Lexical and Orthographic Variation in Portuguese Financial Tweets: Annotation, Analysis, and Implications for Embedding Models

Ariani Di Felippo[1,2], Norton Trevisan Roman[1,3], Bryan K. S. Barbosa[1,2], Gabriela Pinheiro de Oliveira[1,2], Clarissa Lenina Scandarolli[1,2]

[1]Núcleo Interinstitucional de Linguística Computacional (NILC)
[2]Universidade Federal de São Carlos, [3]EACH/Universidade de São Paulo

Correspondence: ariani@ufscar.br, norton@usp.br, bryankhelven@ieee.org, gabrielapinheiro@estudante.ufscar.br, clarissa.scandarolli@estudante.ufscar.br

Abstract

Twitter/X remains a key source of user-generated content, requiring Natural Language Processing tools capable of handling non-canonical language. This study presents a manual annotation of lexical and orthographic phenomena in DANTEStocks, a corpus of financial tweets in Brazilian Portuguese, using a hierarchical typology to capture both creative uses and deviations from the standard norm. Results show that orthographic variation is strongly influenced by creative forms, mainly driven by platform- and domain-specific innovations. Standard norm variation is systematic, mostly involving predictable omissions of diacritics and the cedilla, and most tokens exhibit only one phenomenon, reflecting stable and largely independent patterns of variation in this Twitter subgenre. The identified variant forms enabled the construction of a lexicon for evaluating embedding models. We assessed how BERTimbau, Word2Vec, and FastText handle lexical variation in raw, unnormalized data, showing that the lexicon reduces out-of-vocabulary rates and improves coverage. These results highlight model robustness and the value of curated lexical resources in complementing both fixed and data-driven vocabularies.

1 Introduction

The immense popularity of social networks in the last decade has made Twitter/X an attractive source of data for numerous research fields and applications. It is particularly valuable for applications such as sentiment analysis and opinion mining in the stock market domain, where correlations between sentiment signals expressed in tweets and stock indices have even been used to predict market movements (Zhang et al., 2011; Mao et al., 2011; Bollen et al., 2011; Gaskell et al., 2013; Deveikyte et al., 2022).

With the aim of developing such applications for tweets in their original form (*i.e.*, without any normalization), Natural Language Processing (NLP) tools must learn to handle the predominantly non-standard language of this particular user-generated content (UGC) genre (*i.e.*, its short, noisy, and colloquial nature), requiring annotated corpora for training and evaluating them.

For English, researchers have access to several annotated corpora of financial tweets, such as the SemEval-2017 Task 5 dataset (Cortis et al., 2017), Fin-SoMe (Chen et al., 2020), and TweetFinSent (Pei et al., 2022). These datasets typically consist of posts (from Twitter and/or StockTwits) labeled for sentiment or market outlook, with annotations reflecting polarity (*e.g.*, positive/negative/neutral) or financial stance (*e.g.* bullish/bearish). Annotations are provided either by expert annotators or by authors of the posts, and dataset sizes generally range from ~2,000 to 10,000 labeled instances.

A pioneering multi-layered resource of gold-standard annotations of UGC written in (Brazilian) Portuguese is DANTEStocks (v.2.1.1)[1], comprising 4,042 tweets from the stock market domain annotated with morphosyntactic and syntactic information, along with named entities and emotions (Di-Felippo and Roman, 2025).

In this article, we present a pioneering corpus annotation, focusing on lexical and orthographic phenomena in DANTEStocks. This annotation employs a hierarchical taxonomy proposed by Scandarolli et al. (2023), designed to capture both creative uses and deviations from the standard norm. The typology unifies phenomena similar to those described by Sanguinetti et al. (2020a,b, 2022), while also incorporating the four types of operations proposed by Damerau (1964) and later refined for Portuguese by Gimenes et al. (2015).

This study offers two main contributions. First, it presents a manual annotation and analysis of lexical and orthographic variation in Portuguese

[1]https://github.com/UniversalDependencies/UD_Portuguese-DANTEStocks

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 628–637
April 13-16, 2026

financial tweets from DANTEStocks, providing linguistic insights into genre-specific patterns and creativity in this Twitter subgenre. Second, it derives a domain-specific variant lexicon from this annotation and shows that it improves vocabulary coverage when evaluating embedding models on raw, unnormalized data.

