10.11606/issn.2317-9511.v46p6-30

# Termos do mercado financeiro: um estudo do *corpus* DANTEStocks

# Financial Market Terms: a study of the DANTEStocks *Corpus*

Roana Rodrigues[*]

Ariani Di Felippo[**]

Norton Trevisan Roman[***]

Pedro Semcovici[****]

Jackson Wilke da Cruz Souza[*****]

Thiago Alexandre Salgueiro Pardo[******]

*Resumo:* Neste artigo, são apresentados os procedimentos de extração e validação de termos do domínio do mercado financeiro em Português Brasileiro (PB) com base no *corpus* DANTEStocks. Para tanto, tem-se como pressuposto teórico a Teoria Comunicativa da Terminologia, que preconiza que os termos e suas propriedades só podem ser identificados e descritos no seu ambiente natural de ocorrência. Para a

* Universidade Federal de Sergipe (UFS)- roana@academico.ufs.br; https://orcid.org/0000-0002-7748-8716.

** Universidade Federal de São Carlos (UFSCar) - arianid@ufscar.br; https://orcid.org/0000-0002-4566-9352.

*** Escola de Artes, Ciências e Humanidades / Universidade de São Paulo (EACH/USP) norton@usp.br; https://orcid.org/0000-0002-0563-2045.

**** Escola de Artes, Ciências e Humanidades / Universidade de São Paulo (EACH/USP) - pedrosemcovici@usp.br; https://orcid.org/0009-0008-8455-8509.

***** Universidade Federal da Bahia (UFBA), no Instituto de Ciência, Tecnologia e Inovação (ICTI) - jackcruzsouza@gmail.com; https://orcid.org/0000-0003-1881-6780.

****** Universidade de São Paulo - taspardo@icmc.usp.br;https://orcid.org/0000-0003-2111-1319.

extração de candidatos a termos, foram aplicados padrões lexicais, resultando numa lista de 527 candidatos. Em seguida, os candidatos a termos foram analisados e validados por especialistas, culminando em uma lista de 380 termos. Além disso, fez-se a verificação em outros glossários do campo semântico da Economia, em que se constatou a ausência de muitos termos do mercado financeiro. Assim, considerando a relevância dos estudos terminológicos para a Linguística e o Processamento de Línguas Naturais, a lista terminológica construída no presente trabalho possibilita a identificação dos termos da área de domínio (mercado financeiro) e sua separação (e quantificação) em relação às palavras de língua geral.

*Palavras-chave:* Terminologia; Mercado financeiro; Recursos linguístico; Processamento de Línguas Naturais.

*Abstract*: In this article we present the procedures for term extraction and validation in the domain of the Brazilian financial market based on the DANTEStocks *corpus*. We adopted as theoretical framework the Communicative Theory of Terminology, which asserts that terms and their properties can only be identified and described within their natural occurrence context. We employe lexical features to extract term candidates, resulting in a list of 527 candidates. These terms were subsequently analyzed and validated by experts, leading to a final list of 380 terms. Additionally, we made a comparison with other glossaries in Economics domain, revealing the absence of many financial market terms. Considering the significance of terminological studies for Linguistics and Natural Language Processing, the constructed terminological list facilitates the identification and differentiation (including quantification) of domain-specific terms (financial market) from general language words.

*Keywords:* Terminology; Financial market; linguistic resources; Natural Language Processing.

# Introdução

Neste artigo, são apresentados os procedimentos de extração e validação de termos do domínio do mercado financeiro em português brasileiro (PB). É reconhecida a existência de obras terminológicas que abrangem, de alguma maneira, esse léxico, entre as quais pode-se citar o *Glossário de termos neológicos da economia* (ALVES 2001), o *Dicionário de Termos Financeiros e Bancários* (BIDERMAN 2013) e o *Glossário financeiro InfoMoney*[1]. No entanto, esta pesquisa se afasta das demais ao observar os termos (unidades terminológicas e unidades terminológicas complexas) em dados extraídos de um *corpus* de *tweets*, o DANTEStocks (DI FELIPPO *et al*. 2021; PARDO *et al*. 2021).

[1] Glossário *InfoMoney*. Disponível em: https://www.infomoney.com.br/glossario/. Acesso em janeiro de 2023.

Segundo os pressupostos gerais da Teoria Comunicativa da Terminologia (CABRÉ 1999), os termos (isto é, os signos que ocorrem como unidades terminológicas) e suas propriedades só podem ser identificados e descritos no seu ambiente natural de ocorrência, ou seja, nos discursos especializados. Dessa forma, esses princípios teóricos e metodológicos colocam em evidência a importância do uso dos *corpora* (fontes não-estruturadas) em qualquer trabalho terminológico (NASCIMENTO 2003; AGBAGO, BARRIÈRE 2005; CABRÉ *et al*. 2005; ALMEIDA 2006).

De acordo com Nascimento (2003), Barros (2004) e Cabré *et al*. (2005), pode-se, a partir dos *corpora*, fazer observações precisas sobre o real comportamento linguístico, proporcionando informações altamente confiáveis sobre os fatos de uma língua. Por meio de um *corpus*, é possível observar vários aspectos (morfológicos, sintáticos, discursivos etc.) relevantes para uma pesquisa linguística, o que possibilita a descoberta de fatos novos na língua, não perceptíveis pela intuição.

Para tanto, a terminologia tem historicamente estreitado seus laços com a área de Computação: por um lado, devido à utilização de ferramentas computacionais que auxiliam nas etapas de armazenamento, busca, extração e divulgação de termos na web (ALMEIDA; OLIVEIRA 2012); e, por outro, relacionado ao aproveitamento e aplicação de conhecimento terminológico em tarefas de Processamento de Língua Natural (PLN).

O PLN, como explicam Jurafsky e Martin (2008), visa a capacitar a máquina a lidar com as línguas humanas, realizando tarefas variadas, como tradução automática, sumarização de textos e recuperação e extração de informação. Essas tarefas, por sua vez, podem demandar recursos e ferramentas linguístico-computacionais diversos, como léxicos de língua geral e de domínios especializados, ontologias, gramáticas e *corpora*, por exemplo. Produzir e disponibilizar informação terminológica (em formato computacional) para a máquina mostra-se muito importante para auxiliá-la na estruturação e uso do conhecimento de domínio relevante para a realização de tarefas específicas de maneira mais satisfatória. Nesse sentido, Drouin *et al*. (2016) afirmam que é essencial para a mineração de textos as informações terminológicas para que haja descoberta científica e inteligência competitiva.

Há muitos usos e estudos de terminologia em PLN, os quais se enquadram na chamada Terminologia Computacional. Lopes *et al*. (2009), por exemplo, utilizam a extração de termos para auxiliar na construção de ontologias na área da saúde, que, como é bem conhecido na área, são a base de muitos processos computacionais. Seiffe *et al*. (2020) evidenciam o uso desse tipo de ontologia e sua importância, alinhando linguagem técnica e leiga na área de medicina. Fukutome e Harada (2018) discutem os desafios relacionados à tradução automática de termos no domínio de bebidas, lidando com a questão da adaptação cultural na tradução, que vai além da tradução literal. Em trabalhos mais recentes, Lang *et al*. (2021) utilizam a modelagem sofisticada dos *transformers* (baseados nas famosas redes neurais artificiais) para extração multilíngue de termos. Peng *et al*. (2022) vão além e utilizam os recentes modelos de língua baseados em *transformers* para descobrir relações entre termos de domínios especializados. Em particular, esses autores trabalham com o domínio financeiro, relacionado ao domínio do *corpus* estudado neste artigo.

