UNIVERSIDADE FEDERAL DE SÃO CARLOS - UFSCAR
CENTRO DE EDUCAÇÃO DE CIÊNCIAS HUMANAS - CECH
DEPARTAMENTO DE LETRAS - DL
CURSO DE BACHARELADO EM LINGUÍSTICA

CLARISSA LENINA SCANDAROLLI

PROPOSTA E VALIDAÇÃO DE UMA TAXONOMIA DE VARIAÇÕES ORTOGRÁFICAS EM *TWEETS* DO MERCADO FINANCEIRO

SÃO CARLOS/SP
2026

PROPOSTA E VALIDAÇÃO DE UMA TAXONOMIA DE VARIAÇÕES ORTOGRÁFICAS EM TWEETS DO MERCADO FINANCEIRO

Monografia apresentada ao Departamento de Letras da Universidade Federal de São Carlos como Trabalho de Conclusão de Curso para obtenção do título de Bacharel em Linguística.

Orientadora: Prof.ª Dra. Ariani Di Felippo.

SÃO CARLOS/SP

2026

DEDICATÓRIA

Dedico este trabalho ao Thiago, meu amor e companheiro, por estar sempre ao meu lado, compartilhando cada desafio com leveza e afeto;

aos meus pequenos Haru, Buzz, Lily, Flash e Nelson Banguela, fontes inesgotáveis de alegria e conforto em todos os momentos;

à Camila, parceira generosa e organizada, que caminha comigo rumo a novos projetos e sonhos;

à Fernanda, Rafael e Luna, que preenchem meus finais de semana com risadas, diversão e brigas acaloradas jogando Catan;

à minha mãe Miriam, meu pai Mário e meu irmão Gustavo, pela presença, carinho e apoio constante dentro das suas possibilidades, sempre sendo uma família unida pelo amor;

e a todos que fizeram parte do meu processo de descoberta, superação e cura: Criolo, Lara, Guilherme, Itu, Arthur, Luan, Fer do teatro, o grupo de teatro Acaso e minhas terapeutas, pessoas fundamentais que marcaram profundamente minha jornada até aqui.

Por fim, dedico também a mim mesma, pela resiliência em recomeçar tantas vezes, me reerguer novamente e nunca perder a fé na vida.

AGRADECIMENTOS

Gostaria de expressar minha profunda gratidão à minha orientadora, Ariani Di Felippo, pela paciência, acolhimento e compreensão constantes ao longo de todos esses anos. Obrigada por me apoiar e incentivar, especialmente nos momentos mais difíceis, quando precisei interromper este trabalho devido a questões pessoais e de saúde. Sua orientação foi fundamental para que eu pudesse chegar até aqui.

Agradeço imensamente ao grupo de pesquisa DANTEStocks por terem me acolhido com tanta abertura e generosidade. A energia positiva, a motivação constante e a disposição de todos em abraçar minhas ideias "incomuns" vindas dos meus aprendizados em sociolinguística foram essenciais para que minha contribuição pudesse se concretizar. Com vocês aprendi muito sobre pesquisa, colaboração e Inteligência Artificial, e guardarei com carinho cada troca e aprendizado.

A todos vocês, meu mais sincero muito obrigada.

Este trabalho foi executado no Centro de Inteligência Artificial (C4AI-USP) com apoio da Fundação de Apoio à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM Corporation.

EPÍGRAFE

Se houver algo que você realmente queira fazer, faça.
No fim, seus pais, amigos e o restante do mundo ficarão felizes com a sua felicidade. ***Tenha coragem!***

Haemin Sunim

RESUMO

Dada a relevância da plataforma Twitter (agora X) para vários segmentos da sociedade, ferramentas e aplicações de Processamento de Língua Natural (PLN) capazes de lidar com a linguagem primordialmente não canônica do gênero tweet (atualmente post) são fortemente demandadas. Para desenvolvê-las, os corpora anotados (chamados *tweebanks*) são recursos essenciais, assim como a descrição e análise de suas características linguísticas. Nesta Monografia de Conclusão de Curso, o objeto de investigação foi o corpus de 4.048 tweets do mercado financeiro DANTEStocks, que é o primeiro em português com anotação segundo o modelo Universal Dependencies (UD). Mais precisamente, fez-se um levantamento dos fenômenos ortográficos no corpus, considerados “variações” e não “erros”, segundo os pressupostos da Sociolinguística Variacionista. Esses fenômenos foram sistematizados em uma tipologia hierárquica com duas dimensões: “Norma Padrão” e “Norma Inovadora”, que buscam capturar variações das formas canônicas e inovadoras lexicais. Com base na tipologia, os fenômenos observados em 3.614 tokens (encontrados em 1.069 tweets) do DANTEStocks foram manualmente anotados, gerando uma caracterização preliminar do corpus, que evidenciou a predominância da Norma Inovadora, correspondendo a 92,81% dos fenômenos anotados (3.457 ocorrências), reforçando a hipótese de que, no Conteúdo Gerado por Usuário (CGU) de temática financeira, predominam estratégias linguísticas inovadoras que refletem nos tokens, que são códigos e formas sistemáticas de comunicar pelo meio digital, característicos do meio e de determinado contexto social. Com isso, a anotação de variações gráficas no DANTEStocks amplia a compreensão da linguagem dos tweets do mercado financeiro e pode ampliar a tolerância dos modelos de PLN à linguagem não canônica, permitindo que reconheçam formas variantes como linguisticamente válidas e semanticamente informativas.

Palavras-chave: variação linguística; Processamento de Língua Natural; conteúdo gerado por usuário.

ABSTRACT

Given the relevance of the Twitter platform (now X) for various segments of society, Natural Language Processing (NLP) tools and applications capable of handling the primarily non-canonical language of the tweet genre (currently referred to as posts) are in high demand. To develop them, annotated corpora (known as tweetbanks) are essential resources, as are the description and analysis of their linguistic characteristics. In this Undergraduate Thesis, the object of investigation was the DANTEStocks corpus of 4,048 financial market tweets, which is the first in Portuguese to be annotated according to the Universal Dependencies (UD) model. More precisely, a survey of orthographic phenomena in the corpus was conducted, treated as "variations" rather than "errors", in accordance with the theoretical assumptions of Variationist Sociolinguistics. These phenomena were systematized into a hierarchical typology with two dimensions: "Standard Norm" and "Innovative Norm", which seek to capture variations of canonical and innovative lexical forms. Based on this typology, the phenomena observed in 3,614 tokens (found in 1,069 tweets) from DANTEStocks were manually annotated, yielding a preliminary characterization of the corpus. The results evidenced the predominance of the Innovative Norm, accounting for 92.81% of the annotated phenomena (3,457 occurrences), reinforcing the hypothesis that, in User-Generated Content (UGC) on financial topics, innovative linguistic strategies prevail, manifesting in tokens that function as codes and systematic forms of digital communication, characteristic of the medium and of a particular social context. In this way, the annotation of graphic variations in DANTEStocks broadens the understanding of the language of financial market tweets and may enhance the tolerance of NLP models toward non-canonical language, enabling them to recognize variant forms as linguistically valid and semantically informative.

Keywords: linguistic variation; Natural Language Processing; user-generated content.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE QUADROS

1. Introdução

O grande volume de "conteúdo gerado por usuários" (CGU)[1], produzido em tempo real e de forma espontânea no Twitter (atualmente X), caracteriza-se, em grande parte, por seu teor opinativo. Dada a ampla circulação e influência do Twitter, esse conteúdo constitui uma fonte valiosa de informação para diversas áreas, como política e finanças, permitindo, por meio de aplicações computacionais como análise de sentimentos e mineração de opiniões, monitorar percepções, analisar tendências e identificar padrões comportamentais relevantes para a tomada de decisão (Derczynski et al., 2016).

Essas aplicações podem se beneficiar — e até requerer — ferramentas de Processamento de Língua Natural (PLN), sobretudo quando se busca interpretações mais robustas, precisas e contextualmente sensíveis. Tendo em vista que a linguagem dos tweets (renomeados para posts) carrega informações contextuais, afetivas e discursivas relevantes, por vezes codificadas de forma não convencional, ferramentas capazes de processar esse tipo de dado em sua forma original têm sido desenvolvidas. Isso busca superar as limitações dos métodos e modelos existentes, que, em sua maioria, foram treinados com textos em linguagem formal e, portanto, pouco eficazes diante das particularidades dos tweets (Sanguinetti et al., 2023).

A não canonicidade do tweet enquanto "gênero textual"[2] evidencia-se, na ortografia, pela variabilidade lexical, abreviações e contrações informais, neologismos, repetição de letras, uso criativo de maiúsculas, elementos próprios da plataforma (hashtags, menções, retweets e truncamentos lexicais) e de domínio específico (p. ex., cashtags do mercado financeiro) (Zappavigna, 2012; Eisenstein, 2013; Freitas; Barth, 2015; Cardoso, 2019). Na sintaxe, os tweets podem apresentar fragmentação, como sequências de blocos informacionais justapostos, elipses e truncamentos (Sanguinetti et al., 2023).

[1] Conteúdo Gerado por Usuário (CGU) se refere a todo tipo de conteúdo criado e publicado por usuários de plataformas online, por exemplo, Twitter, Instagram, Whatsapp, etc. Esse conteúdo publicado pode ser gerado no formato de texto, vídeo, imagem ou áudio.

[2] Gêneros textuais são formas de uso da linguagem que se constituem historicamente e se estabilizam em função das atividades comunicativas que as originam (Marcuschi, 2008).

Nesse contexto, o desenvolvimento de ferramentas de PLN, como as de *tagging*[3] e *parsing*[4] voltadas ao CGU — em especial a tweets — demanda a disponibilização de corpora anotados, sobretudo do tipo gold standard. Por apresentarem anotações consistentes, geralmente oriundas de processos manuais ou semiautomáticos, esses corpora viabilizam a formulação de regras para abordagens simbólicas e o treinamento supervisionado de modelos baseados em Aprendizado de Máquina (Duran; Pardo, 2024).

A grande maioria dos corpora de tweets possui anotação segundo o modelo gramatical Universal Dependencies (UD) (Nivre et al., 2020), que prevê dois níveis de descrição: (i) morfológico, que compreende lema, categoria gramatical (Part-of-Speech ou PoS tag) e traços gramaticais; e (ii) sintático, que consiste nas relações de dependência entre as palavras ou tokens[5]. Uma vez anotado com essas informações, o corpus passa a ser um *tweebank*, isto é, um treebank de tweets. Uma vez enriquecidos com essas informações linguísticas explícitas, os corpora passam a ser classificados como *tweebanks*.

Para a construção de *tweebanks*, tem-se no trabalho de Sanguinetti et al. (2023) o principal conjunto de diretrizes de anotação UD adaptadas ao CGU, uma vez que as diretrizes originais foram definidas para textos formais. Tais diretrizes baseiam-se em um estudo linguístico prévio no qual as características léxico-ortográficas típicas do CGU foram organizadas em uma taxonomia de tipos (variedade do fenômeno) e subtipos (subcategorização de cada tipo), com base em duas dimensões: canonicidade e intencionalidade (Sanguinetti et al., 2023). O disfarce (disguise), como em “p**a” (“puta”), por exemplo, foi classificado como um subtipo do fenômeno de inovação lexical, que é intencional e não canônico (cf. Figura 1). Para os casos de disfarce, a proposta de diretriz para anotação UD envolve normalizar o lema (isto é, colocá-lo na forma canônica), associar a etiqueta PoS e a relação de dependência correspondentes à palavra no contexto e indicar adicionalmente a forma padrão da palavra e a classificação do fenômeno (no caso, NonCan=SpellVar).

---

[3] *Tagging* é o processo de atribuir rótulos a cada *token* de uma sentença que indicam categorias gramaticais (“partes do discurso”) (Jurafsky; Martin, 2025).

[4] *Parsing* é a tarefa de reconhecer a estrutura sintática de uma sentença, isto é, determinar sua estrutura sintagmática ou as relações de dependência (Jurafsky; Martin, 2025).

[5] *Token* corresponde a uma unidade segmentada do texto que serve de base para a anotação morfossintática e sintática, podendo representar palavras, sinais de pontuação ou outras formas gráficas relevantes. Em corpora de Conteúdo Gerado por Usuário, como tweets, é muito presente em formas não canônicas como abreviações, hashtags, emojis e outros elementos característicos da escrita digital.

Embora Sanguinetti et al. (2023) tenham realizado um trabalho relevante, o emprego do critério de “intencionalidade” é bastante questionável, como apontam os próprios autores. Isso se deve ao fato de que, diante da incerteza inerente à interpretação e da natureza contextual dos tweets, a categorização dos fenômenos quanto à intencionalidade pode apenas ser inferida, já que não é possível confirmá-la exclusivamente a partir da superfície textual. Além disso, a tipologia não contempla plenamente as particularidades da plataforma, como hashtags, menções, retweets e truncamentos lexicais, nem as especificidades do domínio do mercado financeiro, uma vez que busca ser genérica, sistematizando os fenômenos mais recorrentes descritos na literatura.

Diante disso, neste trabalho investigou-se como as variações léxico-ortográficas, aqui denominadas fenômenos idiossincráticos, manifestam-se no corpus DANTEStocks, composto por 4.048 tweets do mercado financeiro em português (Di Felippo; Roman, 2025). Trata-se do primeiro *tweebank* com múltiplas anotações gold standard: (i) morfológica e sintática, segundo o modelo UD; (ii) emoções, segundo o modelo Wheel of Emotions de Plutchik e Kellerman (1986); e (iii) entidades nomeadas, segundo as categorias genéricas do Segundo HAREM (Mota; Santos, 2008). As anotações de PoS e de dependências sintáticas do DANTEStocks já subsidiaram o desenvolvimento de diversas ferramentas para o português (Silva et al., 2020; Di Felippo et al., 2024; Barbosa, 2024).

A análise manual de uma parcela do corpus permitiu identificar uma série de fenômenos, que foram organizados em uma tipologia. As dimensões principais dessa tipologia, fundamentadas nos conceitos de “variação linguística” e “norma” (Bagno, 2007; Labov, 1972; Cagliari, 1998; Coelho, 2010, 2014), foram denominadas Norma Padrão e Norma Inovadora. Para a proposição das classes, tipos e subtipos da Norma Padrão, em particular, adotaram-se critérios objetivos, a saber: o conceito de “caractere” do padrão Unicode[6] (Moran; Cysouw, 2025) e as operações básicas de edição de palavras propostas por Damerau (1964).

A tipologia foi então empregada para a anotação manual do corpus DANTEStocks. Os resultados preliminares do corpus parcialmente anotado foram publicados em Scandarolli et al. (2023). O presente trabalho visa expandir essa publicação, apresentando uma caracterização linguística e estatística dos tweets do

[6] http://www.unicode.org/standard/WhatIsUnicode.html

mercado financeiro anotados em função das variações léxico-ortográficas e das categorias gramaticais a elas associadas.

Para apresentar a pesquisa, este relatório está organizado em seis seções. Na Seção 2, apresenta-se uma breve revisão da literatura sobre CGU, o gênero “tweet” e a caracterização de fenômenos léxico-ortográficos. Na Seção 3, descreve-se o corpus DANTEStocks, com especial atenção às decisões de pré-processamento, sobretudo às diretrizes de tokenização UD, que mantêm relação direta com os fenômenos léxico-ortográficos presentes no recurso. Na Seção 4, apresenta-se a proposta de tipologia, com destaque para os pressupostos teóricos que a fundamentam. Na Seção 5, discorre-se sobre a anotação e a caracterização linguística do corpus DANTEStocks à luz da tipologia proposta. Por fim, na Seção 6, apresentam-se as considerações finais do trabalho, enfatizando suas contribuições, limitações e perspectivas de trabalhos futuros.

2. Revisão da literatura

2.1. Conteúdo Gerado Por Usuário (CGU)

O avanço da comunicação digital, especialmente a partir do surgimento e crescimento das redes sociais, *blogs*, fóruns e outras mídias interativas, fomentou uma transformação profunda na forma como os indivíduos produzem e compartilham conteúdo. Nesse cenário, cunhou-se o termo *user-generated content* (UGC)*,* isto é, "conteúdo gerado por usuários" (CGU), para se referir a todo tipo de conteúdo criado e publicado por usuários da *web* na forma de texto, vídeo, imagem ou áudio (Krumm *et al.*, 2008).

No que tange ao conteúdo textual, o termo CGU recobre um *continuum* de subgêneros segundo Sanguinetti *et al*. (2023), o qual, no geral, caracteriza-se pelo uso da linguagem do cotidiano nas mídias digitais. A variação dos subgêneros textuais de CGU depende, em grande medida, das convenções e limitações impostas pelo meio ou plataforma em que são veiculados. Embora possuam fenômenos característicos gerais e reconhecidos, conforme apontado por diversos autores (p.ex.: Foster, 2010; Eisenstein, 2013; Sanguinetti *et al.*, 2023), a natureza informal e a amplitude dos subgêneros tornam o seu processamento automático uma tarefa complexa.