The remainder of this article is organized as follows: Section 2 provides a review of related work. In Section 3 we present our corpus and the annotation method used to build it, whereas Section 4 presents our results. A discussion on the coverage of these results in terms of reduction in the amount of out-of-vocabulary words left behind by three different automatic models is carried out in Section 5, with our final remarks being presented in Section 6.

2 Related work

According to Damerau (1964), over 80% of all words with orthographic deviations present a single phenomenon which, in turn, falls into one out of four categories: *omission* (a missing letter), *insertion* (an extra letter), *substitution* (an incorrect letter in place of the correct one), and *transposition* (the swapping of two adjacent letters). The fact that these categories were devised primarily for English was nevertheless pointed out, among others, by Gimenes et al. (2015), who took into account variations in Brazilian Portuguese.

In their research, Gimenes et al. (2015) analysed a corpus of blog posts (*i.e.*, UGC) and a corpus of dialogue summaries collected through a web experiment (*cf.* (Roman et al., 2013)), expanding the original categories to accommodate diacritic, cedilla, and space-related variations. Their results not only confirmed Damerau's findings when these variations are not accounted for, but also highlighted the key role of diacritics in Portuguese nonstandard spellings which, when dealt with, render Damerau's categorisation unsuited for this language.

Still in the realm of UGC, Sanguinetti et al. (2022, 2020a,b) proposed a typology of linguistic devices typical of social media content. This typology sought to systematise recurrent non-canonical phenomena across different UGC sources, genres (particularly Twitter), domains and languages, with the goal of supporting the development of specific annotation guidelines for UGC within the Universal Dependencies[2] (UD) framework, thereby ensuring cross-linguistic consistency in morphosyntactic and syntactic annotation.

In this typology, the noncanonical phenomena (*i.e.*, linguistic devices not commonly found in standard text) are classified as either intentional or unintentional. Unintentional phenomena are grouped into two main types: (i) *typos*, and (ii) *medium-dependent phenomena* (*i.e.*, alterations resulting from the electronic medium) such as autocorrection[3] (*e.g.*, changing English "concise" to Irish "coicíse") and truncation[4].

The intentional phenomena are divided into five types (and subtypes): (i) *encoding simplification* encompasses phenomena aimed at reducing the writing effort, such as diacritic omission (*e.g.*, GA *Leigh aris!→Léigh arís!* ("Read again!"), vowel omission (*e.g.*, EN *ppl→people*), phonetization (*e.g.*, EN *4ever→forever*), abbreviation (*e.g.*, EN *govt→government*), and spelling variation (*e.g.*, FR *je sé→je sais* ("I know")); (ii) *marks of expressiveness* are orthographic variations used for expressive purposes, such as punctuation reduplication (*e.g.*, EN *nice!!!*), (iii) graphemic stretching (*e.g.*, EN *superrrrr*), emoticons and smileys; (iv) *foreign influence* are code-switching (*e.g.*, IT *non fare la bad girl* ("don´t be a bad girl")) and transliteration (*e.g.*, GA *áicbheaird→amscaí*, whose pronunciation mimics the English word "awkward"), and (v) *lexical innovation*, including anthimeria (*e.g.*, IT *tuittare→twittare* ("to tweet")), disguise (*e.g.*, IT *caxxo→cazzo*) and transliteration.

Building on prior work, our study introduces a typology tailored to DANTEStocks designed to capture lexical and orthographic phenomena in financial discourse. The typology (i) prioritises a concrete, phenomenon-based description over interpretive or intentional criteria, thereby enhancing objectivity and reproducibility in annotation; (ii) extends further Gimenes et al. (2015)'s categories; and (iii) combines Sanguinetti et al. (2022)'s general framework with domain-specific patterns.

3 Materials and methods

This study builds on the DANTEStocks corpus (v2.1.1), which comprises 4,042 tweets (80,998 tokens) mentioning any stock listed in Ibovespa[5],

[2]UD (Nivre et al., 2020; de Marneffe et al., 2021) is a framework for morphosyntactic and syntactic annotation used to describe the grammatical structure of sentences across languages in a standardized way.

[3]The automatic replacement of words taken as misspellings by their allegedly correct form.

[4]The shortening of words (typically by dropping their ending) by the writer to save space under length limitations.

[5]The main index of the Brazilian Stock Exchange.

collected in 2014. Because the data precede Twitter's 2017 character limit expansion, all tweets conform to the original 140-character restriction. First released by Silva et al. (2020) annotated with emotions, the corpus was later extended with other stand-off layers, giving rise to DANTEStocks.