É interessante notar o apelo que a modelagem de domínio especializado tem em PLN, mesmo que a área esteja em um momento de pesquisa e desenvolvimento em que o aprendizado automatizado a partir de grandes quantidades de dados seja a abordagem dominante. Evidência disso é que, recentemente, na esteira dos grandes modelos de língua, dos quais as versões do GPT (do inglês, *Generative Pre-trained Transformer*) são seus mais populares representantes, foi proposto um modelo customizado para o domínio financeiro, chamado de BloombergGPT (Wu *et al*. 2023), que produziu resultados melhores do que outros modelos em várias tarefas realizadas no referido domínio.

Vale também mencionar que esta pesquisa se insere no Projeto PoETISA[2] (POrtuguese processing - Towards Syntactic Analysis and parsing), projeto de longo prazo que visa ao desenvolvimento de recursos e ferramentas do estado-da-arte para o processamento sintático do PB. Sendo assim, e considerando a relevância dos estudos terminológicos para a Linguística e o PLN, a lista terminológica construída no presente trabalho possibilita a identificação dos

[2] Disponível em: https://sites.google.com/icmc.usp.br/poetisa. Acesso em janeiro de 2023.

termos da área de domínio (mercado financeiro) e sua separação (e quantificação) em relação às palavras de língua geral. Além disso, embora o levantamento de termos tenha ocorrido com base em um *corpus* de *tweets*[3], verifica-se que o seu uso apresenta restrições de compreensão por leigos, ou seja, por pessoas que não são especialistas da área. Nesse sentido, este trabalho também contribui com uma etapa inicial e necessária no que se refere às preocupações relacionadas à simplificação textual e, sobretudo, à acessibilidade terminológica (FINATTO 2022), ao buscar não só levantar os termos como também quantificar sua utilização em textos populares.

Sendo assim, neste artigo serão discutidas as seguintes questões:

- Quais os aspectos qualitativos e quantitativos da lista de termos extraída do *corpus* de tweets DANTEStocks?; e
- Quais as particularidades da lista de termos gerada em comparação com obras terminológicas já realizadas anteriormente?

Para a discussão dos pontos acima apresentados, o trabalho está organizado da seguinte maneira: na primeira seção é apresentada uma breve descrição do *corpus* adotado nesta pesquisa. Em seguida, são relatados os dados gerais, quantitativos e qualitativos, dos termos do mercado financeiro extraídos do referido *corpus*, além do processo de validação realizado por especialistas. Na seção três, são descritas as particularidades dos termos do DANTEStocks, considerando seus dados gerais e a sua comparação com outras obras terminológicas da economia e do mercado financeiro. Nas Considerações finais, são retomadas as questões de pesquisa, as contribuições dos dados criados e apontados os trabalhos futuros.

[3] Os *tweets*, atualmente *posts* X, são mensagens oriundas de uma plataforma de *microblogging* e são descritos como um gênero bastante popular e caracterizado por apresentar desvios no que diz respeito ao uso da norma padrão da língua (*cf*. EISENSTEIN 2013). Segundo Freitas e Barth (2015), o *tweet* enquanto gênero parece possuir resquícios de outros gêneros (como notícia, propaganda, bilhete etc.) que foram modificados para atender às necessidades comunicativas na rede.

# 1. A respeito do *corpus* de *tweets* DANTEStocks

A lista de termos do mercado financeiro brasileiro elaborada nesta investigação foi extraída a partir da análise manual do DANTEStocks, nome dado ao *corpus* de *tweets* contendo comentários feitos por pessoas envolvidas com o mercado financeiro, acerca de alguns dos papéis que compõem o iBOVESPA, principal indicador da bolsa de valores oficial do Brasil.

Esse *corpus* foi compilado a partir da coleta automática de postagens do Twitter[4] do ano de 2014 com base nos *tickers*[5] das ações que compõem o índice. Naquele momento, a compilação do *corpus* tinha como principal finalidade a anotação e estudo de emoções (SILVA; ROMAN; CARVALHO 2020). Em trabalhos posteriores, o mesmo *corpus* também foi anotado morfossintaticamente, com base nas diretrizes da *Universal Dependencies*, doravante UD (NIVRE 2015; NIVRE *et al*. 2020), no âmbito do projeto DANTE (*Dependency-ANalised corpora of TwEets*)[6], de onde deriva a sua atual denominação.[7]

Conforme apresentado por Di Felippo *et al*. (2021, 2022) com base em Nivre (2015), pode-se afirmar que, de maneira geral, a UD possui uma perspectiva lexicalista da sintaxe e é caracterizada por ser um modelo gramatical com diretrizes universais, possibilitando e facilitando a realização de estudos comparados entre diferentes línguas naturais. Especificamente, esse modelo prevê a anotação de um *corpus* em 2 níveis. No nível morfológico, especificam-se 3 informações: lema, etiqueta morfossintática (*part-of-speech* ou PoS) e traços lexicais/gramaticais (*features*). No nível sintático, a anotação

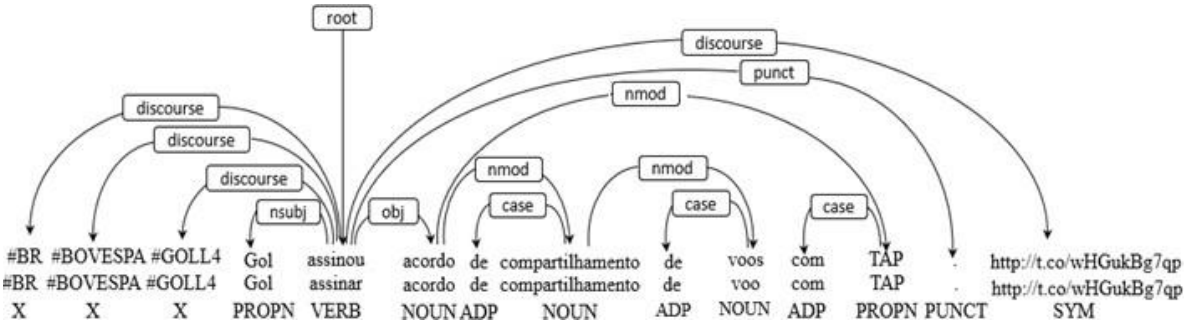
[4] Desde março de 2023, o *Twitter* passou a ser chamado de *X*; porém, foram mantidas as maneiras de publicação, repostagem e construção dos *tweets*, que passaram a ser chamados de *posts X*.
[5] Um ticker é um código alfanumérico que representa a empresa e o tipo de ação. O ticker VALE5, por exemplo, indica ações preferenciais da Vale do Rio Doce.
[6] Projeto de pesquisa que objetiva construir *corpora* de *tweets* anotados segundo a UD e outros tipos de anotações linguísticas.
[7] Salienta-se que a escolha de trabalhar nesta pesquisa com o *corpus* DANTEStocks se deu pelos seguintes fatores: (i) o *corpus* está disponível livremente; e (ii) o *corpus* já possui informações acerca da emoção percebida em cada *tweet* e etiquetas morfossintáticas. Essas características fazem com que os resultados aqui apresentados possam ser utilizados em conjunto com esses recursos provindos de projetos anteriores, o que permite uma diversidade de pesquisas futuras sobre variados temas e áreas.

se dá por relações de dependência (*deprels*) e a representação de uma estrutura de dependências é arbórea, em que uma palavra é o *root* (raiz).