A crescente importância do conteúdo veiculado pelos diferentes subgêneros CGU e a linguagem não canônica que o caracteriza motivaram a construção de *corpora* anotados de CGU em várias línguas. No período de 2011 a 2019, Sanguinetti *et al.* (2023) identificaram 30 *corpora* de CGU com anotação sintática de referência construídos para diversas línguas europeias, inglês americano, árabe, chinês, híndi e outras (Figura 1).

A maioria desses *corpora* é composta, parcial ou totalmente, por *posts* extraídos do Twitter/X e segue o modelo gramatical UD. Além do alcance das opiniões veiculadas na plataforma, outras razões para a proeminência dos *tweebanks* foram a facilidade de obtenção dos dados via *Application Programming Interface* (API) e a política de uso dos dados para fins acadêmicos adotada até muito recentemente pela plataforma. Até 2023, pesquisadores acadêmicos tinham acesso gratuito à API para coletar grandes volumes de dados, sendo possível acessar/compilar todos os tweets públicos desde 2006, sem limite de tempo.

2.2. O Gênero Textual “Tweet”

O Twitter, renomeado para X em 2023, é uma rede social e plataforma de *microblogging*, que permite a publicação de mensagens curtas[7] chamadas *tweets* (agora, *posts*). Desde seu lançamento em 2006, o Twitter se consolidou como um importante canal de comunicação em tempo real. Segundo o DataReportal[8] – plataforma *online* que oferece relatórios detalhados e *insights* sobre o cenário digital global –, o Twitter atingiu 586 milhões de usuários ativos em janeiro de 2025, ocupando a 7ª posição no *ranking* mundial das mídias sociais. No Brasil, há 16 milhões de usuários, sendo a 9ª mais popular[9].

Um *tweet* é uma unidade discursiva autônoma ou parte de uma sequência (*thread*), podendo conter texto, imagens, vídeos, *links* e outros elementos interativos (como *hashtags*[10], menções, *emojis,* etc.). Diante disso, diz-se que, enquanto (sub)gênero textual, ele é uma forma breve, multimodal e altamente contextualizada de comunicação digital. Embora inclua características de outros gêneros (como notícia, *blog*, SMS (*short message service*), conversa informal, bilhete, citação etc.), o *tweet* possui características particulares que o distinguem dos demais gêneros digitais, sobretudo pela limitação de espaço, dinamismo e uso frequente de recursos interacionais e hipertextuais. Autores como Eisenstein (2013), Zappavigna (2012) e Cardoso (2019) destacam que o *tweet* é um gênero adaptado ao ritmo e às práticas da cultura digital, que se insere em um ecossistema comunicativo altamente interacional e efêmero.

A combinação entre esses gêneros menos formais e as características da plataforma favorece a predominância da informalidade no Twitter. Como destacam Freitas e Barth (2015), ainda que muitos *tweets* apresentem traços da norma culta, o formato reduzido limita o uso de construções mais elaboradas, incentivando a inserção de hipertextos, *links* e construções reduzidas. Soma-se a isso o fato de que a comunicação ocorre em um ambiente em que os interlocutores compartilham um repertório digital comum, o que permite o uso de códigos, gírias e referências

[7] A partir de 2017, o limite de caracteres para cada *tweet* passou de 140 para 280, podendo ser maior com assinaturas ou em *threads.*

[8] https://datareportal.com/reports/digital-2025-global-overview-report

[9] https://datareportal.com/reports/digital-2025-brazil

[10] *Hashtag* é uma etiqueta textual precedida pelo símbolo # (cerquilha), usada para marcar palavras-chave ou tópicos em plataformas de mídias sociais como o Twitter/X.

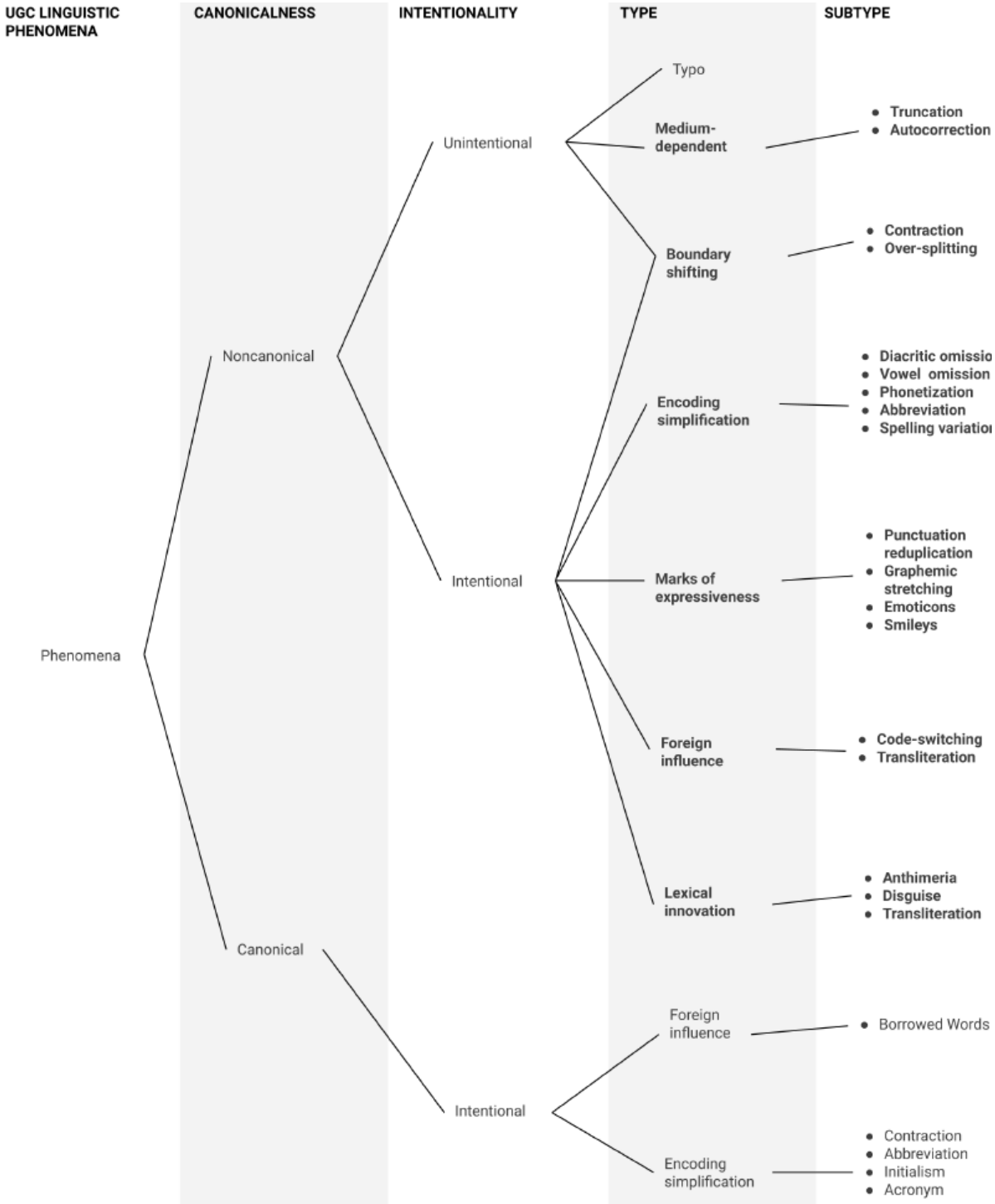
específicas sem necessidade de explicitação. Essa informalidade não apenas contribui para a agilidade na troca de informações, mas também reforça uma atmosfera discursiva marcada pela espontaneidade, autenticidade e proximidade entre os usuários.

Em suma, pode-se dizer que o *tweet* apresenta as seguintes características gerais: (i) extensão reduzida; (ii) estrutura multimodal; (iii) interatividade intensa, por meio de curtidas, retuítes, respostas e menções; (iv) marcas de oralidade e informalidade, como pontuação não convencional, vocabulário coloquial e *emojis*; (v) forte hipertextualidade, com uso de *hashtags* e URLs; (vi) organização seriada, por meio de *threads* que ampliam o conteúdo; (vii) temporalidade marcada, com foco em eventos do presente; (viii) ampla variação funcional, incluindo usos informativos, opinativos, promocionais, humorísticos, narrativos e militantes.

2.3. Caracterização Léxico-Ortográfica De CGU/Tweets

Sanguinetti *et al.* (2023) organizaram as particularidades lexicais e ortográficas de CGU (especialmente, *tweets*) em uma taxonomia de tipos (variedade do fenômeno) e subtipos (subcategorização de cada tipo) com base em duas dimensões: canonicidade e intencionalidade (Figura 1). Por “canonicidade”, entende-se à ocorrência ou não do fenômeno em textos formais. “Intencionalidade”, por sua vez, diz respeito ao fato de o fenômeno ter ocorrido de forma inadvertida/acidental ou não. Na Figura 2, lista-se exemplos dos diferentes fenômenos em várias línguas, sem, entretanto, indicação das dimensões canonicidade, intencionalidade e tipos.

Figura 1. Diagrama dos fenômenos linguísticos em UGC**.**



Fonte: Sanguinetti et al. (2020) https://arxiv.org/abs/2011.02063.

Figura 2. Exemplos de fenômenos CGU em diferentes línguas**.**

Phenomenon	Lang	Attested example	Standard form	Gloss
Encoding simplification				
Diacritic omission	GA	*Leigh aris!*	*Léigh arís!*	'Read again!'
	TR	*Istanbuldaki agaclar*	*İstanbul'daki ağaçlar*	'trees in Istanbul'
Vowel omission	EN	*ppl*	*people*	'people'
	TR	**slm**	*selam*	'hi'
Phonetization	EN	*Happy Birthday* **2** *me*	*Happy Birthday to me*	'Happy Birthday to me'
	TR	**1** *az*	*biraz*	'some'
	DE	*k* **1** *Mensch hat so* **1** *Thailandhass*	*kein Mensch hat so einen Thailandhass*	'nobody has such a hatred of Thailand'
Spelling variation	FR	*je* **sé**	*je sais*	'I know'
	GA	**gura** *míle*	*go raibh míle*	'thank you very much'
	FR	*tous mes* **examen** **son** *normaux*	*tous mes examens sont normaux*	'All my examinations are normal'
	IT	**anno** *mangiato*	*hanno mangiato*	'(they) have eaten'
Abbreviation	EN	**govt**	*government*	'government'
	DE	**zuggm**	*zugegebenermaßen*	'admittedly'
Boundary shifting				
Contraction	FR	**nimp** *quoi*	*n'importe quoi*	'rubbish'
Over-splitting	FR	**c a dire**	*c'est-à-dire*	'namely'
	TR	**gele bilirim**	*gelebilirim*	'I can come'
Marks of expressiveness				
Punct. reduplication	FR	*Joli* ***!!!!!!***	*Joli !*	'nice!'
	IT	*chi* ***?!?!?!***	*chi?*	'who?'
Case variation	GA	*is* ***BREÁ le daoine***	*is breá le daoine*	'people love'
Graphemic stretching	EN	*super***rrrrrrrrr**	*super*	'great'
	IT	***siiiiiiiiiiii***	*sì*	'yes'
Emoticons/smileys	-	*:-)* *<3*	–	–
	GA	*<3 mór*	*Grá mór*	'Lots of love'
Lexical innovation				
Disguise	IT	***caxxo***	*cazzo*	'fuck'
	TR	***mok / b.k / b*k***	*bok*	'shit'
	DE	***Verfi**t*** lange Reise	*Verfickt lange Reise*	'fucking long trip'
Anthimeria	IT	**tuittare**	*twittare*	'to tweet'
	EN	feel free to *PM*	*personal message*	'to send a message'
	DE	**achtisch**	*EN eightish*	'about 8 o'clock'
Foreign language influence				
Transliteration	GA	**áicbheaird**	*amscaí*	'awkward'
	TR	**taymlayn**	*zaman akışı*	'timeline'
Medium-dependent phenomena				
Truncation	GA	*thart fa' 53* **nó...**	*thart fa' 53 nóiméad*	'over 53 mi...(minutes)'
Autocorrection	GA	**concise**	*coicíse*	'fortnight'

Fonte: Sanguinetti et al. (2020) https://arxiv.org/abs/2011.02063.

Os fenômenos não canônicos (parte superior da Figura 1) e não intencionais, isto é, que ocorrem tipicamente em CGU de forma acidental, são de três tipos. Um deles são os erros de digitação *(typo*). Outro tipo são os fenômenos dependentes do meio (*medium-dependent*), isto é, aqueles realizados pela própria plataforma; seus subtipos são truncamento (isto é, quebra de uma palavra pelo limite de caracteres) e autocorreção, que pode consistir na filtragem ou substituição de palavras tabu ou ofensivas. O deslocamento de fronteira (*boundary shifting*) é o único fenômeno dependente do meio que pode ser não intencional ou intencional. Ele se refere a alterações no número de *tokens* comparado à ortografia padrão, sendo que seus subtipos são a contração (p.ex.: a expressão multipalavra do francês "*n'importe*" é contraída para "*nimp*") e a supersegmentação (*over-splitting*), que ocorre quando um único *token* da língua padrão é desmembrado em vários *tokens* (p.ex.: a expressão em francês "*c'est-à-dire*" ("isto é"/"ou seja") aparece supersegmentada em "*c a dire*").

Os fenômenos não canônicos e exclusivamente intencionais são:

a) simplificação de código, cujos tipos são (i) omissão de diacrítico (p.ex.: a expressão irlandesa *Léigh arís!* ("leia de novo") ocorre como *Leigh aris!*) e vogal (p.ex.: *people* ("pessoa") para *ppl*), (ii) fonetização[11] (p.ex.: a expressão em inglês *Happy Birthday to me* ("Parabéns pra mim") ocorre como *Happy Birthday 2 me*), (iii) abreviação (p.ex.: *government* ("governo") para *govt*) e (iv) variação de grafia (p.ex: *je sais* em francês ("eu sei") ocorrer como *je sé*).
b) marcas de expressividade (*marks of expressiveness*), que são usadas para indicar emoção ou ênfase; os tipos são (i) prolongamento grafêmico (p.ex.: *yesssss* ("sim")), repetição de pontuação (p.ex.: ?????) e *emoticons* (como xD para risada intensa) ou *emoji* (p.ex.: ☹).
c) influência estrangeira, cujos tipos são (i) mistura de línguas (*code-switching*) em um único *tweet* (p.ex.: "*non fare la bad girl*" ("não seja a garota má") substitui a forma padrão "*non fare la cattiva ragazza*") e (ii) transliteração[12] (p.ex.: "áicbheaird" parece uma forma de imitar foneticamente a palavra inglesa "awkward" em irlandês).

[11] Processo de representar sons da fala com símbolos ou grafias que imitam a pronúncia em vez da ortografia padrão.
[12] Processo de representar os sons de uma palavra de uma língua usando os caracteres de outro sistema de escrita, preservando sua forma fonética.

d) Inovação lexical, cujos tipos são (i) disfarce (*disguise*), isto é, formas alternativas, abreviadas ou censuradas para evitar o uso explícito de palavrões, seja por censura, eufemismo ou para suavizar a expressão (p.ex.: *caxxo* (“caralho/porra”) ao invés da forma padrão *cazzo* em italiano), (ii) anthimeria[13] (p.ex.: o verbo *to tweet* adaptado para *tuittare* em italiano) e o já explicado (iii) transliteração.

Os fenômenos canônicos, isto é, que também são observados em textos formais (na parte de baixo da Figura 1), são sempre intencionais, e podem ser de dois tipos: (i) “influência estrangeira” (*foreign influence*), que tem nos empréstimos lexicais seu único subtipo, e (ii) “simplificação de código” (*encoding simplification*), que envolve economia de esforço na escrita, como ocorre nos subtipos contração, abreviação, inicialismo e acrônimo.

Embora Sanguinetti *et el.* (2023) tenham feito um trabalho importante, o emprego do critério “intencionalidade” é bastante questionável, como apontam os próprios autores. Diz-se isso porque, diante da incerteza inerente à interpretação e a natureza contextual dos *tweets*, a categorização dos fenômenos quanto à intencionalidade pode ser apenas inferida, já que não é possível confirmá-la apenas pela observação da superfície do texto. Além disso, a tipologia não engloba fenômenos da plataforma, como *hashtags,* menções, *retweet* e truncamentos (lexicais) e do domínio do mercado financeiro.