In its current version, DANTEStocks[6] includes: (i) annotation of emotions (by Silva et al. (2020)); (ii) morphosyntax (part-of-speech tags) and syntax (dependency relations) under the UD model (by Di-Felippo et al. (2024)), and (iii) named entities (by Zerbinati et al. (2024) and Piai et al. (2025)). The resource has also been partially annotated with *Abstract Meaning Representation* (AMR) graphs, providing it with a semantic layer (Ceregatto and Di-Felippo, 2025).

In all of DANTEStocks' annotation layers, each tweet was treated as the basic unit of analysis rather than being segmented into smaller units such as sentences, clauses, or phrases. Tweets were also not normalised, so as to preserve their original form. These design choices preserve the authenticity of the data and allow for the analysis of linguistic phenomena as they naturally occur in UGC, while also supporting the development of multilingual systems capable of handling real-world data.

3.1 The typology of UGC phenomena

We employed a refined version of the typology proposed by Scandarolli et al. (2023), theoretically grounded in the concept of linguistic *norm* (Coseriu, 1952), *i.e.* the set of socially accepted and recurrent linguistic usages within a speech community. This framework enables a linguistically principled interpretation of variation, distinguishing between *standard norm variation* and *innovative norm* categories (Table 1)[7]. Although intentionality may underlie these two categories, the focus remains on describing the linguistic processes that give rise to variation.

3.1.1 Standard Norm Variation

This category refers to orthographic variations from the standard language as described by Damerau: *substitution*, *omission*, *insertion*, and *transposition*. However, rather than restricting these operations to alphabetic letters, as in Damerau's original formulation, in this work a letter is defined as a Unicode character[8]. This adaptation allows the same set of operations to be applied to different types of elements (such as letters, spaces, hyphens, diacritics, and symbols), thus providing a more comprehensive and encoding-independent framework for describing orthographic variation in UGC.

In addition, and following Gimenes et al. (2015), substitutions, omissions, and insertions were further subdivided to account for variations involving cedilla, other diacritics, space, hyphen, and those affecting other kinds of characters. The transposition type, however, was not subdivided, since no further distinctions were required.

3.1.2 Innovative Norm

This category encompasses emerging or creative forms that reflect ongoing processes of linguistic innovation and adaptation, as well as domain- and medium-specific lexical items. It comprises 6 types, partly adapted from Sanguinetti et al. (2022): (i) abbreviation, (ii) neologism, (iii) mark of expressiveness, (iv) homophone writing, (v) medium-dependent token, and (vi) domain-specific issue.

Abbreviations are shortened forms of single or multiwords. *Initialisms* are abbreviations composed of the initial letters of a phrase, with each letter pronounced separately (*e.g.*, *conselho fiscal→cf* ("board of auditors")). *Shortenings* result from the omission of parts of a word and may involve the loss of initial (*e.g.*, *estou→tou* ("(I) am")), medial (*e.g.*, *também→tbem* ("too")), or final (*e.g.*, *para→p* ("to / for")) segments. *Contractions* occur when two closed-class words or multiword expressions merge into a single orthographic token, as in *oq* (*o* + *que* ("what")). This type also includes *truncations* and *alphanumeric abbreviation* (such as *1T14→primeiro trimestre de 2014* ("first trimester of 2014")).

Neologisms refer to any newly formed word. In DANTEStocks, neologisms arise from three main word-formation processes: (i) *blending*, when parts of two (or more) words are merged to create a new term that may retain a compositional meaning derived from the original words or acquire a new connotation, such as *Ibolixo* (which merges *Ibov(espa)* and *lixo* ("trash")); (ii) *derivation*, involving affixation, as in *diretassa*, where the augmentative suffix *-assa* (an informal variant of *-aça*) is added to *direta* ("direct") to convey an intensified or humorous meaning, and (iii) *foreign influence*, referring to

[6] https://sites.google.com/icmc.usp.br/poetisa/porttinari-2-1

[7] A single utterance (or word) may simultaneously instantiate multiple categories within this classification.