Na Figura 1, ilustra-se a anotação UD de um *tweet* do *corpus* DANTEStocks. Vale salientar que a versão atual do *corpus* ainda não possui anotação de *deprels*. O exemplo da Figura 1 foi anotado, com base em Sanguinetti *et al*. (2020), apenas para ilustrar os construtos do modelo. Nessa figura, as etiquetas de PoS[8] estão em caixa alta, como NOUN para *acordo*. Acima, estão os lemas, como *voo* para *voos*. As *deprels* estão indicadas por setas rotuladas que se originam no *head* e se destinam ao dependente. Na Figura 1, *acordo* é dependente de *assinou* e estes estão conectados pela *deprel*[9] **obj** (objeto direto[10]). O verbo *assinou* é o *root* da representação.

A UD também fornece uma extensa lista de traços que codificam traços (*features*) lexicais e gramaticais das palavras. Os traços não constam na Figura 1, mas, segundo a UD, *acordo*, por exemplo, tem os traços-valores: Gender=Masc e Number=Sing.

Figura 1: Exemplo de *tweet* do DANTEStocks anotado via UD.




Fonte: anotado via UD.

A versão do DANTEStocks a que se teve acesso na presente pesquisa data de novembro de 2022 e possui 4.048 *tweets* anotados morfossintaticamente seguindo as diretrizes da UD. A unidade mínima de análise na anotação foi o *tweet* (e não a sentença ou outro segmento), devido às propriedades inerentes

[8] A versão 2.0 da UD dispõe de 17 etiquetas de PoS e de critérios para o emprego/anotação de cada uma delas.
[9] A UD 2.0 provê 37 *deprels* e critérios para o emprego de cada uma delas. As diretrizes estão disponíveis em: https://universaldependencies.org/guidelines.html. Acesso em fevereiro de 2023.
[10] Relação entre o predicado verbal e o segundo argumento core do verbo (o primeiro é nsubj).

ao gênero. Segundo Silva *et al.* (2021), os *tweets*, por vezes, não possuem construções sintáticas equivalentes à análise formal da língua, além de apresentarem oralidade, informalidade, instantaneidade enunciativa do sujeito e da forma de referência a ele e atos interlocutórios, aproximando-se de uma estrutura de diálogo. O Quadro 1 apresenta os principais fenômenos e particularidades do *corpus* em questão, sobrepondo-se características gerais do gênero (*microblogging*) a propriedades do domínio de conhecimento (*mercado financeiro*), segundo Di Felippo *et al.* (2021):

Quadro 1: Fenômenos UGC no *corpus* DANTEStocks

| **Fenômeno** | **Definição** | **Exemplo** |
|---|---|---|
| Simplificação de código | fenômenos *ergográficos*, que reduzem o esforço de escrita de um único token | ausência/adição/substituição de diacrítico (*milhao* em lugar de *milhão*);<br>omissão de letra (*qdo* em lugar de *quando*);<br>erro ortográfico/digitação (*comrpa* em lugar de *compra*) |
| Abreviação | forma reduzida de palavras; contração (de elementos gramaticais), siglas acrônimos ou inicialismos | contração (*pq* em lugar de *por quê*);<br>acrônimo/inicialismo (*BB* em lugar de *Banco do Brasil*; *cf* em lugar de *conselho fiscal*) |
| Expressão de sentimento | fenômenos que emulam o sentimento expresso pela prosódia, expressão facial ou gesto na interação | alongamento grafêmico (*noossaa* em lugar de *nossa*);<br>autocensura (*p**a* em lugar de *puta*) |
| Influência de língua estrangeira | vocábulo formado com base em outra língua | *estopar*: parar investimento, deriva de *stop*, do inglês. |
| Expressão de oralidade | palavra da grafia relacionada à comunicação (fala) informal, às vezes com função humorística | coloquialismo (*guvêrno* em lugar de *governo*);<br>*hahahaha*: risos |
| Elemento metalinguístico | elemento recorrente no Twitter, como *hashtags*, menções, marcas de retweet, URL, trucamentos (quebras de palavras) | hashtags (*#PETR4*: indexador de tópico ou assunto);<br>menção (@garimpodeacoes: perfil de usuário) |

| | | |
|---|---|---|
| Fenômeno de domínio (mercado financeiro) | fenômeno lexical/gráfico que diferencia os tweets do DANTEStocks: tickers, cashtag, numerais, índices etc. | índice de desvalorização (*+2,09%* em lugar de *subiu 2,09%*);<br>expressão (temporal) híbrida (*1T14* em lugar de *primeiro trimestre de 2014*) |

Fonte: elaboração própria com base nos dados de Di Felippo *et al*. (2021).

É fundamental considerar os fenômenos apresentados no Quadro 1 no momento da seleção dos candidatos a termos do mercado financeiro extraídos do DANTEStocks, pois se referem a propriedades características e recorrentes em *tweets*, de maneira geral, e no *corpus* de análise em questão.

# 2. Metodologia: extração e validação de termos do DANTEStocks

Para a obtenção da lista de candidatos a termos do mercado financeiro extraída do *corpus* DANTEStocks, realizaram-se os seguintes procedimentos:

1. Análise geral dos *tokens* que constituem o *corpus*;
2. Extração automática de n-gramas a partir de padrões convencionais, com o auxílio da ferramenta UDConcord[11];
3. Análise manual dos dados gerados, a fim de se obter uma lista inicial de candidatos a termos da área;
4. Validação dos candidatos a termos com base nas listas elaboradas nas etapas anteriores por três especialistas da área;
5. Análise da concordância da validação dos dados entre os especialistas e elaboração da lista final de termos da área; e
6. Detalhamento das propriedades da lista terminológica criada e sua comparação com outras duas bases de dados (glossários terminológicos) que abrangem o mercado financeiro de alguma maneira.

A respeito da *primeira etapa*, foi possível ter acesso à frequência dos tokens que compõem o DANTEStocks. Nesse sentido, destacaram-se os

[11] Disponível em: https://udconcord.icmc.usp.br/. Acesso em janeiro de 2023.

fenômenos do domínio do mercado financeiro, caracterizados, principalmente, pela presença de numerais relacionados a preços, porcentagens e datas (R$13,55, 3,44%, 31/12/2013), tickers (*PETR4, Vale5, ITUB4*) e nomes próprios (*iBovespa*, *Brasil*); além dos elementos metalinguísticos que são recorrentes no *Twitter* como URLs, marcas de *retweet* e menções (*http*, rt, @usuário_twitter).