Assim, para o *corpus* DANTEStocks, descrito na sequência, fez-se uma análise das particularidades léxico-ortográficas, sistematizando-as em uma tipologia hierárquica (cf. Seções 4 e 5).

[13] Anthimeria é uma figura de linguagem que consiste em usar uma palavra de uma classe gramatical para exercer a função de outra.

3. O *corpus* DANTEStocks

3.1. Caracterização geral do corpus

O DANTEStocks é um *corpus* de CGU com múltiplas camadas de anotação *gold standard* em língua portuguesa. A versão atual é composta por 4.048 *tweets* sobre o mercado financeiro, totalizando aproximadamente 81 mil *tokens*. Ele integra o *treebank* Porttinari[14], que está sendo desenvolvido no âmbito do projeto POeTiSA[15] (*POrtuguese processing – Towards Syntactic Analysis and parsing*) da frente de NLP2 (*Natural Language Processing for Portuguese*) do Centro de Inteligência Artificial (C4AI) da Universidade de São Paulo. O projeto POeTiSA visa contribuir para o avanço de recursos e ferramentas de PLN voltados à análise sintática do português por meio da criação do *corpus* multigênero Porttinari, anotado conforme o modelo UD. Em outras palavras, o DANTEStocks corresponde ao *subcorpus* de CGU que compõe o Porttinari, que atualmente está em sua versão 2.0.

Mais precisamente, o DANTEStocks resulta do refinamento e da anotação morfossintática-UD do *corpus* originalmente compilado por Silva *et al*. (2020) em 2014. A compilação dos *tweets* foi feita com base na ocorrência de ao menos um *ticker* de uma das 73 ações que compunham o índice Bovespa à época. Um *ticker* é o código alfanumérico (normalmente quatro letras e um número) que representa a empresa e o tipo da ação, como “PETR4”, que representa a ação preferencial da Petrobras. Os *tickers*, aliás, são comumente empregados no mercado financeiro em substituição aos nomes das empresas e organizações.

Como a compilação deste *corpus* ocorreu em 2014, a extensão máxima dos *tweets* do DANTEStocks é de 140 caracteres, devido à limitação imposta pela plataforma Twitter à época. Isso impacta diretamente a análise do *corpus*, pois muitas sentenças estão truncadas e fragmentadas, principalmente sentenças formadas por cabeçalhos de notícias que são postadas pelos usuários na plataforma.

Ademais, é importante destacar que os *posts* em questão estão em sua forma original, isto é, eles não foram submetidos sem nenhum processo de segmentação em unidades estruturais menores (sentenças ou sintagmas) ou mesmo de normalização lexical e frásica. Com isso, o *corpus* possui uma mistura de linguagem padrão e não-padrão.

[14] https://sites.google.com/icmc.usp.br/poetisa/porttinari-2-0

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

Essa mistura pode ser evidenciada pelo fato de que os *tweets* apresentam estruturais internas bastante variadas, incluindo (Di-Felippo *et al.*, 2021): (i) uma ou mais sentenças bem delimitadas (1) e (2); (ii) ausência de pontuação ou pontuação empregada inadequadamente (3) e (4); (iii) fragmentação textual (5); e (iv) colagens de trechos jornalísticos ou manchetes de outras fontes (6).

(1) Sera k petr4 já entrou na baixa?
(2) PETR4 subiu na bolsa 13,50. Muito bem, surpreso com o resultado.
(3) #PETR4 #PETROBRAS a R$13,13. Pronto! O #PT conseguiu fazer propaganda eleitoral antecipada O que a @user4 tem a dizer sobre isso?
(4) Bom dia Marcos, Alguma previsão para petr4?!
(5) #GGBR4 Suportes e resistências http://t.co/Azw6yIEVI9
(6) Logística, ex-LLX, anuncia prejuízo de R$ 135,8 milhões em 2013: A Prumo Logística, ex-LLX (LLXL3), divu... http://t.co/LwmlKPqssk.

Sobre seus aspectos lexicais, os quais têm relação direta com este trabalho, discorre-se, na sequência, sobre as estratégias empregadas no processo de *tokenização,* posto que algumas das características das palavras no DANTEStocks resultam desse processo. Embora o *corpus* possua várias camadas de anotação, como mencionado, descreve-se aqui apenas a gramatical segundo o modelo UD, uma vez que a anotação dos fenômenos léxico-ortográficos foi feita nos arquivos gerados por essa anotação e a seleção do conjunto de dados para a proposição da tipologia resultou do processo de anotação morfossintática.

3.2. *Tokenização*

A *tokenização* do *corpus* foi feita com base nos pressupostos do modelo UD, pois esse processo é condição para as anotações de PoS e de dependências sintáticas estabelecidas pelo modelo. A *tokenização* é o processo responsável por dividir o texto em um conjunto de segmentos significativos, chamados *tokens*. Assumindo a perspectiva lexicalista da sintaxe do modelo UD, as relações de dependência ocorrem entre palavras, o que significa que não há necessidade de segmentar as palavras em

morfemas. No entanto, é necessário delimitar as unidades básicas de anotação, denominadas “palavras sintáticas”[16].

Para *tokenizar* os 4.048 *tweets*, aplicou-se uma abordagem semiautomática, ou seja, tokenização automática seguida de revisão manual. Para isso, desenvolveu-se uma ferramenta baseada em regras chamada DANTE Tokenizer (Silva, E. *et al.*, 2020). Esse *tokenizador* é uma versão do TweetTokenizer[17] do NLTK ampliada com regras específicas para preservar o conteúdo original dos *tweets* e respeitar as diretrizes-UD propostas por Sanguinetti *et al*. (2023).

Com base nos pressupostos da UD, o *tokenizador* do DANTEStocks possui, por exemplo, uma regra para separar sinais de pontuação das palavras adjacentes quando estes formam um único *token* (exceto no caso de abreviações).

No que diz respeito às peculiaridades do português, a visão lexicalista da UD impõe a separação ou desmembramento de clíticos e contrações. Assim, um único *token* ortográfico como “fez-se” foi dividido em três *tokens* individuais (“fez” “-” “se”), uma vez que corresponde a múltiplas palavras (sintáticas). O mesmo ocorreu com contrações convencionais (p.ex. “na” > “em” “a”) e contrações não canônicas, que são muito frequentes no *corpus* (p.ex.: “pq” > “p” “q” e “oq” > “o” “q(ue)”).

No geral, a *tokenização* do *corpus* buscou seguir a delimitação original com base nos espaços em branco, incluindo casos de fonetização, *hashtags*, menções, *emoticons* e URLs. Para ilustrar isso, considere o uso não dos sinais de pontuação como pictogramas. O TweetTokenizer original do NLTK dividiria ocorrências como “:)” em vários sinais de pontuação (isto é, “:” e “)”). Como os *emoticons* podem substituir palavras da língua padrão, eles não devem ser divididos em mais de um segmento. Assim, uma regra foi adicionada para garantir o reconhecimento correto dessas cadeias de pontuação como *tokens* únicos ou individuais.

Regras semelhantes foram criadas para garantir o reconhecimento adequado de ocorrências de fenômenos específicos do domínio como *tokens*, incluindo *tickers*, *cashtags*[18], números decimais com parte fracionária indeterminada (p.ex.: “18,xx”) e

[16] Palavra sintática (do inglês, *syntactic word*) é a unidade mínima a que corresponde uma função sintática (https://universaldependencies.org/u/overview/tokenization.html).

[17] https://www.nltk.org/api/nltk.tokenize.html

[18] *Cashtag* é um marcador textual que utiliza o símbolo $ seguido do código de uma ação (por exemplo, “$Petr4”). Assim como as *hashtags* (#) agrupam tópicos gerais, as *cashtags* permitem a indexação e rastreamento de menções a ações nos *posts*, facilitando o monitoramento de tendências, análises e opiniões do mercado.

expressões temporais alfanuméricas. Para os *tickers,* por exemplo, uma regra foi criada para reconhecer cadeias alfanuméricas do tipo “petr4” (1) como um *token* único.

Por outro lado, mudanças de valores das ações (p.ex.: “+2,10%” > “+” “2,10” “%”) e valores monetários com formatos não convencionais (p.ex.: “R$20,00” > “R$” “20,00”) foram divididos em mais de um *token*. Tais ocorrências foram decompostas porque os símbolos matemáticos (como “+” e “-”), símbolo de porcentagem (%) e símbolos de moeda (como “R$”) podem ser substituídos por palavras comuns. Em resumo, a tokenização do DANTEStocks buscou preservou o conteúdo original dos *tweets*, com poucas exceções.

3.3. Anotação Gramatical

3.3.1. O modelo Universal Dependencies

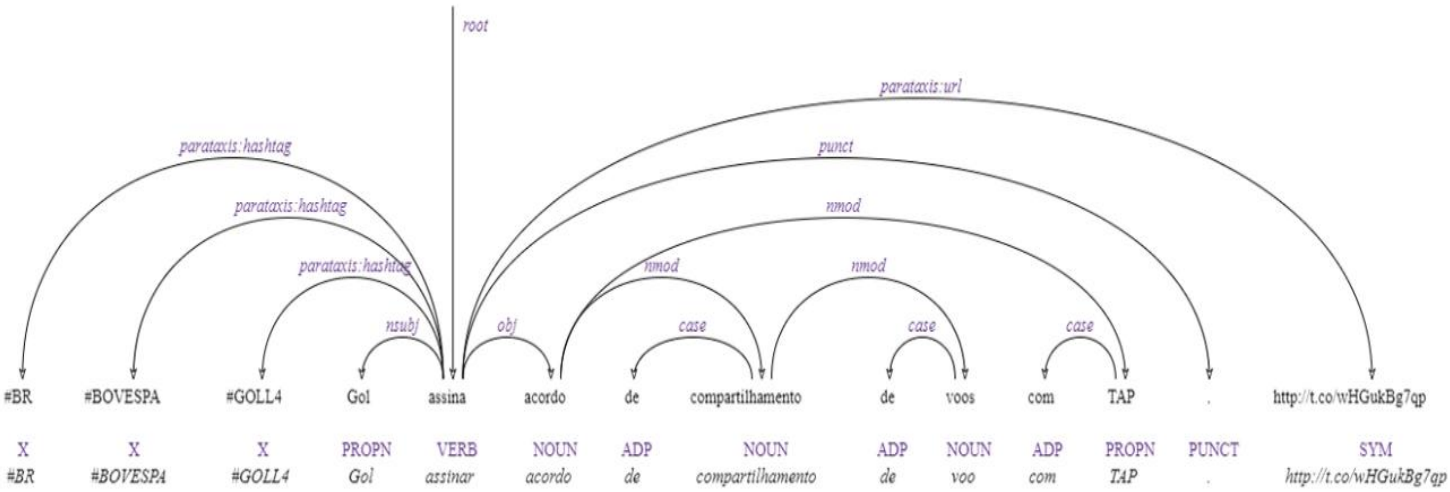
A maioria dos *tweebanks,* incluindo o DANTEStocks, possui anotação segundo modelo gramatical UD (Nivre *et al.,* 2020), que tem cada vez mais se torna uma referência padrão para anotação de *corpus* devido à sua adaptabilidade a diferentes domínios e gêneros (Sanguinetti *et al*., 2023). Trata-se de um modelo ou esquema de anotação gramatical voltado à representação estruturada e comparável dos elementos morfossintáticos das línguas naturais.

Esse modelo define a descrição em dois níveis: morfológico e sintático. No nível morfológico, cada unidade textual (*token*) é anotada com um lema (ou forma canônica), uma etiqueta de categoria gramatical (ou *tag* PoS) e um conjunto de traços morfológicos (*features*) que codificam informações gramaticais adicionais, como tempo, número, gênero, aspecto, entre outros. No nível sintático, prevê-se a descrição das relações de dependência (*deprel*) entre palavras/*tokens*. Uma dependência é estabelecida entre uma palavra sintaticamente dependente e outra palavra da qual ela depende (*head*). A versão atual da UD (v2) engloba 17 *tags* PoS[19] e 37 *deprels*[20].

Na Figura 3, ilustra-se a anotação-UD de um *tweet* do *corpus* DANTEStocks no formato arbóreo.

[19] https://universaldependencies.org/u/pos/index.html
[20] https://universaldependencies.org/u/dep/index.html

Figura 3. Exemplo de tweet com anotação-UD em formato de árvore.



Fonte: Barbosa (2024).

As etiquetas ou *tags* PoS são indicadas em caixa alta, como NOUN para “acordo”. Acima das formas flexionadas, estão os lemas, como “voo” para o plural “voos”. As relações de dependência sintática (*deprels*) são representadas pelas setas que partem do núcleo (*head*) e apontam para o dependente. A palavra “compartilhamento”, por exemplo, é dependente de “acordo” pela *deprel* ***nmod*** (modificador nominal), sendo essa relação mediada pela preposição “de”. A preposição, por sua vez, é dependente de “compartilhamento” e está ligada a ele pela *deprel* ***case***, que marca relações introduzidas por elementos funcionais como preposições. O verbo “assinou” constitui o *root* da estrutura, ou seja, a raiz sintática do *tweet*, servindo como núcleo a partir do qual se organizam as demais relações de dependência.

Como resultado da anotação-UD, tem-se um arquivo correspondente no formato CoNLL-U. Nele, o enunciado é representado em uma tabela de 10 colunas, em que cada linha corresponde a um *token* do enunciado (no caso, *tweet*) anotado. Na Figura 4, tem-se o arquivo CoNLL-U correspondente ao *tweet* da Figura 1. Da esquerda para a direita, as colunas correspondem às seguintes informações:

1. identificador do *token* (ID)
2. forma de superfície da palavra ou *token* (FORM)
3. lema (LEMMA)
4. etiqueta de classe gramatical universal (UPoS)

5. etiqueta específica da língua[21] (XPoS)
6. traços gramaticais (FEATS)
7. palavra à qual o *token* está subordinado sintaticamente (HEAD)
8. relação sintática estabelecida com o *head* (DEPREL)
9. relação *enhanced* (ou enriquecida)[22] (DEPS)
10. informações adicionais (MISC).

Figura 4. Exemplo de anotação de um tweet do DANTEStocks no formato CoNLL-U.

ID	FORM	LEMMA	UPoS	XPoS	FEATS	HEAD	DEPREL	DEPS	MISC
1	#BR	#BR	X	_	_	5	parataxis:hashtag	_	_
2	#BOVESPA	#BOVESPA	X	_	_	5	parataxis:hashtag	_	_
3	#GOLL4	#GOLL4	X	_	_	5	parataxis:hashtag	_	_
4	Gol	Gol	PROPN	_	_	5	nsubj	_	_
5	assina	assinar	VERB	_	Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin	0	root	_	_
6	acordo	acordo	NOUN	_	Gender=Masc|Number=Sing	5	obj	_	_
7	de	de	ADP	_	_	8	case	_	_
8	compartilhamento	compartilhamento	NOUN	_	Gender=Masc|Number=Sing	6	nmod	_	_
9	de	de	ADP	_	_	10	case	_	_
10	voos	voo	NOUN	_	Gender=Masc|Number=Plur	8	nmod	_	_
11	com	com	ADP	_	_	12	case	_	_
12	TAP	TAP	PROPN	_	_	6	nmod	_	SpaceAfter=No
13	.	.	PUNCT	_	_	5	punct	_	_
14	http://t.co/wHGukBg7qp	http://t.co/wHGukBg7qp	SYM	_	_	5	parataxis:url	_	SpaceAfter=No

Fonte: A autora (2025).

3.3.2. As anotações morfológica e sintática do DANTEStocks

A anotação morfológica também foi feita de forma semiautomática (Silva *et al*. 2021). As etiquetas de classe gramatical (PoS *tags*) geradas pelo analisador UDPipe 2 (Straka, 2018), treinado incrementalmente com base no UD-Portuguese Bosque (Rademaker *et al*. 2017) e em *tweets*, foram analisadas manualmente por três anotadores, e os casos de discordância entre eles foram resolvidos por um linguista sênior com base em diretrizes adaptadas para o português padrão (Duran, 2021) e para *tweets* (Di-Felippo *et al*., 2022). Assim como foi concebida, a anotação morfossintática não necessitou de métricas de concordância entre anotadores para medir sua confiabilidade. Como resultado, todas as 17 etiquetas do modelo UD ocorrem no DANTEStocks, sendo PUNCT, NOUN e PROPN as mais frequentes com cerca de 16%, 15% e 14% do total, respectivamente.

[21] Essas etiquetas não foram contempladas na anotação conduzida no POeTiSA.
[22] O DANTEStocks ainda não possui anotação dessas relações.