[8] https://www.unicode.org/versions/Unicode15.0.0/

Class	Type	Subtype	Attested example	Standard form	Gloss
Standard Norm Variation	Substitution	*Diacritic by other*	**sō**	*só*	`only/just´
		Hyphen by space	**segunda feira**	***segunda-feira***	`Monday´
		Other	**Awuele** shooting star	***Aquele*** *shooting star*	`that shooting star´
	Omission	*Cedilla*	**lancamento** das notas	***lançamento*** *das notas*	`notes issuing´
		Diacritic	capital **proprio**	*capital* ***próprio***	`equity capital´
		Hyphen	**presal**	*pré-sal*	`pre-salt (oil fields)´
		Space	**180d**	***180 dias***	`180 days´
		Other	**valu**	***valeu***	`thanks´
	Insertion	*Diacritic*	*#PETR4* **fêz** *uma Onda 2*	`#PETR4 **fez** uma Onda 2´	`#PETR4 made a Wave 2´
		Hyphen	data-folha	*Datafolha*	--
		Space	sub onda	*subonda*	`PETR4 had a minor dip´
		Other	montar um **Streaddle**	*montar um* ***Straddle***	`to set a Straddle´
	Transposition	--	vc se manteve na **comrpa**?	*vc se manteve na* ***compra****?*	`did you stick with stocks?´
Innovative Norm	Abbreviation	*Initialism*	membros do Cf	*membros do* ***Conselho fiscal***	`board of auditors´
		Shortening	(eles) falam **q** por **enqt**	(eles) *falam* ***que*** *por* ***enquanto***	`(they) say that´ `for now´
		Contraction	**Oq** vc acha @user?	***O que*** vc acha @user*?*	`What do you think @user?´
		Truncation	ação sobe **fo**...	*ação sobe* ***forte***...	`Stock rises sharply´
		Alphanumeric	**1T14**	primeiro trimestre de 2014	`first trimester of 2014´
	Neologism	*Blending*	44.6k no **Ibolixo**	-	`44.6 thousand in Ibotrash´
		Derivation	**diretassa** do morgan	***diretaça*** *do morgan*	`straight from morgan´
		Foreign influence	#itub4 **estopou**	*#itub4* ***estopou***	´#itub4 stopped´
	Expressiveness	*Graphemic stretching*	**chooooooram**!	***Choram***	`Cry!´
		Dialectal variation	De **zóio**!	*De* ***olho****!*	`(I am) keeping an eye!´
		Symbolism	**:)** muito **$**	- *muito* ***dinheiro***	`smile´ `much money´
		Capitalization	**LINNDAA**	***linda***	`beautiful´
		Disguise	essa **p******a**	*essa* ***puta***	`this bitch´
	Homophon eWriting	*Phonetization*	é **d+**	*é* ***demais***	`(it is) awesome´
		Graphemic substitution	**neh**	***né***	`isn't it´
		Onomatopoeia	**hahaha**	-	-
	Medium-dependen ttoken	*Hashtag*	Presidente da **#PETR4**	-	`President of #PETR4´
		At-mention	né, **@user**?	-	`isn´t it, @user?´
		URL	http://t.co/OQ3rDdWilf	-	-
		RT	**RT** @user...	-	-
		Code-switching	E ponto final! **PERIOD**!	E ponto final! **PONTO**!	`Full stop! PERIOD´
	Domain-specific token	*Ticker*	**PETR4** subiu	-	`PETR4 went up´
		Cashtag	$**PBR** testando	-	`$PBR (is) testing´
		Numeric approximation	**18,xx**	-	`18,xx´

Table 1: Taxonomy of UGC phenomena in DANTEStocks (Di-Felippo and Roman, 2025).

words from foreign lexical bases but following Portuguese morphological principles, such as *stopar*, which derives from the English expression "stop loss" or "stop gain".

Marks of expressiveness comprise: (i) *graphemic stretching*, where letters are repeated for emphasis or emotion (*e.g.*, *beleza→belezaaaa* ("nice")); (ii) *dialectal variation*, where words are spelled according to a regional or nonstandard variety (*e.g.*, *nóis* instead of *nós* ("we"); (iii) *symbolism* refers to the use of graphical signs, including emoticons, emojis, and other symbols that can function as substitutes for words; (iv) *capitalization*, used to convey emphasis or shouting; and (v) *disguise*, where offensive or taboo words are partially masked (*e.g.*, *puta→p**a* ("bitch").

Homophone writing encompasses orthographic deviations based on sound similarity. Within this group, (i) *phonetization* occurs when alternative letters or symbols are used to reproduce the sound of a word, such as *d+* for *demais* ("awesome")); (ii) *graphemic substitution* involves replacing diacritics or diacritic-bearing letters with phonetically similar characters, as in *neh* instead of *né* ("isn't it") or *ñ* instead of *não* ("no"); and (iii) *onomatopoeia* reproduces expressive noises (*e.g.*, *hahaha*).