Considerando-se as particularidades do *corpus*, para a eleição dos candidatos a termos, foram estipuladas as seguintes diretrizes para os casos de 1-grama: (i) exclusão de elementos gramaticais e nomes próprios, considerando a lista de frequência dos tokens; (ii) análise manual dos termos anotados como NOUN (substantivo comum), com frequência maior que 5 no *corpus*[12]; (iii) análise de uma lista prévia de palavras elencadas pela equipe de pesquisadores da fase de anotação do *corpus* (DI FELIPPO *et al.* 2022); e (iv) análise de duas listas de palavras com até 4 caracteres, a fim de identificar os candidatos a termos que aparecem no *corpus* como abreviaturas, siglas e acrônimos[13]. Já para os casos de n-grama, extraíram-se bigramas, trigramas e tetragramas do *corpus* a partir de padrões morfossintáticos específicos, conforme será relatado nas etapas seguintes.

Destacam-se a atenção e o interesse, nesta pesquisa, pelos possíveis casos de abreviaturas, siglas ou inicialismos e acrônimos do *corpus*. Com base em trabalhos anteriores (ALVES 2001; LIMA 2019), entende-se que *abreviatura* é uma forma encurtada de uma única palavra (*cp*: compra; *máx.*: máximo), *sigla* ou *inicialismo* se refere à redução de um sintagma (palavras compostas) sob a forma de suas letras iniciais (*JSCP*: juros sobre capital próprio; *IFR*: índice de força relativa) e *acrônimo* é a redução do sintagma (palavras compostas) sob a forma de sílabas pronunciadas como uma palavra autônoma (*SELIC*[14]: sistema especial de liquidação e custódia).

Na *segunda etapa* da pesquisa, a versão do *corpus* DANTEStocks, anotada morfossintaticamente segundo as diretrizes da UD (em formato CoNLL-U), foi

[12] Optou-se por uma frequência relativamente baixa devido à dimensão do *corpus*, que possui apenas 4.048 *tweets*.
[13] As listas totalizaram aproximadamente 450 entradas (*types*). Ao todo, 146 casos (*token*s) foram considerados potenciais termos do mercado financeiro e, por isso, foram incluídos na lista de candidatos a termos para posterior validação de especialistas.
[14] *SELIC* não consta na lista de termos nesta pesquisa, porque foi anotado como nome próprio no DANTESTocks e aqui só foram analisados os casos de nomes comuns.

submetida à ferramenta online UDConcord. Nela, foi possível fazer as buscas pelos padrões morfossintáticos que mais comumente representam termos em língua portuguesa[15].

Quanto à compilação dos candidatos, os *tweets* contendo os padrões morfossintáticos exemplificados no Quadro 2, que são, na verdade, sequências de *tags* PoS (por exemplo, NOUN+ADP+NOUN), foram organizados em planilhas de maneira automática, conforme se ilustra no Quadro 2. A planilha em questão contém 5 colunas, denominadas: (i) *Antes* (trecho do *tweet* que antecede o candidato a termo), (ii) *Lexema em análise* (isto é, o candidato a termo), (iii) *Depois* (trecho do *tweet* que sucede o candidato a termo), (iv) n-grama (isto é, a extensão do candidato, podendo ser bigrama, trigrama ou tetragrama) (v) e *Padrão* (isto é, a sequência de *tags* que permitiu a compilação do candidato).

Quadro 2: Organização dos lexemas do DANTEStocks para análise[16]

| Antes | Lexema em análise | Depois | N-grama | Padrão |
|---|---|---|---|---|
| Me lembra os | **junk bonds** | da de a PDVSA @edmilson #PETR4 fundo do de o poço RT Petrobras prepara megacaptação de US$ 12 bi http://t.co/UyCIRKCzjO | bigrama | NOUN+NOUN |
| | **Índice de ações** | MSCI Brasil, do de o Morgan Stanley, q é seguido por grandes fundos retirou da de a carteira a construtora MVRE3 e a Hering - HGTX3 ... | trigrama | NOUN+ADP+NOUN |
| #BOVESPA #BRML3 BR Malls publica suas | **demonstrações financeiras** | anuais de 2013. http://t.co/i8ar5BflH1 | bigrama | NOUN+ADJ |

Fonte: elaboração própria.

[15] Ressalta-se que a anotação de PoS do referido *corpus* foi feita, em trabalhos anteriores, com base em dois manuais, um deles desenvolvido para subsidiar a anotação morfossintática segundo a UD de textos em português (DURAN 2021) e o outro definido especificamente para a anotação de PoS do DANTEStocks (DI FELIPPO *et al.* 2022).

[16] Os dados foram organizados em planilhas separadas com base no padrão de busca e só posteriormente os candidatos a termos foram unificados em uma única planilha.

A *terceira etapa* constituiu-se da análise manual dos padrões de n-grama em todo o *corpus* (cf. Tabela 1). Na Tabela 1 apresentam-se os dados gerais dos candidatos a termos com relação aos padrões de busca, apresentando as informações referentes aos *tokens* (número total de palavras encontradas), *types* (número de palavras diferentes) e número total de candidatos a termos.

Tabela 1: Relação entre os *padrões de busca* no DANTEStocks e os *candidatos a termos* elencados

| n-grama | Padrão[17] | Exemplo | *Tokens* | *Types* | Candidatos a termos |
|---|---|---|---|---|---|
| **Bigrama** | NOUN+ADJ | *venda coberta* | 1.293 | 483 | 139 |
| | NOUN+NOUN | *série C* | 436 | 139 | 24 |
| | ADJ+NOUN | *longo prazo* | 843 | 330 | 8 |
| **Trigrama** | NOUN+ADP+NOUN | *engolfo de baixa* | 1.161 | 661 | 110 |
| | NOUN+NOUN+ADJ | **mercado spot chinês*[18] | 6 | 6 | 0 |
| | NOUN+ADJ+ADJ | *assembleia geral ordinária* | 150 | 26 | 12 |
| **Tetragrama** | NOUN+ADJ+ADP+NOUN | *FAD anormal para venda* | 67 | 48 | 2 |
| | NOUN+ADP+NOUN+ADJ | *aumento de capital social* | 224 | 72 | 8 |
| **Total** | | | 16.114 | 3.871 | **303** |

Fonte: elaboração própria.

[17] Segundo a anotação UD, as *tags* NOUN, ADJ e ADP indicam, respectivamente, as seguintes classes de palavras: nome/substantivo, adjetivo e preposição. Ressalta-se que ADP (adposição) é um termo da UD que contempla itens lexicais que ocorrem antes (preposições) ou depois (posposições) de um complemento constituído por substantivo, pronome, sintagma nominal ou oração subordinada substantiva. Como no português só ocorrem preposições (e não posposições), a *tag* ADP deve ser tomada como sinônimo de preposição.

[18] O asterisco (*) indica que o exemplo não é considerado um termo nesta pesquisa.

