento ações - Gráfico diário - **16h** Analise se romper: AEDU3 13,70 BRML3 19,19 EMBR3 20,46 ESTC3 22,25 ESTC3 22,25 HGTX3 27,53

(84) RT @juniormayhe: @LidPSDBsenado e o PT acabando com a Petrobrás RT @ale_chumer: Desde **24/04/2009 até hoje** Ibovespa +11,58% PETR4 -47,7% DJ...

Datas seguidas de horas, conforme ilustrado no exemplo (85), constituem uma única referência temporal do tipo TEMPO_CALEND, uma vez que essas expressões representam uma única referência temporal integrada e devem ser anotados como pertencentes a uma única entidade. Além disso, expressões específicas do domínio financeiro que sinalizam marcações temporais, como "INTRADAY" (86), também devem ser anotadas.

(85) JBSS3 comprar a R$ 7.57 indicado em **18/03/2014 15:23** e finalizou a venda com resultado de R$ -0.13 ou -1.72% http://t.co/kgt1YiTbF7

(86) **INTRADAY** PETR4: Suportes 13,67 e 14,04 e resistências 14,60 e 14,79 **INTRADAY** VALE5: Suportes 26,71 e 26,93 e resistências 27,37 e 27,59

Diretrizes de delimitação para a categoria TEMPO:

- **Delimitação das entidades temporais:** No DANTEStocks, o critério de delimitação consiste em identificar a menor expressão capaz de representar a entidade sem perda de sentido. Em outras palavras, a EN será a expressão temporal mínima necessária para a interpretação. Assim, no exemplo (87), anota-se como entidade temporal apenas "ontem" e não " dia de ontem".

  (87) 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

- **Separação de entidades temporais complexas**

  Tomando como exemplo o *tweet* (88):

  (88) #VALE5 - Análise #Ichimoku - pregão de **sexta-feira, 28 de fevereiro**. http://t.co/kEreB1xgU4

  Uma expressão temporal complexa pode ser dividida em unidades menores, desde que sejam atendidos simultaneamente os seguintes critérios:

  1. Cada expressão componente deve ser sintaticamente válida quando combinada de modo independente com o evento que modifica.
  2. Cada combinação "evento + expressão temporal mínima"pode ser deduzido da expressão "evento + expressão temporal complexa".

  No exemplo (88), observar-se que as expressões "sexta-feira" e "28 de fevereiro" são sintaticamente válidas de forma independente quando associadas ao evento "pregão":

  (89) #VALE5 - Análise #Ichimoku - pregão de sexta-feira. http://t.co/kEreB1xgU4
  #VALE5 - Análise #Ichimoku - pregão de 28 de fevereiro. http://t.co/kEreB1xgU4

  Embora cada expressão mínima seja sintaticamente válida quando combinada com o evento "pregão", nenhuma delas, de forma independente, é logicamente implicada pela expressão complexa "pregão de sexta-feira, 28 de fevereiro". Isso ocorre, porque a expressão complexa faz referência a um único ponto específico no tempo, isto é, a sexta-feira que coincide com o dia 28 de fevereiro, ou seja, não é possível deduzir, a partir da expressão mínima, que o evento também ocorreu isoladamente em todas as sextas-feiras ou em todos os dias 28 de fevereiro.

4.3.9 VALOR

- **CLASSIFICAÇÃO**

  Esse tipo refere-se a classificações desportivas, ordinais normais e outras. Enumerações de parágrafos, tópicos e outras secções não devem ser etiquetadas.

  (90) #vale5 (mensagem: **950904**) http://t.co/wfR8HEPu4k

  (91) #CVCB3 registrou hoje seu **oitavo** pregão de alta enquanto #GOAU4 acumulou sua **sexta** baixa seguida. #LIGT3 #CMIG4 #ALPA4 com 5 altas!

- **QUANTIDADE**

  Corresponde a quantidades absolutas ou relativas, abrangendo percentuais, números isolados e, quando aplicável, suas respectivas unidades de medida. As unidades consideradas na delimitação desse tipo de entidade referem-se a propriedades físicas, como distância, tempo, luminosidade, área, volume, peso e massa, e não a itens contáveis. Assim, no *tweet* (92), apenas o numeral "4" é anotado como entidade. O mesmo critério se aplica ao exemplo (93), em que apenas "70,2 milhões" é incluído na anotação da entidade, enquanto "de ações" não é considerado parte dela.

  (92) Nº de papéis tomados de BBAS3 subiu **45%** nos últimos **4** pregões.

  (93) CSN boa oportunidade com o anúncio de recompra de **70,2 milhões** de ações pela empresa. Valeu ficar bem de olho. #CSNA3

  (94) Estatais brasileiras estão entre as maiores alta do dia: Banco do Brasil **+4,32%**, PETR3, **+3,64%**, PETR4, **+3,80%**, ELET3,**+3,23%** e ELET6, **+2,84%**

  Quando uma entidade do tipo COISA–INDICADOR estiver acompanhada de um valor que expressa sua medida, esse valor deve ser anotado como VALOR-QUANTIDADE. No exemplo (95), os indicadores "Dividend yield", "P/VP" e "PL" são anotados como COISA–INDICADOR, e seus valores "11,9%", "0,77" e "3,4" como VALOR-QUANTIDADE. O mesmo vale para o exemplo (96). Os índices "INDX" e "IVBX" são COISA–INDICADOR, e seus valores numéricos, VALOR-QUANTIDADE.