Medium-dependent phenomena arise from the platform and include *hashtags, at-mentions, retweet markers*, URLs, and *code-switchings*. Although URLs and alternation between languages (code-switching) within the same utterance may also occur in texts written in standard language, they were considered here because they are highly characteristic of the domain and genre.

Domain-specific phenomena include linguistic and orthographic patterns that are characteristic of financial discourse, such as: (i) *ticker*, a five- or six-character alphanumeric string that identifies a com-

pany's stock (*e.g.*, *PETR4* for Petrobras' preferred shares); (ii) *cashtag*, *i.e.* a stock ticker preceded by a dollar sign ($)[9]; and (iii) *numeric approximation*, where final digits are omitted or replaced by placeholders (*e.g.*, *xx*) to indicate approximate or unspecified numerical values.

3.2 Data selection and annotation procedure

Due to time constraints, the annotation focused on the first 3,073 tweets (61,586 tokens) of the 4,042 (80,998 tokens) tweets in DANTEStocks (approximately 76% of the corpus). The annotation was carried out on a tweet-by-tweet and token-by-token basis by a Linguistics graduate with experience in Corpus Linguistics and NLP. The manual annotation sessions were held four times per week, each lasting one hour and covering 100 tweets. The annotation process spanned a total of two months.

To do so, DANTEStocks' CoNLL-U files[10] were adapted to keep only the columns relevant to token interpretation: ID, FORM, LEMMA, and UPoS. Two additional columns were then added to annotate the types and subtypes of each norm variation. All annotations were manually performed directly in a spreadsheet application, using abbreviated tags as exemplified in Table 2.

# sent_id = dante_01_446652807991791616l					
# text = PETR4, alta de 10% em 3 dias... rapááááz					
ID	**FORM**	**LEMMA**	**UPoS**	**I-N**	**S-N-V**
1	PETR4	PETR4	PROPN	**T**	
2	,	,	PUNCT		
3	alta	alta	NOUN		
4	de	de	ADP		
5	10	10	NUM		
6	%	%	SYM		
7	em	em	ADP		
8	3	3	NUM		
9	dias	dia	NOUN		
10	...	...	PUNCT		
11	rapááááz	rapaz	NOUN	**GS**	**I-D**

Table 2: Example of corpus annotation (I-N: innovative norm, S-N-V: standard norm variation, T: ticker, GS: graphemic stretching, I-D: insertion of diacritic).

Each *innovative norm* was annotated directly at the subtype level (*e.g.* ticker, truncation, and graphemic stretching), while the corresponding type (and class) was automatically inferred afterwards. In contrast, each instance of *standard norm variation* was assigned both a type and, when applicable, a subtype (*e.g.* insertion–diacritic), as shown in Table 2. In cases like *rapááááz*, where multiple diacritical marks are inserted as a result of graphemic prolongation, only one instance of diacritic insertion is counted, since the focus is on identifying types/subtypes rather than their frequency within a token.

4 Results

Overall, 10,271 tokens displayed some orthographic variation (nearly 17% of the 61,586 tokens in the annotated part of the corpus). A total of 10,465 instances of lexical and orthographic variation were annotated across these tokens, yielding an average of 1.02 phenomenon per token.

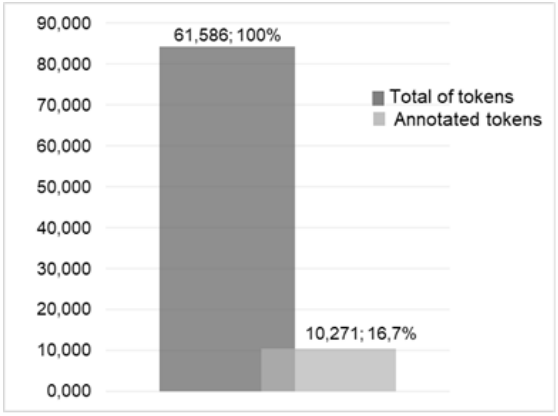

Figure 1: Proportion of annotated tokens in the corpus.

Of the 10,465 occurrences, 9,588 (91.6%) correspond to the *innovative norm*, spanning 9,560 tokens, whereas 877 (8.4%) fall under the *standard norm variation*, covering 711 tokens (Figure 2).

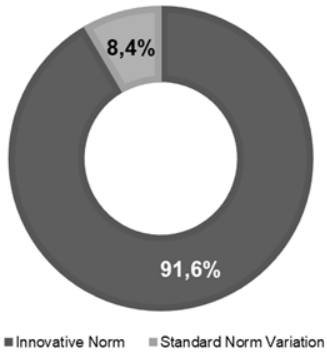

Figure 2: Distribution of norms in DANTEStocks.