Os lemas e os traços gramaticais foram obtidos de forma semiautomática do léxico PortiLexicon-UD (Lopes *et al*., 2022). A revisão manual dos lemas foi extensa devido à elevada taxa de palavras *out-of-vocabulary.* Com relação aos traços gramaticais, o cenário foi diferente, pois a extração dessas informações do PortiLexicon-UD foi guiada pelas classes gramaticais e lemas previamente validados, o que reduziu bastante o esforço de revisão manual. A maioria das correções referiu-se a erros resultantes de ambiguidade dos traços verbais: forma verbal (*VerbForm*), modo (*Mood*), tempo (*Tense*), gênero (*Gender*), número (*Number*) ou pessoa (*Person*).

As relações de dependência foram anotadas em duas etapas semiautomáticas (Di-Felippo; Roman, 2025). Primeiramente, criou-se um *subcorpus gold-standard* com 1.000 *tweets*, anotado com o UDPipe 2 (treinado sobre o UD-Portuguese-Bosque) e revisado manualmente com base em diretrizes para o português padrão (Duran, 2022) e *tweets* (Di-Felippo *et al*., 2024). O restante do *corpus* foi anotado por meio da customização do *parser* Stanza (Qi *et al.,* 2020) para o DANTEStocks. A combinação do *corpus* Porttinari-base (Duran *et al.,* 2023) com o *subcorpus* de referência do DANTEStocks foi utilizada como conjunto inicial de treinamento do Stanza. O modelo de *parsing* resultante foi então utilizado para anotar automaticamente um primeiro lote de dados (dos 3.048 *tweets* restantes). Esse primeiro lote foi revisado manualmente e incorporado ao conjunto anterior, sendo utilizado para uma nova rodada de treinamento do Stanza. Esse ciclo iterativo de treinamento e anotação continuou de forma incremental até que o último (de um total de 6) lote fosse anotado e revisado.

Para fornecer uma medida de confiabilidade da anotação do DANTEStocks, um segundo especialista em PLN (também com experiência em anotação UD) revisou manualmente a anotação automática de 100 *tweets* aleatórios, com base nas mesmas diretrizes. As árvores de dependência analisadas por esse segundo anotador podiam pertencer ao *subcorpus* de referência ou ter sido geradas pelo Stanza em alguma de suas interações.

A medida de *Inter-Annotator Agreement* (IAA) foi calculada utilizando o coeficiente *Kappa* (Cohen, 1960; Carletta, 1996) em dois cenários distintos. No primeiro, o objetivo foi avaliar separadamente a anotação de *head* (palavra núcleo) e *deprel*. Os resultados do coeficiente *Kappa* foram de 0,96 para *head* e 0,97 para *deprel*. No segundo cenário, a avaliação considerou a combinação *head* + *deprel*,

obtendo um valor de Kappa de 0,95. O IAA por tipo de *deprel* foi medido por meio da métrica de *concordância total* (Sobrevilha Cabezudo, 2015), uma vez que o coeficiente *Kappa* não é adequado em casos de distribuição desequilibrada das relações. Obteve-se 100% de *concordância total* para mais da metade das 46 diferentes *deprels* (incluindo sub-relações) observadas na amostra de 100 *tweets*. Aliás, das 37 relações de dependência previstas no esquema UD, três não foram utilizadas no corpus (*clf, compound* e *dep*), sendo *punct* a mais frequente e *reparandum* a menos frequente.

4. Proposta de tipologia para fenômenos léxico-ortográficos

4.1. Materiais e métodos

O ponto de partida para a análise dos **fenômenos léxico-ortográficos**[23] e conseguinte proposição de uma tipologia foi um subconjunto de 1.363 *tokens* marcados como *Typo=Yes* durante a etapa de anotação morfossintática do *corpus.* Essa marcação foi feita porque tais *tokens* apresentavam ortografia não convencional para a qual ainda era preciso definir a forma mais adequada de anotação-UD. Isso quer dizer que, durante a fase de revisão conduzida pelos três anotadores, seguida pela adjudicação do especialista sênior, essas palavras/tokens foram identificadas e isoladas, para que pudessem ser adequadamente tratadas uma vez que diretrizes de anotação morfológica, incluindo lematização, categoria gramatical (PoS) e de traços gramaticais, fossem definidas. Mediante a análise manual linear dos 1.363 *tokens*, isto é, *token* por *token,* os fenômenos foram identificados e organizados em uma tipologia.

4.2. O modelo tipológico

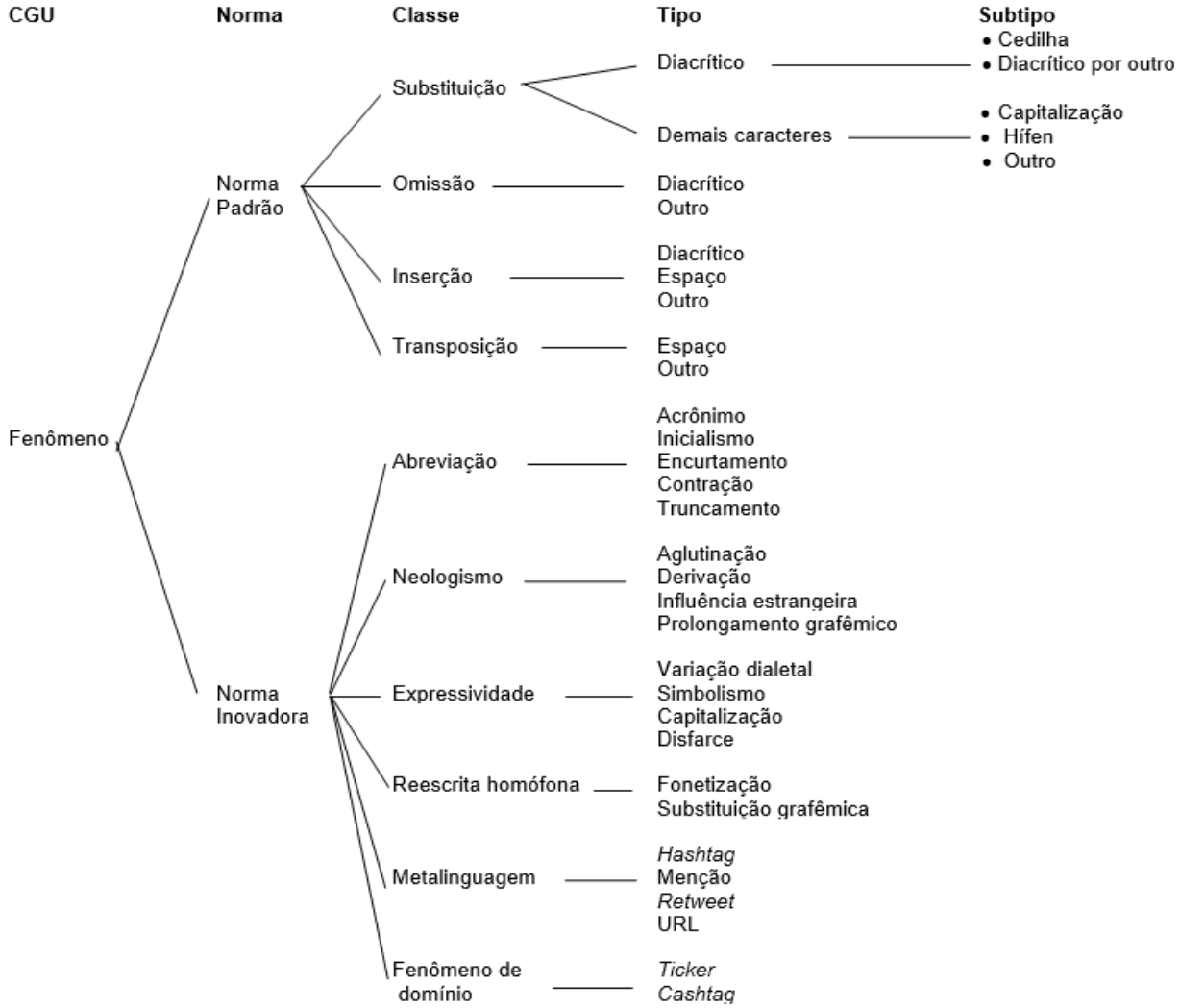
Com base nos conceitos de "variação linguística"[24] e "norma"[25] da Sociolinguística Variacionista (Bagno, 2007; Labov, 1972; Cagliari, 1998; Coelho, 2010, 2014), os fenômenos léxico-ortográficos não foram considerados "erros", mas variações linguísticas (ou idiossincrasias lexicais) que caracterizam o gênero *tweet*, e organizados em duas dimensões principais denominadas Norma Padrão e Norma Inovadora (Figura 5). A Norma Padrão engloba variações relativamente previsíveis quanto à ortografia padrão, isto é, engloba formas que se aproximam da canônica. A Norma Inovadora captura variações resultantes de práticas linguísticas que surgem espontaneamente entre usuários, com função expressiva, afetiva, humorística etc.

[23] Os **fenômenos léxico-ortográficos** correspondem a idiossincrasias linguísticas de natureza lexical no âmbito do PLN, as quais, em geral, são normalizadas e/ou padronizadas para possibilitar sua utilização em tarefas de aprendizado de máquina. Tais fenômenos referem-se a "palavras mal formadas" ou a "erros ortográficos" que não se enquadram no léxico e na gramática-padrão previamente estabelecidos nas bibliotecas básicas de PLN. Observa-se que idiossincrasias linguísticas ocorrem em todos os níveis de análise linguística; entretanto, neste trabalho, propõe-se a especificação e a descrição apenas das idiossincrasias lexicais, mais especificamente aquelas associadas à ortografia de cada token.

[24] Formas alternativas (ou variantes) de um mesmo item linguístico dentro de uma mesma comunidade de fala, condicionadas por fatores sociais, estilísticos ou linguísticos.

[25] Conjunto de regularidades compartilhadas por uma comunidade de fala sobre o uso linguístico.

Figura 5. Proposta tipológica de fenômenos léxico-ortográficos em tweets.



Fonte: Scandarolli *et al.* (2023).

4.2.1. Norma Padrão

Para organizar os fenômenos da Norma Padrão, adotou-se o trabalho de Damerau (1964), que foi pioneiro no PLN ao focar na verificação automática da ortografia (do inglês, *spell checking*). O objetivo do autor foi desenvolver um método computacional para detectar e corrigir automaticamente variações gráficas, tratadas por ele como "erros de digitação" (do inglês, *typographical error (typo*)), com base na comparação entre a palavra digitada e um vocabulário conhecido.

Damerau verificou que as formas variantes podiam ser explicadas com base no número mínimo de operações necessárias para transformá-las em palavras na forma padrão. Tais operações são substituição, omissão, inserção e transposição. A formalização (e implementação algorítmica) das 4 operações ficou conhecida como "distância de Damerau-Levenshtein" (Wagner; Fischer, 1974) e foi amplamente empregada em várias tarefas, como correção ortográfica automática (em inglês, *spell chekers*) e reconhecimento óptico de caracteres (do inglês, *Optical character recognition* – OCR), é considerado um clássico.

As 4 operações básicas foram definidas como "classes" da Noma Padrão (Figura 1). No entanto, diferentemente de Damerau, cujas operações são feitas sobre o conceito de "letra" (isto é, caractere ASCII), adotou-se aqui o conceito de "caractere" do padrão Unicode, que é a unidade mínima de escrita com significado semântico ou funcional, codificada com um valor numérico único (em inglês, *code point*). Considerando a forma de normalização do padrão denominada *Normalization Form Decomposed* (NFD) ("Forma de Normalização Decomposta"), o Unicode separa os caracteres compostos (como "ã") em partes básicas. Dessa forma, o conceito de "caractere" engloba letra, dígito, acento, sinal de pontuação, *emoji*, símbolo, espaço, caractere de controle (quebra de linha, tabulação) e outros.

Cada operação de Damerau atua então sobre caracteres Unicode e não letras e isso impacta a interpretação da operação/fenômeno. Ao se analisar, por exemplo, a palavra "mâe" considerando letras, a mudança em relação à forma canônica "mãe" é uma substituição simples de "ã" por "â". Porém, considerando os caracteres Unicode na forma NFD, essa alteração corresponde à troca do diacrítico combinante por outro (isto é, til (U+0303) pelo acento circunflexo (U+0302)), mostrando que a operação é, na verdade, a substituição do acento e não da letra base.

Com a adoção do conceito de "caractere", buscou-se garantir precisão e granularidade linguística na interpretação das operações, permitindo identificar padrões específicos de variação e auxiliando ferramentas automáticas, por exemplo, a interpretar o texto de forma mais eficiente. A seguir, descrevem-se as 4 classes e seus respectivos tipos e subtipos, tomando como ponto de partida as formas canônicas de referência. Vale ressaltar que a classe substituição é a única com subtipos.

Substituição

Um caractere é substituído por outro, mantendo a estrutura do token.

Tipos/Subtipos:

- **Substituição de Diacrítico:**
  - **Cedilha:** substituição do cê-cedilha (“ç”) por outro caractere, como em "acougue" ("açougue"); embora a cedilha (“¸”) seja um diacrítico com *code point* específico na representação NFD do Unicode, optou-se por dar a ela um tratamento particular, considerando o cê-cedilha como alvo da substituição, pois, de forma contrária a outros diacríticos no português, a cedilha não existe como sinal isolado, mas sempre em associação com o caractere “c” ou “C”, havendo nos teclados do formato ABNT2 (amplamente empregado no Brasil) uma tecla própria.
  - **Diacrítico por outro:** substituição de um diacrítico por outro, como em "mâe" (ao invés de "mãe"), em que o diacrítico “~” (til) foi substituído pelo acento circunflexo (“^”).
- **Substituição de Demais caracteres:**
  - **Capitalização:** “dilma” ao invés de “Dilma”
  - **Hífen:** “cruz credo” no lugar de “cruz-credo”
  - **Outro: “**hirário” em vez de “horário”.

Omissão

Um caractere é omitido.

Tipos:

- **Omissão de Diacrítico:** "esta" por "está"
- **Omissão de Outro (caractere):** "açõe" por "ações"

Inserção

Um caractere não previsto na grafia padrão é inserido.

Tipos:

- **Inserção de Diacrítico:** “Petrobrás” no lugar de “Petrobras”
- **Inserção de Espaço:** “a final” e não “afinal”
- **Inserção de Outro (caractere):** “*Streaddle*” ao invés de “*Straddle*”[26]

[26] Estratégia de negociação financeira que envolve a compra simultânea de uma opção de compra (call) e uma opção de venda (put) com o mesmo preço de exercício e data de vencimento (https://en.wikipedia.org/wiki/Straddle).

Transposição

A ordem de dois caracteres é invertida.

Tipos:

- **Transposição de Espaço:** “meua migo” no lugar de “meu amigo”
- **Transposição de Outro (caractere):** “acrodo” e não “acordo”

4.2.2 Norma Inovadora

A Norma Inovadora designa o conjunto de práticas linguísticas não canônicas que se manifestam de modo sistemático, recorrente e funcional em contextos de comunicação digital, especialmente no gênero textual tweet. Fundamentada nos pressupostos da Sociolinguística Variacionista (Labov, Bagno), essa noção parte do princípio de que a variação linguística é inerente ao uso da língua e que as diferentes normas refletem regularidades compartilhadas por comunidades de fala específicas, e não desvios aleatórios ou falhas individuais. Assim, a Norma Inovadora não é concebida como erro ortográfico, mas como uma norma de uso legitimada pelo contexto social, tecnológico e discursivo em que se insere.

Essa norma emerge como resposta às condições próprias do meio digital, tais como a limitação de caracteres, a velocidade da interação, a multimodalidade e a forte dimensão “interacional” e expressiva das redes sociais. Caracteriza-se pela flexibilização das convenções ortográficas tradicionais e pela aproximação entre escrita e oralidade, incorporando estratégias que visam economia linguística, expressividade, marcação identitária e adequação pragmática. Entre os fenômenos que a compõem estão abreviações, neologismos, reescritas fonéticas, variações grafêmicas, uso expressivo de maiúsculas, prolongamento de caracteres, simbolismo gráfico (emojis e emoticons), além de formas de censura criativa e disfarce lexical.

A Norma Inovadora abrange ainda elementos metalinguísticos e técnicos próprios da plataforma Twitter/X, como hashtags, menções, retweets e links/URLs, bem como vocabulário especializado do domínio discursivo em questão, como tickers e cashtags no contexto do mercado financeiro.

Como será revelado mais adiante, a predominância dessa norma no corpus analisado evidencia que a inovação linguística constitui o padrão dominante no Conteúdo Gerado por Usuário, reforçando a necessidade de modelos teóricos e computacionais que reconheçam essas formas como linguisticamente válidas. Desse

modo, a descrição da Norma Inovadora contribui tanto para uma compreensão mais fiel da linguagem digital contemporânea quanto para o desenvolvimento de ferramentas de Processamento de Linguagem Natural mais robustas e sensíveis à variação linguística. A seguir, descrevem-se as classes e seus respectivos tipos.