Apesar de não constar na Tabela 1, que ilustra os dados gerais da busca no *corpus* pelos n-gramas, é importante mencionar que foi extraído maior número de candidatos a termos do tipo *unigrama* (NOUN), ou seja, candidatos a termos constituídos por um único nome (224 casos), conforme descrito na *primeira etapa* da pesquisa. Depois dos unigramas, os dados da Tabela 1 indicam que os candidatos do tipo bigrama (isto é, termo constituído por dois lexemas), foram os casos mais recorrentes (171 casos), com a seguinte distribuição de casos entre os padrões: NOUN+ADJ (139 casos) e NOUN+ADP+NOUN (110 casos). Salienta-se que o número da coluna *types* se refere não apenas aos casos de palavras diferentes do *corpus*, mas também às palavras com variações na escrita, tais como: *aquisição de ações* e *aquisicao de acoes*. Essa decisão foi tomada devido ao interesse em identificar ou levantar a variação ortográfica/gráfica no *corpus*, uma vez que essa é uma das características marcantes encontradas em material textual produzido por usuários (*cf*. Quadro 1). No entanto, na lista de *Candidatos a termos*, essas variantes de forma não foram contabilizadas como entradas individuais, mas sim indicadas como tal em uma coluna específica (*cf*. Tabela 3). Assim, o número total de *Candidatos a termos*, na Tabela 2, refere-se de fato aos casos de *types*.

Na *quarta etapa* da pesquisa, esses dados foram revisados e validados por três especialistas da área, e, na *quinta etapa*, a validação feita pelos especialistas foi analisada com pormenores. Dos 527 candidatos a termos (224 unigramas e 303 n-gramas), 380 (72,1% dos casos) foram anotados como termos por todos os especialistas. Nesse processo, houve diferentes cenários de discordância entre os consultores e, por isso, apenas os casos em que houve total concordância entre eles fizeram parte da lista final. Ainda que essa possa parecer uma decisão extrema, optamos por tomá-la como uma tentativa de tornar a lista de termos o mais objetiva possível, descartando assim vieses que pudessem ser introduzidos pelo estado emocional e personalidade dos anotadores (ALM *et al*. 2005), ou mesmo pela subjetividade natural da tarefa, como já observado na anotação de fatores emocionais (TURNEY; LITTMAN 2003), vieses esses que reduzem a concordância.

É importante destacar que um dos especialistas realizou apontamentos sobre o fato de que alguns termos não são específicos do mercado financeiro, mas sim de outras áreas de domínio correlatas, como economia (*política inflacionária*, *mercado interno*, *superavit primário*), contabilidade (*margem líquida*, *resultados operacionais*) e matemática e estatística (*média anual*, *variação negativa*). No entanto, como essas especificidades não haviam sido consideradas pelos demais especialistas, optou-se por considerar tais termos pertencentes ao mercado financeiro, sendo assim incluídos na lista final.

Ainda na etapa de validação, alguns especialistas sugeriram a inclusão de outros termos (*mercado à vista*, *fato relevante*, *zona de alto risco*) com base na leitura dos *tweets* enviados como exemplo. Devido à falta de validação dos demais consultores, esses casos foram anotados em um documento separado para estudos futuros.

Além disso, destacam-se 3 alterações sugeridas pelos especialistas na lista de termos final: (i) extensão do termo *white soldiers* (bigrama) para *three white soldiers* (trigrama) (NOUN+ADJ+NOUN), (ii) inclusão do termo *linha de tendência de baixa*, que, por se tratar de um pentagrama (NOUN+ADP+NOUN+ADP+NOUN), não havia sido capturado pelos métodos de extração, mas apenas a sua sigla (isto é, o unigrama LTB), (iii) inclusão do bigrama *virar pó*, que, por seguir o padrão VERB+NOUN, também não tinha sido capturado. Na Tabela 2, apresentam-se os dados gerais dos termos validados.

Tabela 2: Relação entre os *padrões de busca* no DANTEStocks e os *termos validados*

| n-grama | Padrão | Exemplos de termos por ordem de maior frequência | Termos validados |
|---|---|---|---|
| **Unigrama** | NOUN | *ação, ativo, compra, alta, papel, preço, intraday, queda, volume, pregão, fundo, objetivo, mercado, lucro, resultado, valor.* | 163 |
| **Bigrama** | NOUN+ADJ | *demonstração financeira, média móvel* | 90 |
| | NOUN+NOUN | *gatilho position, day trade* | 21 |

| | | | |
|---|---|---|---|
| | ADJ+NOUN | *curto prazo, blue chip* | 6 |
| | VERB+NOUN | *virar pó* | 1 |
| **Trigrama** | NOUN+ADP+NOUN | *distribuição de dividendo, recompra de ações, programa de recompra, valor de mercado* | 83 |
| | NOUN+ADJ+ADJ | *assembleia geral extraordinária* | 10 |
| | NOUN+ADJ+NOUN | *three white soldiers* | 1 |
| **Tetragrama** | NOUN+ADP+NOUN+ADJ | *juros sobre capital próprio, aumento de capital social* | 4 |
| **Pentagrama** | NOUN+ADP+NOUN+ADP+NOUN | *linha de tendência de baixa* | 1 |
| **Total** | | | **380** |

Fonte: elaboração própria.

Além das informações apresentadas na Tabela 2, outras propriedades dos termos validados foram analisadas, constituindo a *sexta etapa* deste trabalho. Tais propriedades foram as variações ortográficas/gráficas, as abreviaturas e afins, a frequência e os estrangeirismos, além da comparação da lista final de termos com outras obras terminológicas. Esses dados estão mais bem descritos na próxima seção.

# 3. Particularidades dos termos do DANTEStocks

No que se refere à *sexta etapa* da pesquisa, verificaram-se as particularidades dos termos retirados do DANTEStocks, a fim de descrever, de maneira geral, a lista de lexemas proposta. Na Tabela 3, exemplificam-se as propriedades, distribuição e anotação dos dados validados pelos especialistas.

Tabela 3: Propriedades dos termos extraídos do DANTEStocks

| Termo | Ngram | Padrão | Exemplo | Variações | Abrev. | Sigla | Acr. | Freq. | Estr. | Econ. | Info. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ação | Uni | NOUN | PETR4 a 13 pratas, agora me caiu a ficha : estratégia de marketing brilhante do de o PT , associando o preço da de a *ação* ao a o seu número de legenda ! | #ações, acao, açõ, ação, acoes, Açoes, ações | - | - | - | 600 | - | - | + |
| intraday | Uni | NOUN | *INTRADAY* PETR4: Suportes 13 e 13,14 e resistências 13,58 e 13,88 INTRADAY VALE5 : Suportes 27,44 e 27,77 e resistências 28,67 e 29,24 | - | Intra | - | - | 88 | + | - | + |
| demonstrações financeiras | Bi | NOUN ADJ | #BOVESPA #BRML3 BR Malls publica suas* demonstrações financeiras* anuais de 2013. http://t.co/i8ar5BflH1 | Demontracoes Financeiras, DFs | - | + | - | 75 | - | - | + |
| linha de tendência de baixa | Penta | NOUN ADP NOUN ADP NOUN | Vale5 - Diário: *Linha de Tendência de baixa * , zona de sobre-venda ! http://t.co/m5dEkcVXf3 | LTB, ltb | - | + | - | 12 | - | - | - |
| valor de mercado | Tri | NOUN ADP NOUN | @GutoAbranches @denisebarbosa @valor_economico *Valor de mercado *do de o #Facebook #FB é maior do de o que da de a #Petrobrás PETR4 http://t.co/fvrM2s4hzC | Valor d mercado | - | - | - | 9 | - | - | + |

Fonte: elaboração própria.