The 9,588 *innovative norm* phenomena are distributed across types and subtypes, as shown in Table 3. All 6 types (abbreviation, neologism, expressiveness, homophone rewriting, medium- and

[9]Cashtags are used as reference to financial instruments, whereas hashtags serve broader topical purposes.

[10]CoNLL-U is a traditional column-based UD-encoding style (https://universaldependencies.org/format.html)

domain-specific phenomenon) and 24 subtypes are attested in the corpus.

Type	Subtype	Qt.	
Abbreviation	Initialism	77	767
	Shortening	492	
	Contraction	8	
	Truncation	187	
	Alphanumeric	3	
Neologism	Blending	15	23
	Derivation	6	
	Foreign influence	2	
Expressiveness	Graphemic stretching	15	540
	Dialectal variation	38	
	Symbolism	127	
	Capitalization	352	
	Disguise	8	
Homophone Writing	Phonetization	12	79
	Graphemic substitution	30	
	Onomatopoeia	37	
Medium-dependent token	Hashtag	1911	4,425
	At-mention	969	
	URL	1287	
	RT	257	
	Code-switching	3	
Domain-specific token	Ticker	3459	3,754
	Cashtag	293	
	Numeric approximation	2	
		TOTAL	9,588

Table 3: Statistics for the *innovative norm* in the corpus.

Among the identified types, medium-dependent (4,425 cases) and domain-specific tokens (3,754) are the most frequent ones, accounting for approximately 85% of all innovative norm occurrences. Following them come abbreviation (767) and expressiveness (540), reflecting the brevity of tweets (originally enforced by a strict 140 character limit and still favoured as a stylistic norm) and the affective tone typical of financial discourse on Twitter. In contrast, neologism is marginal, with only 23 occurrences (0.23% of the occurrences).

Figure 3 shows the overall frequency with which each *innovative norm* subtype occurs along the corpus. Ticker and hashtag are the most popular subtypes, totalling 3,459 and 1,911 instances, respectively. The prominence of tickers stems from the corpus compilation strategy, while the high frequency of hashtag reflects their indexical function. URLs and at-mentions are also common, with 1,287 and 969 instances, respectively. This reflects the referential and interactive nature of tweets from the stock market domain.

In contrast, foreign influence and numeric approximation are rare, with only two occurrences each. The former are *stopando*, from the English *stop* (as in *stop loss*) and formed with the productive Portuguese verbal suffix *-ar* ("to stop out"); and *scalperzinho*, from *scalper* (a trader who profits from small price movements) combined with the diminutive suffix *-zinho*, meaning "little scalper". Examples of the latter are *20,xx* and *9,XX*.

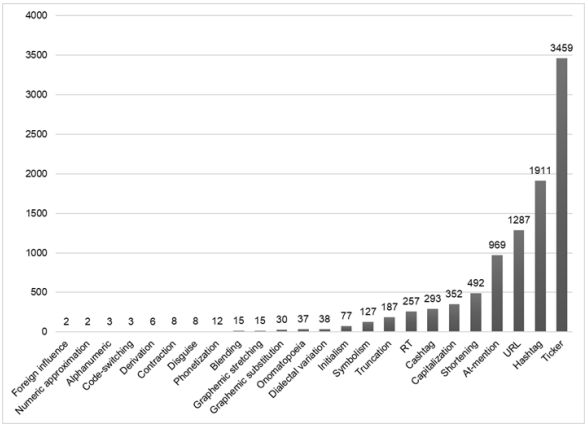

Figure 3: Distribution of the *innovative norm* subtypes.

A further observation regards the number of phenomenon types per token, which reflects the occurrence of multiple innovative norm operations within a single token. In this sense, virtually all annotated tokens (9,534 out of 9,560 – *i.e.* 99.7%) exhibit a single type, whereas 24 display two and 2 contain three types.

The most recurrent pairings of types within tokens, each occurring three times, are: (i) graphemic stretching and dialectal variation, illustrated by *bunituuu* for *bonito* ("beautiful"); (ii) shortening and phonetization, as in *krk* for *caraca* ("holy..."); and (iii) initialism and capitalization, as in *PQP* for *puta que pariu* ("bloody hell!"). The only two tokens exhibiting 3 distinct phenomena are: (i) *plz* ("please"), which combines code-switching, shortening, and phonetization, and (ii) *MTOOOO*, instead of *muito* ("very"), which combines shortening, capitalization, and graphemic stretching. Distributions concerning *innovative norm* remain consistent even when repetitive forms are excluded (*i.e.* when only distinct lemmas are kept).