  (95) **Dividend yield** de BBAS3: **11,9%**. P/VP, **0,77**. PL, **3,4**. Inacreditável... http://t.co/eg4c8sGhKW

  (96) Rastreamento ações - Gráfico diário - 11h. Analise se romper: CPFE3 18,73 ECOR3 13,70 HYPE3 16,31 **INDX 11479,53 IVBX 7026,11** LIGT3 18,92

- **MOEDA**

  Abrange valores monetários, devendo incluir a unidade monetária caso esteja explicitamente presente no texto.

  (97) #ecor3 fechando **15** acima **12,53** já fica interessante

  (98) 07/03/2014 - 17:19: Maiores Altas : HYPE3 1,99% **R$ 14,81**, VIVT4 1,28% **R$ 44,76**, TIMP3 0,78% **R$ 11,61**, NATU3 0,28% **R$ 35,80**, UGPA3 0,24% **R$ 52,38**.

Diretriz de delimitação para a categoria VALOR:

- Modificadores, como preposições, quantificadores e expressões como "mais de", "menos de", "aproximadamente", "quase", entre outras, não são considerados na delimitação da entidade. Observando os exemplos a seguir, apenas as expressões destacadas em negrito, respectivamente "100%", "R$ 12 milhões" e "6%", fazem parte da delimitação.

(99) $ESTC3 - ESTÁCIO (ESTC3): Lucros disparam em mais de **100%** http://t.co/H8GtUMZow0

(100) ah, e quase que eu ia me esquecendo... Lembram da 'inominável'? Pois bem... já comprou mais de **R$ 12 milhões** de #PETR4 no dia de hoje...

(101) Goll4 quase **6%** de alta hj com a bolsa caindo, piores, CYre3 -5,21 e Rsid3 -4,86 até o leilão

5 Agradecimentos

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.

Os autores deste trabalho agradecem 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.

Referências

CARDOSO, N.; SANTOS, D. Directivas para a identificação e classificação semântica na colecção dourada do harem. *In*: SANTOS, D.; CARDOSO, N. (ed.). **Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área**. [*S.l.: s.n.*]: Linguateca, 2007. p. 211–238. Documento original publicado no sítio do HAREM a 29 de Março de 2006. Republicado como Relatório técnico DI-FCUL TR-06-18: Departamento de Informática, Faculdade.

DI-FELIPPO, A.; ROMAN, N. T. DANTEStocks: a multi-layered annotated corpus of stock market tweets for Brazilian Portuguese. **Brazilian Journal of Applied Linguistics**, Corpus Linguistics: Studies and Applications, p. 1–23, 2025. To appear.

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. BPLN, 2024. book chapter 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.

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.

HAGèGE, C.; BAPTISTA, J.; MAMEDE, N. Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O segundo harem. *In*: MOTA, C.; SANTOS, D. (ed.). **Apêndice B**. [*S.l.: s.n.*]: Linguateca, 2008. p. 289–308.

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

MARNEFFE, M.-C. de *et al.* Universal Dependencies. **Computational Linguistics**, MIT Press, Cambridge, MA, v. 47, n. 2, p. 255–308, jun. 2021. Disponível em: https://aclanthology.org/2021.cl-2.11/.

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. Disponível em: https://www.sciencedirect.com/science/article/pii/S0920548912001080.

MOTA, C.; SANTOS, D. Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O segundo harem. *In*: . [*S.l.: s.n.*], 2008.

NIVRE, J. *et al.* Universal Dependencies v2: An evergrowing multilingual treebank collection. *In*: CALZOLARI, N. *et al.* (ed.). **Proceedings of the Twelfth Language Resources and Evaluation Conference**. Marseille, France: European Language Resources Association, 2020. p. 4034–4043. ISBN 979-10-95546-34-4. Disponível em: https://aclanthology.org/2020.lrec-1.497.

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

SANTOS, D.; CARDOSO, N. **Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área**. [*S.l.: s.n.*]: Linguateca, 2007.

SANTOS, D. *et al.* Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O segundo harem. *In*: MOTA, C.; SANTOS, D. (ed.). **Apêndice A**. [*S.l.: s.n.*]: Linguateca, 2008. p. 277–286.

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**. [*S.l.: s.n.*]: SBC, 2023.

ZERBINATI, M. M.; ROMAN, N. T. **Manual de Anotação de Entidades Nomeadas do DANTEStocks utilizando categorias do Segundo HAREM**. São Paulo, SP, 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*: GAMALLO, P. *et al.* (ed.). **Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1**. Santiago de Compostela, Galicia/Spain: Association for Computational Linguistics, 2024. p. 276–284. Disponível em: https://aclanthology.org/2024.propor-1.28.