Abreviação

Forma reduzida de um token ou expressão.

Tipos:

- **Acrônimo:** *token* pronunciável como uma palavra, formado a partir das letras iniciais (ou partes) de uma expressão ou conjunto de palavras., p.ex.: “Cemig” (“Companhia Energética de Minas Gerais”)
- **Inicialismo:** *token* formado pelas letras iniciais de várias palavras e pronunciado letra por letra, p.ex.: “lp” advindo de “longo prazo”
- **Encurtamento:** *token* que se caracteriza pela ausência das letras finais de uma palavra/*otken,* como “ult” (ao invés de “última/último”)
- **Contração:** *token* que se caracteriza pela ausência de letras intermediárias, como em “enqt” (“enquanto”)
- **Truncamento:** *token* quebrado no final do *tweet* por limite de caracteres, geralmente seguido de reticências, p.ex.: “divu (...)” (“divulgou”)

Neologismo

Palavras novas ainda não institucionalizadas.

Tipos:

- **Aglutinação:** *token* resultante da junção de outras palavras, como “Ibolixo” (“Ibovespa” + “lixo”)
- **Derivação:** adição de afixo (formal ou informal) a uma raiz existente, como em “diretassa”
- **Influência estrangeira:** *token* criado por influência de outra língua; p.ex.: “estopar” é uma forma aportuguesada do verbo “*to stop*” em inglês, com origem nas expressões “*stop loss*” e “*stop gain*”, que no mercado financeiro são ordens de venda.

Expressividade

Elementos gráficos que transpõem à linguagem escrita aspectos da oralidade e da emoção (Crystal, 2006).

Tipos:

- **Prolongamento:** repetição de caracteres, comumente vogais, como em “noooossaaa” (“nossa”)
- **Variação dialetal:** *token* que reflete o modo como certos grupos pronunciam a palavra em contextos regionais ou populares, como “malmita” (“marmita”), em que houve a troca dos caracteres (consoantes líquidas) “l” e “r”.
- **Simbolismo:** caractere simbólico em substituição a uma palavra (como os *emojis* e *emoticons*) ou parte dela (p.ex.: “m+” ao invés de “mais”)
- **Disfarce:** substituição de letras por *caracteres* especiais para censura, como em “m*” (“merda”)
- **Capitalização:** que se caracteriza pelo uso de maiúsculas para expressar ênfase, como em “Mensal da PETROFUMO”[27]

Reescrita Homófona

Variações motivadas pela fonética ou simplificação de escrita.

Tipos:

- **Fonetização:** ocorre quando a forma escrita tenta representar diretamente os sons da fala, como “krai” (“caralho”)
- **Substituição grafêmica:** fenômeno em que uma letra é usada para substituir sinais gráficos (diacríticos), que normalmente marcam tonicidade ou timbre na norma padrão, como “neh” (“né”)

Metalinguagem

Elementos próprios da linguagem de plataforma social Twitter.

Tipos:

[27] “Petrofumo**”** (“petro” (petróleo) + “fumo” (fumaça)) sugere algo que “queima” dinheiro ou valor, ou “fumaça” como símbolo de algo nebuloso, problemático, ou que gera perdas.

- ***Hashtag*:** palavra ou expressão precedida do símbolo “#”, usada para marcar ou agrupar conteúdos em plataformas digitais como o Twitter, p.ex.: “#PT”
- **Menção:** símbolo “@” seguido do nome de um usuário, p.ex.: “@user”
- ***Retweet*:** emprego da abreviação “RT” antes de um *post* para sinalizar que o conteúdo foi replicado
- **Link/URL:** endereço eletrônico de uma página na internet, usado para compartilhar conteúdo, como nos exemplos (5) e (6)

Fenômeno de Domínio

Vocabulário específico do mercado financeiro.

Tipos:

- ***Ticker*:** “PETR4”
- ***Cashtag*:** “$PETR3”

É importante ressaltar que a diferença entre Capitalização sendo tipo de Expressividade e Capitalização sendo do subtipo de Substituição por Demais Caracteres reside na natureza e definição das diferentes normas. Na Norma Padrão, a Capitalização subtipo de Substituição implica que no âmbito da tipologia proposta, um caractere Unicode foi substituído por outro na sua versão capitalizada, porém não mantendo uma relação direta e previsível com a ortografia canônica da língua, isto é, com as formas legitimadas por gramáticas, dicionários e convenções ortográficas institucionalizadas. Enquanto por outro lado, o recurso de Capitalização tipo de Expressividade, ocorre como um recurso expressivo característico do meio digital que normalmente demonstra ênfase, exaltação ou exagero.

5. Anotação do DANTEStocks

5.1. Seleção dos Dados e Metodologia

Uma vez que a tipologia foi definida, procedeu-se à anotação efetiva dos primeiros 1.069 *tweets,* o que equivale a aproximadamente 26% do *corpus*. Neste subcorpus foi feita a anotação sobre um arquivo CoNLL-U, convertido em tabela (arquivo .xlsx)[28] e seguindo uma estratégia linear, foi anotado *tweet* a *tweet* (e *token a token*) de cima para baixo e da esquerda para a direita.

O arquivo CoNLL-U utilizado na anotação (Figura 6) difere-se do arquivo original (Figura 4) pelo ocultamento de algumas colunas, a saber: colunas 5 (XPoS), 7 (HEAD), 8 (DEPREL) e 9 (DEPS) e, ademais, pela inserção de *quatro novas colunas à direita*, cada uma delas destinada às informações previstas na tipologia, isto é, **norma**, **classe**, **tipo** e **subtipo** de um fenômeno.

Figura 6. Arquivo CoNLL-U modificado para a anotação.

	A	B	C	D	F	J	K	L	M	N
1	ID	FORM	LEMMA	UPOS	FEATS	MISC	Norma 1	Classe 1	Tipo 1	Subtipo 1
25011	# sent_id = dante_01_446339383189577728l									
25012	# text = olha os R$13,50 se aproximando na #PETR4 ! Uia ! Não esperava ver isso acontecendo h									
25013	1	olha	olhar	VERB	Mood=Imp|Nu	_				
25014	2	os	o	DET	Definite=Def|	_				
25015	3	R$	R$	SYM	_	_				
25016	4	13,5	13,5	NUM	NumType=Ca	_				
25017	5	se	se	PRON	Case=Nom|P	_				
25018	6	aproximando	aproximar	VERB	VerbForm=Ge	_				
25019	7-8	na	_	_	_	_				
25020	7	em	em	ADP	_	_				
25021	8	a	o	DET	Definite=Def|	_				
25022	9	#PETR4	#PETR4	PROPN	_	_	Innovative...	Metalingua...	Hashtag	
25023	10	!	!	PUNCT	_	_				
25024	11	Uia	uia	INTJ	_	_				
25025	12	!	!	PUNCT	_	_				
25026	13	Não	não	ADV	_	_				
25027	14	esperava	esperar	VERB	Mood=Ind|Nu	_				
25028	15	ver	ver	VERB	VerbForm=Inf	_				
25029	16	isso	isso	PRON	Gender=Mas	_				
25030	17	acontecendo	acontecer	VERB	VerbForm=Ge	_				
25031	18	hoje	hoje	ADV	_	_				
25032	19	não	não	ADV	_	_				
25033	20	...	...	PUNCT	_	_				
25034	21	o.O	o.O	SYM	_	SpacesAfter=\n	Innovative...	Expressivi...	Simbol...	

Fonte: A autora (2025).

[28] A tabela onde se encontram os 1069 tweets do DANTEStocks anotados segundo a metodologia descrita neste trabalho, assim como as estatísticas que foram apuradas consequentemente à anotação se encontram no link https://docs.google.com/spreadsheets/d/1JNppB7prBS0kkuaKGl0c-TbiCL4n5uW_K7i2pS-QotE/edit?usp=sharing .

Há ainda a possibilidade da ocorrência de dois fenômenos diferentes para um mesmo token, seja Norma Padrão ocorrendo junto com Norma Inovadora, dois fenômenos diferentes de Norma Padrão ou dois fenômenos de Norma Inovadora concomitantemente (Figura 7 - observado no *token* “OBJ”). Para este caso, acrescenta-se outras *quatro colunas* contendo as categorias da tipologia à direita. É possível encontrar no corpus anotado até três fenômenos diferentes concomitantes para um único token, esses casos serão discutidos mais adiante na seção 5.2.3.

Figura 7. Exemplo de *token* com variação de Norma Inovadora e Norma Padrão.

	A	B	C	D	F	J	K	L	M	N	O	P	Q	R
1	ID	FORM	LEMMA	UPOS	FEATS	MISC	Norma 1	Classe 1	Tipo 1	Subtipo 1	Norma 2	Classe 2	Tipo 2	Subtipo 2
24711	# sent_id = dante_01_4463294035100999681													
24712	# text = #petr4 não é lindo? OBJ CUMPRIDO!!!! RT @Live_Trade: 13,28 no curto 5													
24713	1	#petr4	#petr4	PROPN	_	_	Innova…	Metalin…	Hashtag					
24714	2-3	né	_	_	_	_								
24715	2	não	não	ADV	_	_								
24716	3	é	ser	AUX	Mood=Ind|Nu	_								
24717	4	lindo	lindo	ADJ	Gender=Mas	_								
24718	5	?	?	PUNCT	_	_								
24719	6	OBJ	objetivo	NOUN	Gender=Mas	FullForm=objetiv	Innova…	Expres…	Capita…		Innov…	Abre…	Encu…	
24720	7	CUMPRIDO	cumprir	VERB	Gender=Mas	_	Innova…	Expres…	Capita…					
24721	8	!	!	PUNCT	_	_								
24722	9	!	!	PUNCT	_	_								
24723	10	!	!	PUNCT	_	_								
24724	11	RT	RT	X	_	_	Innova…	Metalin…	Retweet					
24725	12	@Live_Trade	@Live_Trade	PROPN	_	_	Innova…	Metalin…	Menção					
24726	13	:	:	PUNCT	_	_								
24727	14	13,28	13,28	NUM	NumType=Ca	_								
24728	15-16	no	_	_	_	_								
24729	15	em	em	ADP	_	_								
24730	16	o	o	DET	Definite=Def|	_								
24731	17	curto	curto	ADJ	Gender=Mas	_								
24732	18	5	5	NUM	NumType=Ca	SpacesAfter=\n								

Fonte: A autora (2025).

5.2. Análise Estatística dos Dados Anotados

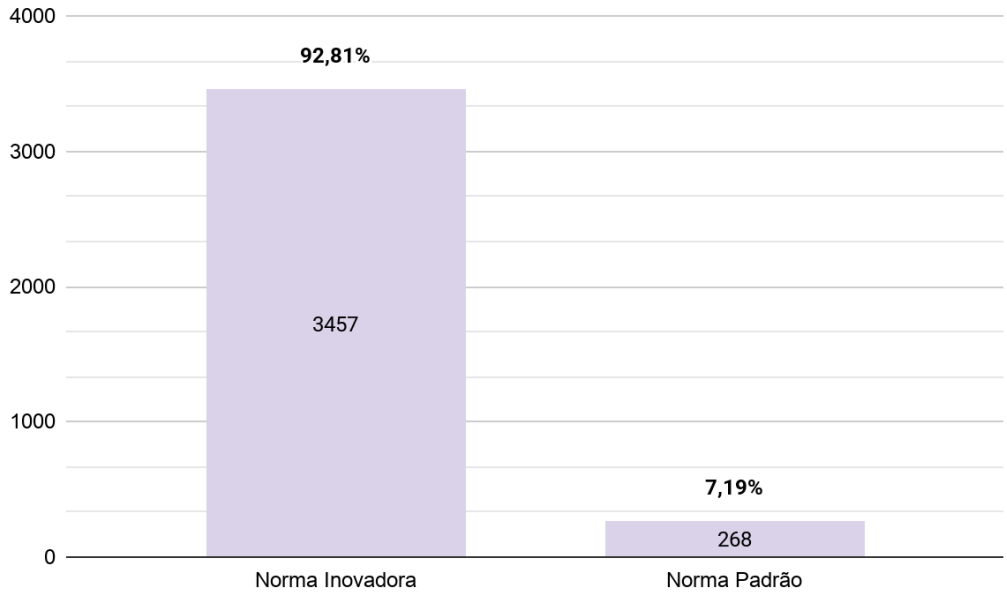
Nesta seção, apresenta-se a análise estatística dos dados anotados no *corpus* DANTEStocks, com o objetivo de identificar os fenômenos léxico-ortográficos idiossincráticos mais recorrentes e compreender sua distribuição em relação às categorias gramaticais do modelo UD (isto é, as *tags* PoS), às categorias da tipologia proposta e aos subtipos técnicos definidos para os fenômenos da Norma Padrão. A análise visa avaliar a aderência da taxonomia desenvolvida aos dados empíricos e explorar implicações linguísticas e computacionais das idiossincrasias anotadas.

5.2.1 Distribuição geral por norma

A análise da distribuição geral dos fenômenos anotados no corpus DANTEStocks evidencia uma predominância expressiva da Norma Inovadora em relação à Norma Padrão. Do total de fenômenos identificados no subcorpus anotado, 92,81%

correspondem à Norma Inovadora, enquanto apenas 7,19% foram classificados como pertencentes à Norma Padrão (Figura 8). Esses valores refletem um cenário em que as ocorrências desviantes em relação à ortografia normativa não se concentram majoritariamente em erros involuntários, mas em estratégias sistemáticas de escrita que se afastam conscientemente do padrão.

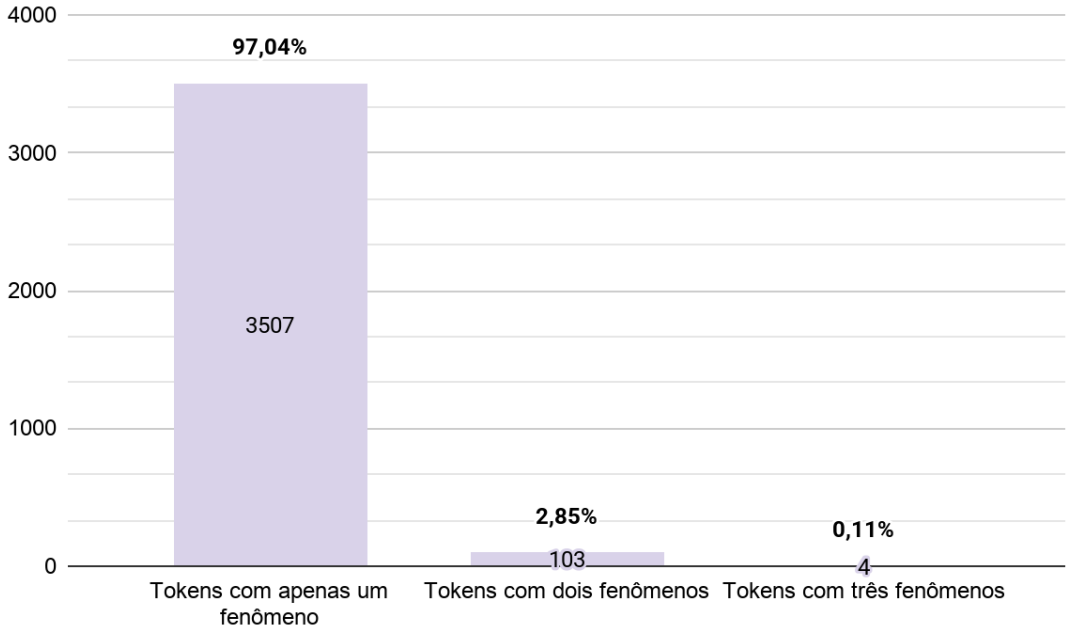
Figura 8. Distribuição do total de fenômenos categorizados por Norma.



Fonte: A autora (2025).

Essa predominância torna-se ainda mais significativa quando se considera a relação entre fenômenos e tokens. Foram anotados 3.725 fenômenos distribuídos em 3.614 tokens que estão contidos nos primeiros 1069 tweets do corpus DANTEStocks. Há no corpus anotado um total de 107 tokens em que mais de um fenômeno foi anotado. Desse total, 103 tokens tem dois fenômenos anotados para um único token (2,85% do total de tokens anotados) e 4 tokens tem três fenômenos (0,11% do total de tokens anotados) concomitantes. Isso indica que, na maior parte dos casos, cada token apresenta apenas um fenômeno idiossincrático. Embora numericamente minoritários, esses casos de sobreposição revelam que a escrita digital no domínio financeiro admite combinações complexas de estratégias linguísticas, aspecto que será explorado de forma mais detalhada na Seção 5.2.3.

Figura 9. Distribuição da quantidade de fenômenos para cada token anotado.