Conforme se verifica na Tabela 3, cada termo do DANTEStocks foi descrito em função de algumas propriedades, as quais são descritas na sequência com a indicação do título da coluna correspondente.

i. Ngrama: tipo ou extensão do n-gram (isto é, unigrama, bigrama etc.)
ii. Padrão: sequência de *tags* PoS usada como padrão morfossintático para a extração automática dos candidatos do tipo n-gramas>1 (por exemplo: NOUN+ADJ)
iii. Exemplo: *tweet* que exemplifica a ocorrência do termo no DANTEStocks
iv. Variações: listagem das diferentes expressões linguísticas do termo no *corpus*, incluindo formas com problemas relacionados a diacríticos (por exemplo: ausência de acentuação), abreviaturas, siglas e acrônimos
v. Abreviatura: indicação se o termo é uma abreviatura ou não (por meio dos polos +/-)
vi. Sigla: indicação se o termo é uma sigla ou não (por meio dos polos +/-)

vii. Acrônimo: indicação se o termo é um acrônimo ou não (por meio dos polos +/-)

viii. Frequência (do termo no *corpus*): frequência simples de ocorrência do termo no *corpus* DANTEStocks

ix. Estrangeirismo: indicação se o termo é um estrangeirismo ou não (por meio dos polos +/-)

x. Econ.: ocorrência do termo no *Glossário de termos neológicos da economia* (ALVES 2001)

xi. Info.: ocorrência do termo no *Glossário Financeiro InfoMoney*.

No que se refere à variação ortográfica, verifica-se, como apontado por Di Felippo *et al.* (2021), a ocorrência de vários termos nos quais ocorrem o fenômeno denominado *simplificação do código*, sobretudo a ausência de diacrítico, em especial, a cedilha e o til (*acao, operacao*) e o acento agudo (*indice*). É importante registrar tais variações devido às especificidades dos textos produzidos por usuários, que, de maneira geral, caracterizam a linguagem não-padrão desse tipo de material textual. Esse registro contribui com a anotação (manual e automática) desse tipo de conteúdo.

Também no que se refere à simplificação do código, muitos termos (54 casos) aparecem no *corpus* como abreviaturas, siglas ou acrônimos, a saber:

- Há um total de 27 termos abreviados, incluindo variações ortográficas (*compra* > *co, cp; mínima*> *mín, min*) e de pontuação (*fechamento* > *fech., vencimento* > *venc.; prévia* > *prev, rompimento* > *rp*); os termos abreviados são unigramas na maioria dos casos, mas o fenômeno da simplificação também ocorre em termos compostos (ou n-gramas>1) (*tendência de alta* > *tend. alta; ponto de entrada* > *pto de entrada*);
- Há 24 termos que aparecem como siglas no *corpus* (*P/L, PMI*) e 3 casos de acrônimos (*capex, FAD, OPA*), havendo preferência pelo emprego de siglas/acrônimos em substituição aos termos por extenso. A título de exemplo, destaca-se o caso de *LTB* (*linha de tendência de baixa*), em que das 12 ocorrências no *corpus*, 11 aparecem como sigla; ou ainda OPA (*Oferta Pública de Aquisição*), em que todas as 14 ocorrências se dão como acrônimos.

Com relação à frequência dos termos, ressalta-se que a grande maioria dos termos (303 casos) apresenta frequência menor que 10 (*acionista, movimento, prévia, variação positiva*); com frequência entre 11 e 30, contabilizam-se 48 termos (*dividendo, gatilho, investimento*); entre 31 e 50 encontram-se 12 termos (*média móvel, stop, capital, baixa*); já com frequência maior que 51, somam-se apenas 17 termos (*ação, ativo, compra, alta, papel, preço, intraday, queda, volume, pregão, demonstrações financeiras, fundo, objetivo, mercado, lucro, resultado* e *valor*).

Dos 380 termos, 66 (17,3%) são *estrangeirismos*, isto é, vocábulos provenientes do inglês, como *aftermarket*, *bond, day trade, joint venture*, *market share, short cap*. O unigrama *scaplerzinho* (< *scapler*), aliás, parece ter passado por um processo de *empréstimo* com o acréscimo do sufixo -inho. Segundo Rodrigues e Vale (2023), o sufixo -inho/-inha pode designar, dentre outros valores, tamanho pequeno ou denotar juízo de valor (depreciação ou apreciação). Além disso, alguns estrangeirismos são compostos pela união de palavras do inglês e do português (*advisor financeiro, agência de rating, call de compra, gatilho position, rating de crédito* etc.). Ainda entre esses termos, 9 se apresentam em formas abreviadas (*intraday* > *intra*), como siglas (*DT, IFR, IPO, ETF, PMI, m&a, SW*) ou acrônimo (*capex*).

Nos dados, verificam-se ainda relações de hiperonímia entre alguns termos. É o caso do hiperônimo *acionista*, que se relaciona a termos mais específicos ou hipônimos, como *acionista controlador* e *acionista minoritário*. Nesses exemplos, o termo hiperônimo é especificado pela ocorrência dos adjetivos *controlador* e *minoritário*. O mesmo ocorre com *call* (*call de compra, call de venda*), *carteira* (*carteira de ações, carteira quantitativa, carteira simulada*), *mercado* (*mercado acionário, mercado de capitais, mercado financeiro, mercado global, mercado internacional, mercado interno, mercado nacional, mercado spot*), entre outros. As relações entre hiperônimos e hipônimos, assim como os possíveis casos de sinônimos, devem ser estudadas em trabalhos futuros, entendidos como questões essenciais para a proposta de elaboração de fichas terminológicas.

Como etapa final da pesquisa, estabeleceram-se relações entre os 380 termos validados nesta investigação e duas bases de dados terminológicas que

dialogam com o mercado de ações: o *Glossário de termos neológicos da economia* (ALVES 2001) e o *Glossário Financeiro InfoMoney*. A escolha por esses trabalhos se deu pela facilidade de acesso aos dados de maneira legível por máquina.

O *Glossário de termos neológicos da economia* apresenta informações gramaticais, siglas e acrônimos, variantes, definições, exemplos (*contexto*), notas, sinônimos e remissivas de 602 termos relacionados à grande área da Economia. Para denotar o caráter neológico, foram agrupados apenas os termos que não estivessem registrados na segunda edição do *Novo dicionário da língua portuguesa* de Aurélio Buarque de Holanda Ferreira, publicado em 1986 pela editora Nova Fronteira (ALVES 2001).