The 877 examples of *standard norm variation* are distributed across types and subtypes as shown in Table 4. Omission is the most frequent type (739 occurrences; 84.2%), whereas transposition is the least frequent one (3 occurrences; 0.4%). The cases of transposition are *deixnado* (*deixando* – "leaving"), *grnade* (*grande* – "big"), and *reias* (*reais* – the Brazilian currency).

Regarding subtypes, diacritic omission and cedilla are the most frequent ones (Figure 4), with 498 and 167 occurrences, respectively. Notably, diacritic-related variations (including cedilla) correspond to approximately 85% of all cases in the cor-

Type	Subtype	Qt.	
Substitution	Diacritic by space	1	40
	Hyphen by space	3	
	Other character	36	
Omission	Cedilla	167	739
	Other diacritic	498	
	Hyphen	17	
	Space	9	
	Other character	48	
Insertion	Diacritic	78	95
	Hyphen	3	
	Space	8	
	Other character	6	
Transposition	Other character	3	3
		TOTAL	877

Table 4: *Standard norm variation* statistics.

pus. This means that over 85% of all single-word variations in DANTEStocks fall into one of Damerau's categories, thereby confirming Damerau's original statistics and the findings by Gimenes et al. (2015) regarding other UGC genres in Portuguese.

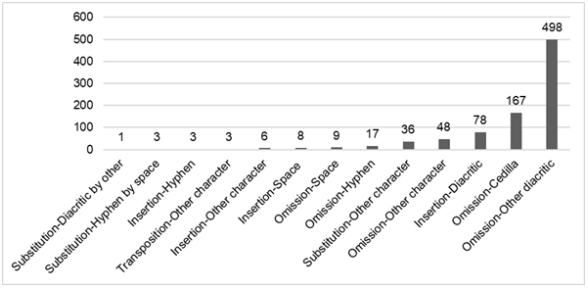

Figure 4: Distribution of *standard norm variation* subtypes.

Most annotated tokens (545; 85% of the 711) exhibit a single phenomenon, with 166 showing two. The most common combination (163 cases) involves the omission of both cedilla and diacritic (*e.g. resolucao* > *resolução* – "resolution"). Thus, we have once again confirmed Damerau's results, in that at least 85% of all tokens with *standard norm variation*, have a single phenomenon.

Beyond individual occurrences of each norm, 14 cases were identified in which phenomena from both norms – *standard* and *innovative* – occurred simultaneously, distributed across 10 distinct tokens. Thus, our investigation involved not only an intra-norm but also an inter-norm analysis, allowing for the identification of interactions between different patterns of orthographic variation. Table 5 presents the category, type and subtype associated with each norm, along with the tokens in which these co-occurrences were observed.

As observed with *innovative norm*, all distributions of *standard norm variation* remain consistent even after the exclusion of duplicate tokens and the accounting for distinct UD lemmas only. Finally, from this annotation we were able to build a lexicon with 2,747 distinct variant forms (from both norms), corresponding to 2,414 unique lemmas.

5 Coverage Analysis and Discussion

To evaluate the impact of our variant lexicon on vocabulary coverage across the 61,586 tokens in the 3,067 annotated tweets, we measured the reduction of out-of-vocabulary (OOV) items as produced by three different embedding models: BERTimbau (Souza et al., 2020), Word2Vec (Mikolov et al., 2013) and FastText (Hartmann et al., 2017). Our intent with this evaluation was to observe how each model's representational properties (contextual *vs.* static and word-level *vs.* subword-level) shape both their vocabulary coverage and the extent to which they benefit from the integration of our lexicon.

The choice for these models was guided by their fundamentally different approaches to lexical information codification, as evidenced by their OOV distributions (Table 6). In this case, BERTimbau, despite its strong contextual capabilities, has a much larger OOV set (23,966 tokens; 39%) than Word2Vec and FastText (both 13,960 tokens; 22.6%). This might be due to its reliance on a fixed WordPiece vocabulary (approximately 30k subwords) learned from formal corpora like brWaC and Wikipedia, which limits recognition of slang, misspellings, or hashtags.