Fonte: A autora (2025).

A predominância de tokens com apenas um fenômeno anotado indica que, na maior parte dos casos, uma única categoria tipológica é suficiente para descrever determinadas ocorrências lexicais. No entanto, também se observa que diferentes categorias podem coexistir na caracterização de um mesmo item, o que culmina na possibilidade de sobreposição classificatória.

Esse fato se explica pela natureza do léxico como um sistema dinâmico e multifacetado, organizado segundo diferentes dimensões de análise, como as dimensões temporal, geográfica, social, funcional, estrutural e semântica. Cada uma dessas dimensões descreve propriedades distintas do mesmo objeto linguístico. Sob a perspectiva da Sociolinguística variacionista, o léxico constitui um domínio socialmente distribuído e historicamente mutável, cujas transformações decorrem do uso que os falantes fazem da língua em contextos específicos. Assim, as categorias analíticas empregadas para sua descrição precisam ser constantemente revistas e ajustadas, a fim de acompanhar a dinamicidade das mudanças semânticas, formais e funcionais que caracterizam o funcionamento do sistema lexical. (LABOV, 2008; AITCHISON, 2001; REY, 1995)

Cabe, contudo, destacar que a interpretação desses dados deve ser feita com

cautela metodológica. Uma vez que a anotação foi realizada por um único anotador e não foi calculada a concordância interanotador, não é possível descartar completamente a influência de julgamentos interpretativos individuais, sobretudo em casos limítrofes entre Norma Padrão e Norma Inovadora. Ainda assim, a magnitude da diferença observada entre as duas normas aponta para uma tendência robusta, que dificilmente se explicaria apenas por viés de anotação, reforçando a relevância da distinção proposta pela taxonomia adotada neste trabalho.

5.2.2. Distribuição das classes, tipos e subtipos

A distribuição interna das categorias revela diferenças estruturais consistentes entre a Norma Inovadora e a Norma Padrão, tanto do ponto de vista quantitativo quanto funcional. Não se trata apenas de uma distinção numérica, mas de uma divergência na própria organização dos fenômenos e nas lógicas que os estruturam. Enquanto a Norma Inovadora se concentra em classes diretamente vinculadas às condições de produção do discurso digital e às especificidades do domínio financeiro, a Norma Padrão apresenta predominância de alterações gráficas pontuais, sobretudo relacionadas à acentuação.

Quadro 1. Distribuição dos fenômenos da Norma Padrão.

Classe	Qt.	Tipo	Qt	Subtipo	Qt.	Exemplo	Lema
Substituição	54	Diacrítico	45	Cedilha	44	Abraco	abraço
				Diacrítico por outro	0	sõ	só
		Demais caracteres	9	Capitalização	0	MAravilha	maravilha
				Hífen	1	pré sal	pré-sal
				Outro	8	Aqwele	aquele
Omissão	169	Diacrítico	158			Economicos	econômico
		Outro	11			Aind	ainda
Inserção	44	Diacrítico	32			analisár	analisar
		Espaço	6			sub onda	subonda
		Outro	6			nóis	nós
Transposição	1	Espaço	0			--	--
		Outro	1			grnade	grande
TOTAL DE FENÔMENOS	**268**						

Fonte: A autora (2025).

Quadro 2. Distribuição dos fenômenos da Norma Inovadora.

Classe	Tipo	Exemplo	Lema	Qt.	Subtotal
Abreviação	Acrônimo	ICON	ICON	9	395
	Inicialismo	PB	price-to-book	70	
	Encurtamento	d	de	108	
	Contração	Abç	abraço	152	
	Truncamento	abai	abaixo	56	
Neologismo	Aglutinação	ibolixo	ibolixo	13	18
	Derivação	diretassa	direto	2	
	Influência estrangeira	stopando	stopar	3	
Expressividade	Prolongamento grafêmico	noosaaa	nossa	33	195
	Variação dialetal	ocê	você	19	
	Simbolismo	=)	=)	20	
	Capitalização	ALEGRIA	alegria	122	
	Disfarce	P**a	puta	1	
Reescrita homófona	Fonetização	hehehe	hehehe	19	37
	Substituição grafêmica	eh	ser	18	
Metalinguagem	Hashtag	#VALE5	#VALE5	731	1639
	Retweet	RT	RT	89	
	Menção	@user	@user	371	
	URL	http://t.co/03Aet66FTO	http://t.co/03Aet66FTO	448	
Fenômeno de domínio	Ticker	ABEV3	ABEV3	1090	1173
	Cashtag	$AEDU3	$AEDU3	83	
TOTAL DE FENÔMENOS					**3457**

Fonte: A autora (2025).

O Quadro 1 e a Figura 10 apresentam a distribuição dos 268 fenômenos classificados como Norma Padrão. Observa-se predominância clara da classe Omissão (169 ocorrências; 63,1%), sobretudo no subtipo Omissão de Diacrítico (158 casos). Em seguida, figuram Substituição (54 ocorrências; 20,2%) e Inserção (44 ocorrências; 16,4%), também com maior incidência de alterações envolvendo marcas acentuais. A

classe Transposição, com apenas uma ocorrência, revela impacto estatisticamente residual.

No Quadro 2 e na Figura 11, observa-se que a Norma Inovadora totaliza 3.457 ocorrências, distribuídas de maneira altamente concentrada em duas classes: Metalinguagem (1.639 ocorrências) e Fenômeno de Domínio (1.173 ocorrências). Juntas, essas categorias correspondem a aproximadamente 80% dos casos registrados. Essa concentração indica que a inovação linguística no corpus não se dispersa de modo aleatório pelo léxico comum, mas se organiza prioritariamente em torno de dois eixos estruturantes: os recursos próprios da arquitetura comunicativa da plataforma e as unidades lexicalizadas do campo financeiro.

No interior da classe Metalinguagem, ainda conforme o Quadro 2, destacam-se hashtags (731 ocorrências), URLs (448) e menções (371). A predominância desses tipos revela que parte expressiva da variação gráfica observada está vinculada a mecanismos de indexação temática, de endereçamento discursivo e de articulação intertextual. Esses elementos não apenas acompanham o conteúdo proposicional dos enunciados, mas integram a própria estrutura textual, configurando-se como componentes constitutivos da escrita digital.

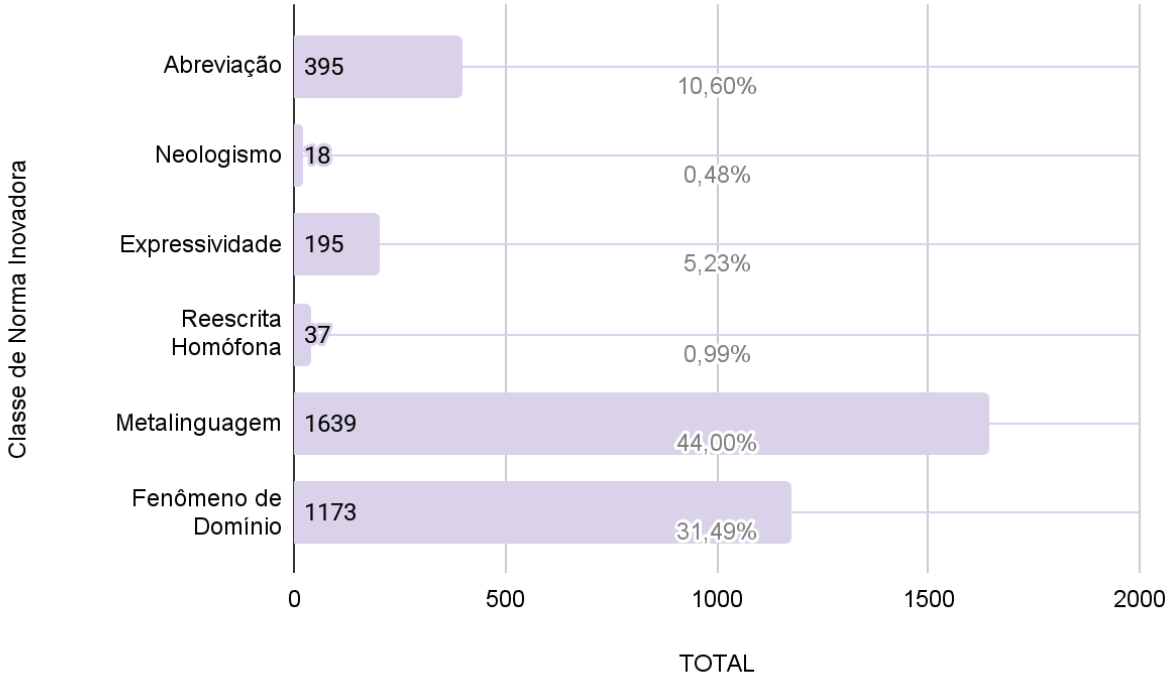
De modo semelhante, a classe Fenômeno de domínio, também detalhada no Quadro 2, é amplamente dominada por tickers (1.090 ocorrências) e cashtags (83). A elevada frequência dessas formas evidencia a centralidade de unidades especializadas na dinâmica comunicativa do corpus. Tais ocorrências não apenas refletem o recorte temático adotado na constituição dos dados, mas indicam a consolidação de convenções gráficas próprias da comunidade discursiva financeira, nas quais a condensação simbólica e a precisão referencial desempenham papel central.

As classes Abreviação (395 ocorrências) e Expressividade (195 ocorrências), igualmente registradas no Quadro 2, configuram espaços relevantes de intervenção sobre o léxico comum. No caso das abreviações, sobressaem encurtamentos (108), contrações (152) e inicialismos (70), apontando para uma tendência consistente de economia linguística. Já na classe Expressividade, a capitalização (122) e o prolongamento grafêmico (33) evidenciam estratégias de intensificação pragmática e de marcação avaliativa. Ainda que quantitativamente secundárias em relação às duas classes principais, essas categorias demonstram que a inovação também incide sobre

a materialidade gráfica como recurso expressivo.

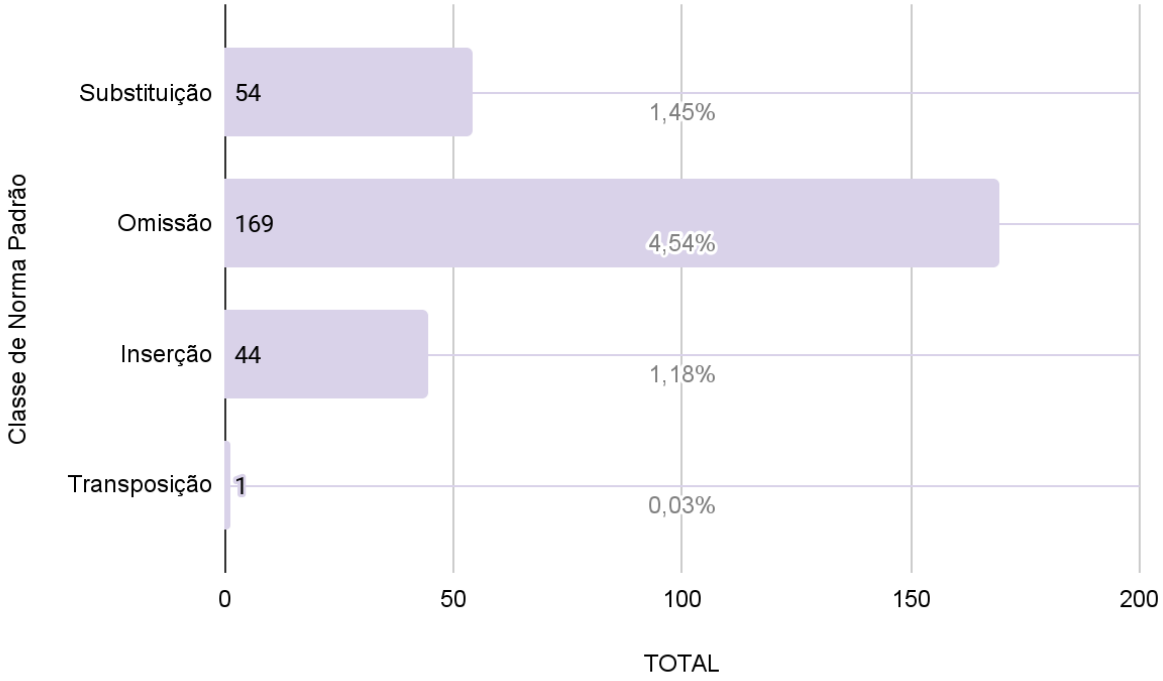
A leitura comparativa dos Quadros 1 e 2 torna evidente a assimetria entre as duas normas. Enquanto a Norma Inovadora apresenta forte concentração em categorias funcionalmente estruturadas que são relacionadas à indexação digital, à especialização temática e à manipulação estratégica da forma gráfica no discurso digital, a Norma Padrão caracteriza-se por alterações predominantemente microestruturais, dispersas e restritas à camada ortográfica.

Figura 10. Distribuição percentual das classes de Norma Inovadora em relação ao total de fenômenos anotados (n = 3.725).



Fonte: A autora, 2026.

Figura 11. Distribuição percentual das classes de Norma Padrão em relação ao total de fenômenos anotados (n = 3.725).



Fonte: A autora, 2026.

Assim, a análise articulada dos dados sintetizados nos Quadros 1 e 2 em conjunto com a distribuição percentual nas Figuras 10 e 11 reforça a hipótese de que a variação gráfica no corpus DANTEStocks não pode ser compreendida prioritariamente como resultado de erro aleatório uma vez que as estatísticas de Norma Inovadora revelam organização concentrada e funcionalmente orientada, enquanto na Norma Padrão manifestam-se sobretudo alterações pontuais que não apresentam consistente função discursiva sistemática. A diferença entre ambas não é apenas quantitativa, mas qualitativa, refletindo modos distintos de relação com o código escrito no contexto da comunicação financeira digital e, portanto, a necessidade de classificação do que normalmente é descartado no tratamento de dados de PLN pois é considerado “ruído” ou “erro”.

5.2.3. Os casos de sobreposição

A anotação do recorte do DANTEStocks revelou a existência de tokens nos quais mais de uma norma incide simultaneamente. Em alguns casos, observam-se até três

fenômenos coexistindo em um único item lexical. A sistematização desses dados encontra-se nos Quadros 3, 4, 5 e 6, que permitem examinar de modo mais preciso como diferentes tipos se combinam e quais padrões emergem dessas associações.

O Quadro 3 apresenta 50 ocorrências em que duas normas inovadoras se sobrepõem no mesmo token. A combinação mais frequente envolve Contração (Abreviação) e Prolongamento Grafêmico (Expressividade), com 22 casos, como em *rsrsr*. Trata-se de forma particularmente relevante, pois articula economia linguística e intensificação expressiva em um único movimento gráfico. Também se observam combinações entre Abreviação e Capitalização, como em *VC*, bem como entre Reescrita Homófona e Expressividade, como em sequências do tipo *kkkkkkk*. Esses dados indicam que a inovação raramente opera de modo isolado, pois os recursos inovadores empregados tendem a acumular funções, condensando economia, marcação afetiva e posicionamento discursivo numa única palavra, por exemplo.

Quadro 3. Fenômenos com duas normas inovadoras.

Tipo 1	Tipo 2	Qt.	Exemplo	Lema
Contração (Abreviação)	Prolong. Grafêmico (Expressividade)	22	rsrsr	rsrsr
Contração (Abreviação)	Substituição Grafêmica (Reescrita Homófona)	2	d+	demais
Contração (Abreviação)	Capitalização (Expressividade)	1	VC	você
Encurtamento (Abreviação)	Substituição Grafêmica (Reescrita Homófona)	5	ñ	não
Fonetização (Reescrita Homófona)	Prolong. Grafêmico (Expressividade)	4	kkkkkkk	kkkkkk
Fonetização (Reescrita Homófona)	Variação dialetal (Expressividade)	3	qua	qua
URL (Metalinguagem)	Truncamento (Abreviação)	4	http://t.…	http://t.…
Inicialismo (Abreviação)	Capitalização (Expressividade)	1	PQP	PQP
Aglutinação (Neologismo)	Capitalização (Expressividade)	1	PETROFUMO	petrofumo
Derivação (Neologismo)	Capitalização (Expressividade)	1	CAPETALIZAÇÃO	capitalização
Variação dialetal (Expressividade)	Capitalização (Expressividade)	1	FUDER	foder
Influência estrangeira (Neologismo)	Contração (Abreviação)	1	plz	plz
Variação dialectal (Expressividade)	Encurtamento (Abreviação)	1	prejú	prejuízo
Encurtamento (Abreviação)	Capitalização (Expressividade)	1	OBJ	objetivo
Prolong. Grafêmico (Expressividade)	Capitalização (Expressividade)	1	FUUUUUUU	FUUUUUUU
Inicialismo (Abreviação)	Capitalização (Expressividade)	1	PQP	PQP
TOTAL		**50**		

Fonte: A autora (2026).