Por sua vez, o *Glossário Financeiro InfoMoney* é uma obra eletrônica, que pertence e está disponível no Site InfoMoney; um canal especializado em mercados, investimentos e negócios do Brasil. Ao todo, foram extraídos do site 1.171 termos e 160 siglas/acrônimos, totalizando 1.331 verbetes. A busca pelos termos se dá em ordem alfabética e cada entrada possui apenas uma breve definição e, em alguns casos, remissivas externas ao glossário, direcionando o usuário a notícias publicadas no próprio site.

Considerando as particularidades (e o nível de especificidade) de cada obra terminológica, estabeleceu-se a sua comparação com os termos levantados do DANTEStocks, obtendo-se os dados dispostos na Tabela 4.

Tabela 4: Relação entre os termos do DANTEStocks e de outras bases de dados

| Relações | Exemplos | Termos comuns |
|---|---|---|
| DANTEStocks – **Econ**omia – **Info**Money | *curto prazo, volatilidade, liquidez, longo prazo, market share, fundo de pensão, patrimônio líquido* | 7 |
| DANTEStocks – **Econ**omia | *preço médio, médio prazo, mercado acionário, saldo positivo* | 16 |
| DANTEStocks – **Info**Money | *ação, ativo, preço, call, PUT, volume médio diário* | 83 |

Fonte: elaboração própria.

O caráter neológico do Glossário de Economia de Alves (2001) parece justificar a ausência de termos bastante frequentes no mercado financeiro, tais

como *ação, ativo, compra, alta*. O nível de especificidade desse Glossário é mais refinado: não há uma entrada com os termos *abertura* e *índice*, por exemplo, mas termos mais acurados como *abertura comercial, abertura de mercado, índice econômico, índice de custo de vida, índice de inadimplência, índice de inflação*, entre outros. Além disso, os casos coincidentes (DANTEStocks – **Econ**omia) se referem a termos mais amplos e gerais da Economia, ou seja, não são específicos ou de uso exclusivo do mercado financeiro.

Já o Glossário InfoMoney estabeleceu mais proximidade com os dados do DANTEStocks. No entanto, o Glossário InfoMoney não apresenta informações detalhadas sobre a sua elaboração, o que torna difícil realizar reflexões e generalizações sobre os casos de intersecção entre as duas listas de termos. O maior número de termos comuns, no entanto, aponta para a importância da criação de obras terminológicas de domínios específicos, pois, mesmo utilizando com frequência termos de áreas afins (Economia, Matemática, Estatística), é preciso considerar as particularidades dos termos recorrentes no Mercado Financeiro em si.

Salienta-se que entre o Glossário de Economia e o InfoMoney, há a correspondência de 145 termos; um número muito maior se comparado às relações estabelecidas com a lista de termos do DANTEStocks. Isso se deve à quantidade de entradas descritas em cada glossário e, em parte, ao fato de ambas as obras recensear nomes próprios (*BNDES, CVM, Ibovespa, Serasa*, etc.). Conforme apresentado na Seção 2 deste artigo, há muitos casos de anotação de nomes próprios (PROPN) no *corpus*, incluindo nomes de instituições financeiras (*Banco do Brasil, Bradesco*), de taxas (*SELIC*), de fundos (*FGTS*), de sistemas (*home broker*), entre outros, e que não foram considerados nesta investigação. Portanto, considera-se importante retomar os critérios de seleção de termos, em trabalhos futuros, e, se cabível, realizar uma análise minuciosa dos nomes próprios anotados no DANTEStocks.

# Considerações Finais

Nesta pesquisa, realizou-se um procedimento metódico de coleta, validação e descrição geral de termos do mercado financeiro com base nos dados do *corpus* de *tweets* DANTESTocks. Ao todo, 527 candidatos a termos foram extraídos do *corpus* e, posteriormente, analisados e validados por especialistas, culminando em uma lista de 380 termos.

A lista gerada apresenta algumas particularidades, a saber: (i) variações ortográficas típicas do gênero *microblogging*, mais especificamente de *tweets* (*acao, indice*); (ii) casos frequentes de abreviaturas, siglas e acrônimos (*ago, bx, encerr., invest., LTA, OPA*); e (iii) estrangeirismos (*candle, delay, stop loss, valuation*).

Além disso, os termos descritos nesta pesquisa foram comparados com termos que constam no *Glossário de termos neológicos da economia* (ALVES 2001) e no *Glossário Financeiro InfoMoney*, apresentando poucos casos de convergência: 16 e 83 termos em comum, respectivamente. Isso indica a necessidade de refinamento dos termos extraídos do DANTEStocks, a partir da aplicação de critérios bem delimitados de granularidade, considerando-se as relações de *hiperonímia, hiponímia* e *sinonímia* entre os termos anotados. Além disso, os 288 termos encontrados apenas no DANTEStocks apontam para a importância de estudos baseados em *corpus* de textos produzidos por usuários, afinal, se trata de termos realmente recorrentes em contextos mais informais e no dia a dia de pessoas (com maior ou menor grau de expertise) que escrevem textos (*tweets*) relacionados ao domínio do mercado financeiro.

Como trabalhos futuros, é importante ampliar a lista de termos, considerando incluir casos anotados como *nomes próprios* no DANTEStocks. Ademais, verifica-se a possibilidade de elaboração de fichas terminológicas para os termos elencados neste trabalho, assim como a criação de um glossário multilíngue (português, espanhol e inglês) da área.

Esta pesquisa contribui com o estado da arte, no sentido de apresentar uma descrição minuciosa dos procedimentos metodológicos para a obtenção de termos extraídos de um *corpus* de *tweets*; colabora com a descrição do *corpus* DANTEStocks, a partir do estudo detalhado dos tokens que o constituem –

inclusive, a lista de termos já está sendo utilizada no interior do Projeto POeTiSA para procedimentos de análise e anotação do DANTEStocks por investigadores que têm pouco conhecimento sobre o mercado financeiro; e apresenta, como produto, um léxico (lista de termos) elaborado com rigor metodológico, que pode ser utilizado em diferentes pesquisas futuras.

# Agradecimentos

Este trabalho foi realizado no âmbito do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI - http://c4ai.inova.usp.br/), com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM. Este projeto também 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.

# Referências

AGBABO, A., Barrière, C. Corpus construction for Terminology. In: *Proceedings of the Corpus Linguistics Conference*, Birmingham, 2005: 14-17.

ALM, C. O.; ROTH, D.; SPROAT, R. Emotions from text: Machine learning for text-based emotion prediction. In: *Proceedings of HLT/EMNLP 2005*, Vancouver/Canada: Association for Computational Linguistics, 2005: 579-586.

ALMEIDA, G. M. B. A Teoria Comunicativa da Terminologia e a sua prática. *Revista ALFA*, [*s.l.*], v. 50, n. 2, 2006.

ALMEIDA, G. M. B.; OLIVEIRA, L. H. M. Terminology and computational linguistics: new praxes in terminography. *Cahiers de Lexicologie*, [*s.l.*], v. 101, 2012: 139-153.

ALVES, I. M. *Glossário de termos neológicos da economia*. Cadernos de Terminologia, 3 (Reimpressão). São Paulo: Humanitas, 2001.

BARROS, L. A. *Curso básico de Terminologia*. São Paulo: EDUSP. 296p, 2004.

BIDERMAN, M. T. C. *Dicionário de termos financeiros e bancários*. Disal Editora-Bantim, Canato e Guazzelli Editora Ltda, 2013.

CABRÉ M. T. Hacia una teoría comunicativa de la terminología: Aspectos metodológicos. In: CABRÉ, M. T. *La Terminología: Representación y Comunicación*: Elementos para una teoría de base comunicativa y otros artículos. Barcelona: Universitat Pompeu Fabra, 1999: 129-150.

CABRÉ M. T. La terminología, una disciplina en evolución: pasado, presente y algunos elementos de futuro. *Revista Debate Terminológico,* [*s.l.*], n. 1, 2005.

DI FELIPPO, A.; POSTALI, C.; CEREGATTO, G.; GAZANA, L. S.; SILVA, E. H.; ROMAN, N. T.; PARDO, T. A. S. Descrição preliminar do corpus dantestocks: Diretrizes de segmentação para anotação segundo universal dependencies. In: *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. SBC, 2021.

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 (UD). *Relatório Técnico do ICMC 438*. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP, 2022.

DROUIN, P.; GRABAR, N.; HAMON, T.; KAGEURA, K.; TAKEUCHI, K. Introduction. In: *Proceedings of the 5th International Workshop on Computational Terminology,* Osaka/Japan: The COLING 2016 Organizing Committee, 2016.

DURAN, M. S. Manual de anotação de PoS *tags*: orientações para anotação de etiquetas morfossintáticas em Língua Portuguesa, seguindo as diretrizes da abordagem *Universal Dependencies*. *Relatório Técnico do ICMC,* 434. ICMC, USP, São Carlos, 2021.

EISENSTEIN, J. What to do about bad language on the internet. In**:** *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics*: Human Language Technologies, Atlanta/Georgia: Association for Computational Linguistics, 2013: 359-369.

FINATTO, M. J. B. Acessibilidade textual e terminológica, o que é isso? In: FINATTO, M. J. B.; PARAGUASSU, L. B. (Org.). *Acessibilidade Textual e Terminológica*. 1ed. Uberlândia: EDUFU, 2022.

FUKUTOME, N.; HARADA, Y. Flavor Wheel Terminology and Challenges in Translation – Focusing on English and Japanese Vocabulary for Wine, Sake and Soy sauce. In**:** *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*. Hong Kong: Association for Computational Linguistics. 2018.

JURAFSKY, D.; MARTIN, J. H. *Speech and Language Processing:* An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. 2a edição. Prentice Hall. 2008.

LANG, C.; WACHOWIAK, L.; HEINISCH, B.; GROMANN, D. Transforming Term Extraction: Transformer-Based Approaches to Multilingual Term Extraction Across Domains. In: *Findings of the Association for Computational Linguistics*. [*s.l.*]: Association for Computational Linguistics, 2021.

LIMA, E. B. A. *A tradução de siglas e acrônimos em textos acadêmicos de ciências da saúde*. Trabalho de conclusão de curso (Bacharelado em Tradução). João Pessoa: Universidade Federal da Paraíba, 2019.

LOPES, L.; VIEIRA, R.; FINATTO, M. J.; MARTINS, D.; ZANETTE, A.; RIBEIRO JR, L. C. Extração automática de termos compostos para construção de ontologias: um experimento na área da saúde. *RECIIS*: Revista eletrônica de comunicação, informação & inovação em saúde. Rio de Janeiro/RJ., vol. 3, n. 1, 2009.

NASCIMENTO, M. F. B. O papel dos corpora especializados na criação de bases terminológicas. In: CASTRO, I.; DUARTE, I. (orgs.). *Razões e emoções, miscelânea de estudos em homenagem a Maria Helena Mateus*. Lisboa: Imprensa Nacional-Casa da Moeda, vol. II, 2003.

NIVRE, J. Towards a universal grammar for natural language processing. In: *Proceedings of Computational Linguistics and Intelligent Text Processing – Part 1,* Cairo/Egypt: Springer International Publishing, 2015.

NIVRE, J.; MARNEFFE, M.; GINTER, F.; HAJIE, J.; MANNING, C.D.; PYYSALO, S.; SCHUSTER, S.; TYRES, F.; ZEMAN, D. Universal Dependencies v2: An evergrowing multilingual treebank collection. *arXiv preprint arXiv:2004*.10643, 2020.

PARDO, T. A. S.; DURAN, M. S.; LOPES, L.; DI FELIPPO, A.; ROMAN, N. T.; NUNES, M. G. P. Porttinari - a Large Multi-genre Treebank for Brazilian Portuguese. In: *Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana* (STIL). Porto Alegre: Sociedade Brasileira de Computação, 2021.

PENG, B.; CHERSONI, E.; HSU, Y-Y.; HUANG, C-R. Discovering Financial Hypernyms by Prompting Masked Language Models. In: *Proceedings of the 4th Financial Narrative Processing Workshop*. Marseille/France: European Language Resources Association, 2022.

RODRIGUES, R.; VALE, O. A. Falsos Diminutivos do Português Brasileiro e seu Reconhecimento em um Dicionário Computacional de Livre Acesso. *Revista do GELNE*, vol. 25, 2023.

SANGUINETTI, M.; BOSCO, C.; CASSIDY, L.; ÇETINOĞLU, Ö.; CIGNARELLA, A. T.; LYNN, T.; REHBEIN, I.; RUPPENHOFER, J.; SEDDAH, D.; ZELDES, A. Treebanking user-generated content: a proposal for a unified representation in universal dependencies. In: *Proceedings of the 12th International Language Resources and Evaluation Conference*. Marseille/France: European Language Resources Association, 2020.

SEIFEE, L.; MARTEN, O.; MIKHAILOV, M.; SCHMEIER, S.; MÖLLER, S.; ROLLER, R. From Witch's Shot to Music Making Bones - Resources for Medical Laymen to Technical Language and Vice Versa. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille/France: European Language Resources Association, 2020.

SILVA, E. H.; PARDO, T. A. S.; ROMAN, N. T.; DI FELIPPO, A. Universal Dependencies for Tweets in Brazilian Portuguese: Tokenization and Part of Speech Tagging. *In*: Encontro Nacional de Inteligência Artificial e Computacional (ENIAC). Porto Alegre: Sociedade Brasileira de Computação, 2021.

SILVA, F. J. V.; ROMAN, N. T.; CARVALHO, A. M. B. R. Stock market tweets annotated with emotions. *Corpora*, v. 15, n. 3, 2020.

TURNEY, P.; LITTMAN, M. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems,* [*s.l.*], v. 21, n.4, 2003.

WU, S.; IRSOY, O.; LU, S.; DABRAVOLSKI, V.; DREDZE, M.; GEHRMANN, S.; KAMBADUR, P.; ROSENBERG, D.; MANN, G. BloombergGPT: A Large Language Model for Finance. *arXiv:2303.17564*, [*s.l.*], 2023.