Quadro 4. Fenômenos com uma norma inovadora e uma norma padrão.

Tipo 1	Tipo 2	Qt.	Exemplo	Lema
Hashtag (Metalinguagem)	Diacrítico (Inserção)	8	#Petrobrás	#Petrobrás
Capitalização (Expressividade)	Diacrítico (Omissão)	1	MINIMO	mínimo

Fonte: A autora (2026).

Quadro 5. Fenômenos com duas normas padrão.

Tipo 1	Subtipo 1	Tipo 2	Qt	Subtipo 2	Exemplo	Lema
Diacrítico (Substituição)	Cedilha	Diacrítico (Omissão)	40	N/A	Precificacao	precificação
Diacrítico (Inserção)	N/A	Outro (Inserção)	3	N/A	d'?ela	dela
Outro (Omissão)	N/A	Diacrítico (Omissão)	1	N/A	coincidenci	coincidência
TOTAL			**44**			

Fonte: A autora (2026).

Quadro 6. Fenômenos com três fenômenos no mesmo token.

Tipo 1	Tipo 2	Tipo 3	Qt.	Exemplo	Lema
Hashtag (Metalinguagem)	Aglutinação (Neologismo)	Capitalização (Expressividade)	3	#IBOLESMA	#IBOLESMA
Fonetização (Reescrita Homófona)	Prolong. Grafêmico (Expressividade)	Capitalização (Expressividade)	1	KKKKKKKKKKK	KKKKKKKKKK
TOTAL			**4**		

Fonte: A autora (2026).

Ainda no âmbito das normas inovadoras, o Quadro 6 registra quatro casos em que três fenômenos coexistem no mesmo token. Exemplos como *#IBOLESMA*, que reúne Hashtag (Metalinguagem), Aglutinação (Neologismo) e Capitalização (Expressividade), evidenciam um grau mais complexo de articulação. Nesse tipo de

ocorrência, observa-se a convergência entre recurso estrutural da plataforma, criação lexical e intensificação gráfica, configurando um uso multifuncional que integra indexação temática, avaliação implícita e busca por visibilidade discursiva.

O Quadro 4, por sua vez, apresenta nove casos de sobreposição entre uma norma inovadora e uma norma padrão, como em *#Petrobrás*, que combina Hashtag (Metalinguagem) e Inserção de Diacrítico (Norma Padrão). Esses casos são particularmente relevantes porque confrontam uma oposição rígida entre inovação e adequação normativa. A presença simultânea de um recurso típico da escrita digital e de marca ortográfica convencional sugere que a adesão à Norma Padrão não exclui a incorporação de estratégias próprias do ambiente digital. Antes, evidencia-se um uso articulado de diferentes repertórios gráficos conforme as demandas do contexto.

Já o Quadro 5 registra 44 ocorrências de sobreposição entre duas normas padrão, predominando combinações que envolvem alterações de diacríticos, como em *Precificacao*, em que há simultaneamente Substituição e Omissão. Diferentemente dos casos inovadores, aqui as sobreposições permanecem restritas à dimensão micrográfica e não configuram articulações funcionais amplas. Trata-se de acúmulo de alterações ortográficas localizadas, sem evidência de orientação pragmática sistemática.

No total, o corpus apresenta 50 casos de dupla norma inovadora, 9 casos mistos (inovadora + padrão), 44 casos de dupla norma padrão e 4 ocorrências com três fenômenos simultâneos. A distribuição dessas combinações revela uma assimetria significativa: as sobreposições que envolvem a Norma Inovadora tendem a ser funcionalmente integradas, ao passo que as sobreposições restritas à Norma Padrão se concentram em aspectos gráficos pontuais.

A coexistência de abreviação, expressividade, indexação digital e criação lexical em um mesmo token sugere que os usuários mobilizam diferentes camadas do sistema gráfico de forma articulada. A escolha linguística, nesse sentido, não opera apenas no nível de seleção entre variantes, mas também na possibilidade de sobreposição e combinação de mecanismos distintos.Tal configuração dialoga com a perspectiva laboviana de que a variação é sistemática e condicionada por fatores sociais e situacionais.

No corpus analisado, a sobreposição de normas pode ser compreendida como resposta às condições específicas de circulação textual em ambiente digital,

caracterizado por rapidez, visibilidade e competição por atenção. Mesmo nos casos classificados como Norma Padrão, a coexistência de múltiplas alterações sugere que a escrita digital não se organiza segundo uma lógica binária entre erro e acerto, mas segundo gradientes de adequação e funcionalidade.

Assim, os casos de sobreposição evidenciam que a inovação linguística no DANTEStocks não se restringe à introdução de formas isoladas, mas inclui processos de soma estratégica de recursos gráficos. Trata-se de um uso multifuncional da escrita, no qual diferentes padrões coexistem e se articulam para atender simultaneamente a objetivos informacionais, expressivos e indexicais.

5.2.4. Distribuição dos fenômenos por PoS

A análise da distribuição dos fenômenos léxico-ortográficos em função das categorias gramaticais (PoS tags) do modelo *Universal Dependencies* constitui uma das contribuições mais relevantes deste trabalho, por permitir examinar em que medida determinadas classes gramaticais mais aparecem na variação ortográfica no corpus DANTEStocks. O cruzamento entre os 3.725 fenômenos anotados e as 17 etiquetas PoS — acrescidas da categoria SEM UPOS, que corresponde a 20 ocorrências associadas a tokens multi-palavras e linhas sem identificador numérico — resulta em um quadro distribucional assimétrico, no qual três categorias (PROPN, X e SYM) concentram a quase totalidade das ocorrências de Norma Inovadora, enquanto as demais etiquetas apresentam perfis distintos e funcionalmente orientados.

Quadro 7. Distribuição de calor das PoSTags por Classe

UPOS	Sub	Om	In	Tr	Abr	Neo	Exp	Hom	Meta	Dom	TOTAL	%
ADJ	0	15	0	0	13	1	9	0	0	0	38	1,02%
ADP	0	4	0	0	42	0	7	0	0	0	53	1,42%
ADV	0	21	0	0	36	1	2	7	0	0	67	1,80%
AUX	0	22	0	0	1	0	0	1	0	0	24	0,64%
CCONJ	0	0	0	0	0	0	5	0	0	1	6	0,16%
DET	0	2	0	0	7	0	12	0	0	0	21	0,56%
INTJ	0	1	0	0	1	0	4	6	1	0	13	0,35%
NOUN	50	98	6	0	106	5	45	3	14	2	329	8,83%
NUM	0	0	0	0	0	0	6	0	0	0	6	0,16%
PART	0	0	0	0	0	0	0	0	0	0	0	0,00%
PRON	0	3	2	0	33	0	4	0	0	0	42	1,13%
PROPN	2	0	23	0	63	9	19	2	807	1087	2012	54,01%
PUNCT	0	0	2	0	3	0	1	0	0	1	7	0,19%
SCONJ	0	0	0	0	23	0	0	0	0	0	23	0,62%
SYM	0	0	0	0	10	0	19	0	449	1	479	12,86%
VERB	2	3	4	1	8	0	20	0	0	0	38	1,02%
X	0	0	0	0	36	2	42	18	368	81	547	14,68%
SEM UPOS	0	0	7	0	13	0	0	0	0	0	20	0,54%
											3725	100,00%

Legenda: Sub=Substituição; Om=Omissão; In=Inserção; Tr=Transposição; Abr=Abreviação; Neo=Neologismo; Exp=Expressividade; Hom=Reescrita Homófona; Meta=Metalinguagem; Dom=Fenômeno de Domínio.

Fonte: A autora (2026).

Do total de 3.725 fenômenos, 3.036 ocorrências (81,5%) estão associadas a apenas três etiquetas: PROPN (2.012; 54,0%), X (547; 14,7%) e SYM (477; 12,8%). Essa concentração não é fortuita, mas decorre diretamente da natureza linguística dessas categorias no contexto de tweets do mercado financeiro.

A etiqueta PROPN (nome próprio) apresenta o maior volume absoluto de fenômenos do corpus, com 2.012 ocorrências. Esse índice elevado reflete o fato de que os nomes próprios no corpus são, em sua maioria, elementos da Norma Inovadora: 807 ocorrências de Metalinguagem (principalmente hashtags com nomes de empresas, como #VALE5 e #PETR4) e 1.087 de Fenômeno de Domínio (tickers e cashtags), que, juntas, somam 94,3% dos fenômenos associados à etiqueta. As 63 ocorrências de Abreviação em PROPN indicam formas contraídas ou encurtadas de denominações empresariais, enquanto as 23 de Inserção correspondem majoritariamente a casos como #Petrobrás, em que há inserção de diacrítico em token de Metalinguagem — casos de sobreposição de normas, conforme discutido na Seção 5.2.3.

A etiqueta X, que no modelo UD acolhe tokens não classificáveis segundo as demais categorias — sendo, portanto, a etiqueta residual por excelência —, apresenta 547 fenômenos anotados, correspondendo a 91,6% do total de tokens X no subcorpus. Esse percentual é o mais elevado de todo o corpus, indicando que praticamente a totalidade dos tokens etiquetados como X manifesta algum fenômeno léxico-ortográfico. A distribuição interna revela concentração em Metalinguagem (368; 67,3%) e Fenômeno de Domínio (81; 14,8%), com participação relevante de Expressividade (42; 7,7%) e Abreviação (36; 6,6%). Esse perfil confirma que a etiqueta X, no DANTEStocks, funciona como repositório de formas cujas características morfológicas não se encaixam nas categorias canônicas justamente porque são, em sua grande maioria, tokens de natureza metalinguística ou especializada — hashtags sem nome próprio explícito, emoticons, siglas sem lema identificável e cashtags —, todos pertencentes à Norma Inovadora.

A etiqueta SYM (símbolo) concentra 477 fenômenos, com taxa de anotação de 44,4% de seus tokens. A predominância de Metalinguagem (447; 93,7%) revela que os tokens etiquetados como SYM correspondem, em sua maioria, a URLs, que, embora sejam formas graficamente peculiares, são sistematicamente anotadas nessa categoria. Os 19 casos de Expressividade em SYM incluem emoticons e símbolos

matemáticos empregados como recurso avaliativo. A concentração de URLs em SYM e Metalinguagem em X e PROPN configura, na prática, uma divisão funcional entre as três etiquetas dominantes no que se refere aos recursos discursivos da plataforma.

Excluídas as três etiquetas dominantes, os 689 fenômenos restantes distribuem-se por 14 etiquetas, com perfis qualitativamente distintos. A etiqueta NOUN (substantivo) apresenta o maior volume nesse grupo, com 329 ocorrências, sendo a única etiqueta entre as lexicais plenas a reunir fenômenos das duas normas em proporções significativas: 154 ocorrências de Norma Padrão (predominantemente Omissão de diacrítico, com 98 casos, e Substituição, com 50) e 175 de Norma Inovadora (com destaque para Abreviação, com 106, e Expressividade, com 45). Esse duplo perfil indica que os substantivos são suscetíveis tanto a desvios ortográficos involuntários — típicos da Norma Padrão — quanto a estratégias de economia linguística e intensificação pragmática, características da Norma Inovadora.

Entre as etiquetas funcionais, ADP (adposição) e SCONJ (conjunção subordinativa) apresentam perfis reveladores. ADP reúne 53 fenômenos, dos quais 42 (79,2%) são Abreviação — em especial contrações de preposições e artigos como "pra" (< "para a") e "no" (< "em o") — e 4 são Omissão de diacrítico. SCONJ apresenta 23 fenômenos, todos classificados como Abreviação, correspondendo a formas como "oq" (< "o que") e "pq" (< "porque"), que no corpus ocorrem tanto como tokens de ID numérico quanto como multi-word tokens de UPOS "_" — precisamente os casos capturados pela linha SEM UPOS. A concentração de Abreviação nessas etiquetas funcionais atesta que a redução de palavras gramaticais de alta frequência é uma estratégia sistemática de economia linguística no gênero tweet.

As etiquetas ADV (advérbio), AUX (auxiliar) e ADJ (adjetivo) apresentam fenômenos de Norma Padrão em proporção elevada relativa ao seu volume total: ADV reúne 21 Omissões e nenhuma Substituição ou Inserção, enquanto AUX concentra 22 Omissões — quase exclusivamente casos de omissão de diacrítico em formas verbais como "esta" (<"está") e "e" (<"é"), que no corpus exercem função auxiliar. Em ADJ, os 15 fenômenos de Norma Padrão são todos de Omissão, enquanto os 23 de Norma Inovadora distribuem-se entre Abreviação (13) e Expressividade (9). A etiqueta VERB (verbo) apresenta o único caso de Transposição do corpus — a única ocorrência desta classe em todo o subcorpus anotado —, além de 10 fenômenos de Norma Padrão e 28 de Norma Inovadora, sendo Expressividade (20) e Abreviação (8)

os mais frequentes.

A etiqueta INTJ (interjeição) concentra 13 fenômenos, com destaque para 6 casos de Reescrita Homófona — a maior proporção relativa desta classe entre todas as etiquetas, correspondendo a 16,2% do total de 37 ocorrências dessa classe no corpus. Esse resultado é coerente com a natureza das interjeições, que tendem a representar, na escrita digital, formas próximas da oralidade, como fonetizações de expressões afetivas ("hehehe", "kkkk"). A etiqueta PRON (pronome) reúne 42 fenômenos, dos quais 33 (78,6%) são Abreviação — formas como "vc" (< "você") e "oq" (< "o que") —, corroborando a tendência de contração de itens pronominais de alta frequência.

A categoria SEM UPOS, criada para capturar as 22 ocorrências associadas a tokens sem etiqueta PoS válida — multi-word tokens com UPOS "_" e linhas sem identificador numérico —, concentra 13 Abreviações (formas como "pra", "pq", "oq", "pro") e 7 Inserções (formas como "d'?ela", "d'?ele", "d'?eles"). Esses tokens correspondem, no formato CoNLL-U, a contrações que foram segmentadas em sub-tokens, sendo o multi-word token o nível em que a anotação de fenômenos foi realizada. A inclusão desta categoria garante a completude do levantamento, elevando o total de fenômenos rastreáveis à cifra confirmada de 3.725.

Em síntese, a distribuição dos fenômenos por PoS revela uma estrutura de dois regimes: (i) um regime dominante, concentrado em PROPN, X e SYM, caracterizado pela quase exclusividade de fenômenos de Norma Inovadora ligados à arquitetura comunicativa da plataforma (Metalinguagem) e à especialização do domínio financeiro (Fenômeno de Domínio); e (ii) um regime distribuído, que abrange as demais etiquetas e exibe fenômenos de ambas as normas, com padrões funcionalmente coerentes — Omissão de diacrítico em classes de conteúdo (NOUN, ADV, AUX), Abreviação em itens gramaticais de alta frequência (ADP, SCONJ, PRON) e Expressividade em categorias expressivas (VERB, INTJ, NOUN). Essa bipartição evidencia que a variação léxico-ortográfica no DANTEStocks não é homogênea entre as classes gramaticais, mas obedece a lógicas distintas conforme a função discursiva e o grau de especialização dos tokens envolvidos.

6. Considerações finais

Este trabalho partiu de uma inquietação teórica e metodológica central: como caracterizar, de modo rigoroso e linguisticamente fundamentado, as variações ortográficas presentes em tweets do mercado financeiro? A resposta construída ao longo da pesquisa foi a proposição e a aplicação de uma tipologia hierárquica de fenômenos léxico-ortográficos ao corpus DANTEStocks, organizada em torno de duas dimensões — Norma Padrão e Norma Inovadora —, fundamentadas nos pressupostos da Sociolinguística Variacionista (Labov, 1972; Bagno, 2007) e nas operações de edição de Damerau (1964), adaptadas ao padrão Unicode.

Do ponto de vista dos resultados quantitativos, a anotação manual de 1.069 tweets — equivalente a aproximadamente 26% do DANTEStocks — gerou um total de 3.725 fenômenos, distribuídos em 3.614 tokens (20.914 no total do subcorpus), com 107 tokens apresentando mais de um fenômeno simultâneo: 103 com dois e 4 com três. Esses números confirmam que, embora a sobreposição de fenômenos seja minoritária, ela é relevante do ponto de vista linguístico, evidenciando a natureza multifuncional da escrita digital. A predominância da Norma Inovadora — 3.458 ocorrências, correspondendo a 92,83% do total — é robusta o suficiente para não ser atribuída a viés de anotação, mesmo considerando que a anotação foi realizada por um único anotador sem cálculo formal de concordância interanotadores.

Essa predominância expressiva da Norma Inovadora confirma a hipótese de partida de que, no Conteúdo Gerado por Usuário de temática financeira, as práticas ortográficas não canônicas não constituem desvios aleatórios ou erros de digitação, mas estratégias linguísticas sistemáticas e funcionalmente motivadas. A concentração de Metalinguagem (1.639; 44,0% da Norma Inovadora) e de Fenômeno de Domínio (1.173; 33,9%) indica que a maior parte da variação observada não incide sobre o léxico comum, mas sobre recursos estruturais da plataforma Twitter (hashtags, menções, retweets e URLs) e sobre o vocabulário especializado do mercado financeiro (tickers e cashtags). Trata-se, portanto, de uma variação funcionalmente orientada, que reflete convenções consolidadas na comunidade discursiva financeira digital.

A análise por categoria gramatical (PoS), apresentada na Seção 5.2.4, aprofunda essa compreensão ao revelar que a distribuição dos fenômenos não é uniforme entre as etiquetas, mas estruturada em dois regimes distintos. O primeiro,

dominante, concentra-se em PROPN (2.012 fenômenos; 54,0%), X (547; 14,7%) e SYM (477; 12,8%), etiquetas nas quais a Norma Inovadora responde por percentuais próximos ou superiores a 90% das ocorrências e cuja variação é determinada, em larga medida, pela natureza dos tokens — nomes próprios de empresas, elementos metalinguísticos e simbólicos. O segundo regime, distribuído, abrange as demais etiquetas e exibe fenômenos de ambas as normas com padrões funcionalmente coerentes: Omissão de diacrítico predomina em classes de conteúdo (NOUN, ADV, AUX), Abreviação concentra-se em itens gramaticais de alta frequência (ADP, SCONJ, PRON) e Expressividade comparece em categorias expressivas (VERB, INTJ, NOUN).

Do ponto de vista tipológico, a tipologia proposta demonstrou operacionalidade empírica satisfatória: as categorias propostas foram suficientes para cobrir a quase totalidade dos fenômenos observados, e a distinção entre Norma Padrão e Norma Inovadora mostrou-se analiticamente produtiva, ao capturar não apenas diferenças de frequência, mas diferenças qualitativas na organização e na função discursiva dos fenômenos. A Norma Padrão caracteriza-se por alterações microestruturais dispersas — sobretudo Omissão de diacrítico (158 casos) —, sem orientação pragmática sistemática, enquanto a Norma Inovadora apresenta concentração funcional em torno da indexação digital e da especialização terminológica.

A tipologia aqui desenvolvida distingue-se da taxonomia de Sanguinetti et al. (2023), tomada como referência crítica, em dois aspectos fundamentais. Primeiro, ao substituir o critério de "intencionalidade" — de operacionalização problemática em dados de superfície textual, como reconhecem os próprios autores — pelas dimensões de "norma", ancoradas teoricamente na Sociolinguística Variacionista e empiricamente nas operações de Damerau, obteve-se um modelo cuja aplicação é mais objetiva e replicável. Segundo, ao contemplar fenômenos específicos da plataforma Twitter (Metalinguagem) e do domínio financeiro (Fenômeno de Domínio) — ausentes na proposta de Sanguinetti et al. por seu escopo genérico —, a tipologia proposta demonstra maior adequação descritiva para o corpus em questão.

Um resultado metodológico de relevo é a identificação de 22 ocorrências associadas a tokens sem UPOS válida — multi-word tokens (UPOS "_") e linhas sem identificador numérico —, que, se ignoradas, produziriam uma discrepância de 0,6% no total de fenômenos. A criação da categoria SEM UPOS na tabela de distribuição

por PoS garante a completude do levantamento e evidencia uma especificidade do formato CoNLL-U que deve ser considerada em pesquisas futuras que utilizem o mesmo protocolo de anotação: a anotação de fenômenos nos multi-word tokens precisa ser acautelada na fase de construção das fórmulas de contagem, sob pena de subestimação sistemática.

Entre as limitações do trabalho, destacam-se três. A primeira é a ausência de cálculo formal de concordância interanotadores, o que impede a quantificação objetiva do grau de subjetividade introduzido nas decisões de classificação — especialmente em casos limítrofes entre Norma Padrão e Norma Inovadora, ou entre classes como Abreviação e Reescrita Homófona. A segunda é a cobertura parcial do corpus: os 1.069 tweets anotados correspondem a 26,4% do DANTEStocks, de modo que os resultados devem ser interpretados como uma caracterização preliminar, sujeita a revisão com a ampliação da base anotada. A terceira é a não realização da etapa de validação da tipologia em corpus externo — especificamente, um corpus de tweets sobre COVID-19, conforme previsto no plano inicial —, o que limita a avaliação da generalidade da tipologia para outros domínios do CGU em português.

Do ponto de vista das contribuições, este trabalho oferece, em primeiro lugar, uma tipologia testada empiricamente para a descrição de fenômenos léxico-ortográficos em tweets do mercado financeiro em português, passível de extensão a outros subcorpora do DANTEStocks e a outros gêneros de CGU. Em segundo lugar, a anotação de 3.725 fenômenos em 3.614 tokens constitui um recurso linguístico concreto que, integrado ao arquivo CoNLL-U do DANTEStocks, amplia as camadas de anotação disponíveis para o treinamento e avaliação de modelos de Processamento de Língua Natural voltados ao português não canônico. Em terceiro lugar, a análise estatística multidimensional — por norma, por classe, por sobreposição e por PoS — oferece uma caracterização linguística densa do subcorpus, com implicações tanto para a linguística descritiva do português digital quanto para a engenharia de sistemas de normalização e pré-processamento de CGU.

Como caminhos para trabalhos futuros, identificam-se quatro direções prioritárias: (i) a expansão da anotação para os 2.979 tweets restantes do DANTEStocks, com adoção de protocolo de múltiplos anotadores e cálculo de IAA (Inter-Annotator Agreement); (ii) a realização do estudo de validação da tipologia em

corpus de outro domínio, como política ou saúde, para aferir a robustez e a generalidade do modelo tipológico; (iii) a integração da anotação de fenômenos léxico-ortográficos a sistemas automáticos de normalização, com investigação do impacto da informação de classe e tipo sobre a qualidade da normalização produzida; e (iv) o desenvolvimento de um anotador automático de fenômenos léxico-ortográficos treinado sobre os dados produzidos neste trabalho, com vistas à anotação eficiente dos demais tweets do corpus e à sua eventual disponibilização como ferramenta de PLN para o português.

Por fim, este trabalho reafirma um pressuposto da Sociolinguística Variacionista que, embora consolidado teoricamente, ainda encontra resistência nas práticas de pré-processamento de PLN: a variação linguística é inerente ao uso da língua, e as formas não canônicas presentes em tweets não são ruído a ser eliminado, mas informação a ser descrita, compreendida e incorporada aos modelos computacionais. Reconhecer a inteligência linguística presente naquilo que se convencionou chamar de erro é, ao mesmo tempo, um imperativo teórico da Linguística e uma exigência prática de qualquer sistema de PLN que aspire à robustez e à sensibilidade contextual no processamento do português digital contemporâneo.

7. Referências bibliográficas

AITCHISON, Jean. **Language change: progress or decay?** 3. ed. Cambridge: Cambridge University Press, 2001.

BAGNO, M. **Preconceito linguístico**: o que é, como se faz. 37. ed. São Paulo: Loyola, 2007.

BARBERIA, L. H.; SCHMALZ, P. H. S.; ROMAN, N. T. When Tweets Get Viral – A Deep Learning Approach for Stance Analysis of Covid-19 Vaccines Tweets By Brazilian Political Elites. In: SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY (STIL), 14, 2023, Belo Horizonte. **Proceedings** […]. Porto Alegre: SBC, 2023. p. 104-114. DOI: https://doi.org/10.5753/stil.2023.233961.

BARBOSA, B. K. S. **Descrição sintático-semântica de nomes predicadores em tweets do mercado financeiro em português**. 2024. 163f. Dissertação (Mestrado em Linguística) – Universidade Federal de São Carlos, São Carlos, 2024.

CAGLIARI, L. C. *Ortografia e alfabetização*. São Paulo: Contexto, 1998.

CARDOSO, P.C.S. **A conversação pública no Twitter***:* uma análise enunciativo-discursiva. 2019. 265 f. Tese (Doutorado em Linguística) – Universidade Federal de Minas Gerais, Belo Horizonte, 2019.

CARLETTA, J. Assessing agreement on cllassification tasks: The kappa statistic. **Computational Linguistics**, v. 22, n. 2, pages 249–254. MIT Press.

COELHO, I. L.; MONGUILHOTT, I. O. S.; SEVERO, C. G. **Norma linguística do português no Brasil**: 12º período. Florianópolis: LLV/CCE/UFSC, 2014.

COELHO, I. L. **Sociolinguística**. Florianópolis: LLV/CCE/UFSC, 2010. ISBN 978-85-61482-25-1.

COHEN, J. A coefficient of agreement for nominal scales. **Educational and Psychological Measurement**, v. 20, n. 1, p. 37-46, 1960.

CRYSTAL, D. **Language and the Internet**. 2. ed. Cambridge: Cambridge University Press, 2006.

DAMERAU, F. J. A technique for computer detection and correction of spelling errors. **Communications of the ACM**, v. 7, n. 3, p. 171–176, 1964.

DERCZYNSKI, L.; BONTCHEVA, K.; ROBERTS, I. Broad Twitter Corpus: A Diverse Named Entity Recognition Resource. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, 26, 2016, Osaka. **Proceedings** […]. Osaka: ACL, 2016. p. 1169-1179.

DI-FELIPPO, A., NUNES, M. G. V., BARBOSA, B. K. S. A dependency treebank of tweets in Brazilian Portuguese: syntactic annotation issues and approach. In: SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY, 15, Belém/PA. **Proceedings**.... Porto Alegre/RS: SBC, 2024a, p. 192–201.

DI-FELIPPO, A., *et al*. Genipapo - a multigenre dependency parser for Brazilian Portuguese. In: SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY, 15, Belém/PA. **Proceedings**.... Porto Alegre/RS: SBC, 2024b, p. 257–266.

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*.

DI-FELIPPO, A. *et al*. Descrição preliminar do corpus DANTEStocks: diretrizes de segmentação para anotação segundo *Universal Dependencies*. In: JORNADA DE DESCRIÇÃO DO PORTUGUÊS, 7, 2021, [online]. **Anais**.... São Carlos: ICMC-USP, 2021. p. 335–343

.

DI-FELIPPO, A., NUNES, M. G. V., BARBOSA, B. K. S. Diretrizes de anotação de relações de dependência em *tweets* do mercado financeiro. **Relatório Técnico do ICMC 446**. ICMC-USP, abr. 2024. 70p.

DI-FELIPPO, A. *et al.* Diretrizes de anotação de PoS tags em *tweets* do mercado financeiro: orientações para anotação em língua portuguesa segundo a abordagem *Universal Dependencies*. **Relatório Técnico do ICMC 438**. ICMC-USP, 2022. 24p.

DURAN, M. *et al.* The dawn of the PortTinari multigenre treebank: introducing its journalistic portion. In: SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY 14, 2023, Belo Horizonte. **Proceedings**… Porto Alegre: SBC, 2023. p. 115–124.

DURAN, M. S.; PARDO, T. A. S. Anotação de córpus, um lugar privilegiado de observação linguística: o estudo das aposições do português brasileiro segundo o modelo Universal Dependencies. In: ENCONTRO DE LINGUÍSTICA DE CORPUS, 16., 2024, Brasília. **Anais**.... Brasília, 2024. p. 118–123.

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* (UD). **Relatório Técnico do ICMC 434**. ICMC-USP, 2021. 55p.

DURAN, M.S. Manual de Anotação de Relações de Dependência - Versão Revisada e Estendida: Orientações para anotação de relações de dependência sintática em Língua Portuguesa, seguindo as diretrizes da abordagem *Universal Dependencies* (UD). **Relatório Técnico do ICMC 440**. ICMC-USP, 2022. 166p.

EISENSTEIN, J. What to do about bad language on the internet. In: CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS – HUMAN LANGUAGE TECHNOLOGIES, 2013, Atlanta. **Proceedings**.... Atlanta: ACL, 2013. p. 359–369.

FOSTER, J. Automatic Error Correction for Text. 2010. Tese (Doutorado em Computação) – Dublin City University, Dublin, 2010.

FREITAS, T.; BARTH, F. Gênero textual e redes sociais: o caso do Twitter. **Revista Brasileira de Linguística Aplicada**, v. 15, n. 3, 2015.

JURAFSKY, D.; MARTIN, J. H. *Speech and Language Processing*: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 3. ed. Disponível em: https://web.stanford.edu/~jurafsky/slp3/. Acesso em: 1 mai. 2025.

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

LABOV, W. **Sociolinguistic Patterns.** Philadelphia: University of Pennsylvania Press, 1972.

LABOV, William. ***Padrões sociolinguísticos.*** Tradução de Marcos Bagno, Marta Scherre e Caroline Cardoso. São Paulo: Parábola Editorial, 2008.

LOPES, L. *et al*. PortiLexicon-UD: a Portuguese lexical resource according to Universal Dependencies model. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 13, 2022, Marseille, França. **Proceedings**.... Marseille: ELRA, 2022. p. 6635–6643.

MARCUSCHI, L. A. **Gêneros textuais:** definição e funcionalidade. São Paulo: Cortez, 2008.

MORAN, S.; CYSOUW, M. The Unicode Cookbook for Linguists: Managing writing systems using orthography profiles. Berlin: Language Science Press, University of Zurich, 2017. Disponível em: https://langsci-press.org/catalog/book/176. Acesso em: 8 jul. 2025.

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

NIVRE, J. *et al*. UNIVERSAL DEPENDENCIES v2: an evergrowing multilingual treebank collection. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 12, 2020, Marseille. **Proceedings**.... Marseille: ELRA, 2020. p. 4034–4043.

PLUTCHIK, R.; KELLERMAN, H. (Eds.). **Emotion:** Theory, Research and Experience. Nova Iorque: Academic Press, 1986.

QI, P. *et al.* Stanza: A Python natural language processing toolkit for many human languages. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (System Demonstrations), 58, 2020 [*online*]. **Proceedings**.... ACL, 2020. p. 101-108.

Rademaker, A. *et al.* Universal Dependencies for Portuguese. In: INTERNATIONAL CONFERENCE ON DEPENDENCY LINGUISTICS (Depling), 4., 2017, Pisa, Itália. **Proceedings**.... Pisa: Linköping University Electronic Press, 2017. p. 197–206.

REY, Alain. ***Essays on terminology.*** Amsterdam: John Benjamins, 1995.

SANGUINETTI, Manuela et al. Treebanking user-generated content: A proposal for a unified representation in Universal Dependencies. In: **Proceedings of the 12th language resources and evaluation conference**. ELRA, Language Resources Association, 2020. p. 5240-5250. 2020, Marseille, França.

SANGUINETTI, M. *et al*. Treebanking user-generated content: a UD based overview of guidelines, corpora and unified recommendations. **Lang Resources & Evaluation**, v. 57, n. 2, p. 493–544, 2023. Springer-VerlagBerlin, Heidelberg. ISSN:1574-020X

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, 14, 2023, Belo Horizonte. **Anais**.... Porto Alegre: SBC, 2023. p. 240-248.

SILVA, E. H.; *et al.*. Universal Dependencies for Tweets in Brazilian Portuguese: Tokenization and Part of Speech Tagging. *In*: BRAZILIAN CONFERENCE ON ARTIFICIAL AND COMPUTATIONAL INTELLIGENCE, 18, 2021 [online]. **Proceedings**.... Porto Alegre: SBC, 2021. p. 434-445.

SILVA, F. J. V.; ROMAN, N. T.; CARVALHO, A. M. B. R. Stock market tweets annotated with emotions. *Corpora*, v. 15, n. 3, p. 343–354, 2020. ISSN 1755-1676.

SOBREVILLA CABEZUDO, M. A.; PARDO, T. Towards a General Abstract Meaning Representation Corpus for Brazilian Portuguese. In: LINGUISTIC ANNOTATION WORKSHOP, 13, 2019, Florence. **Proceedings**.... Florence: ACL, 2019. p. 236–244.

STRAKA, M. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In: CoNLL 2018 SHARED TASK: MULTILINGUAL PARSING FROM RAW TEXT TO UNIVERSAL DEPENDENCIES, 2018, Brussels. **Proceedings**.... Brussels: ACL, 2018. p. 197–207.

WAGNER, R. A.; FISCHER, M. J. The String-to-String Correction Problem. **Journal of the ACM**, v. 21, n. 1, p. 168–173, 1974.

ZAPPAVIGNA, M. **Discourse of Twitter and social media**: How we use language to create affiliation on the web. London: Continuum, 2012.