The introduction of the variant lexicon substantially alters this landscape. For BERTimbau, the reduction from 23,966 to 12,230 OOV tokens illustrates how lexicon enrichment compensates for the model's limited subword granularity, enabling the recovery of more than half (51%) of the previously unseen items. This indicates that a curated lexical resource can effectively bridge gaps created by data-driven vocabulary truncation, particularly in architectures that do not dynamically generalize from orthographic variation.

Word2Vec and FastText, in turn, exhibit lower OOV rates because they incorporate all full words observed during training and, in FastText's case, they can also generate embeddings for previously unseen forms using subword units. Still, both models benefit from lexicon integration: the drop from 13,960 to 5,618 OOV tokens indicates that 8,342 tokens – nearly 60% of the initial set – were now covered. This outcome highlights the lexicon's

Innovative Norm	Standard Norm Variation	Token	Standard form	Qt.
Capitalization	Insertion-Diacritic	PETROBRÁS	Petrobras	5
		ELETROBRÁS	Eletrobras	
Capitalization	Omission-Diacritic	AMANHA	amanhã	4
		DIARIO	diário	
		MINIMO	mínimo	
		PARABENS	parabéns	
Blending	Insertion-Diacritic	PeTebrás	PeTebras	2
Capitalization	Omission-Space	LONG&SHORT	long & short	1
Graphemic stretching	Insertion-Diacritic	rapáááz	rapaz	1
Hashtag	Insertion-Space	# petr4	#petr4	1
			TOTAL	14

Table 5: *Innovative* and *standard variation norm* intersection.

Embedding Model	Original OOV	OOV after Variant Lexicon	Reduction Token (%)
BERTimbau	23,966 (38.9%)	11,736 (19.1%)	12,230 (51.0%)
Word2Vec	13,960 (22.7%)	5,618 (9.1%)	8,342 (59.8%)
FastText	13,960 (22.7%)	5,618 (9.1%)	8,342 (59.8%)

Table 6: Amount of Out-of-Vocabulary words by each model, with and without accounting for our lexicon.

value not only as a corrective mechanism but as a complementary resource that enhances coverage even for models known for their flexibility with noisy or morphologically rich data.

Although interesting, these results are restricted to approximately the first 3/4 of the corpus, with 969 tweets remaining ($\approx 24\%$) to be analysed. To address this limitation, we ran the three automatic models in these tweets, both with and without the lexicon derived from the annotated part, and verified the amount of OOV tokens resulting from each model. As illustrated in Figure 5, there is a very similar proportion of OOV tokens in both annotated and plain sets. This is another indication of our lexicon coverage and of its importance in reducing OOV tokens from automatic models.

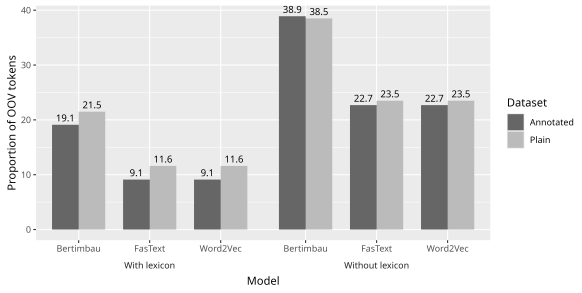

Figure 5: Proportion of OOV tokens in both annotated and non-annotated parts of the corpus, with and without the application of our lexicon of variants.

6 Final remarks

In this work, we argue that, regarding the linguistic characterization of social media language in the stock market domain, most variation corresponds to the innovative norm, driven by medium- and domain-specific adaptations. Abbreviations and expressive forms reflect stylistic conventions, while neologisms are rare. Standard norm variation is systematic, mostly involving simple character omissions, indicating stable rule-governed orthographic patterns in this subgenre.

Moreover, by determining how our lexicon of variants could reduce the amount of OOV tokens from automatic models, we evidenced its role in expanding vocabulary coverage regardless of representation type, thereby facilitating model adaptation to non-standard text and enabling a more robust processing of social media language. All trained models, along with our lexicon, can be found on the POETISA[11] project page. Finally, with respect to future improvements, we intend to test this lexicon in a larger corpus of financial market tweets, along with corpora from other domains.

Acknowledgments

This work was carried out at the Center for Artificial Intelligence of the University of São Paulo (C4AI – http://c4ai.inova.usp.br/), with support by the São Paulo Research Foundation (FAPESP grant #2019/07665-4) and by the IBM Corporation. The project was also supported by the Ministry of Science, Technology and Innovation, with resources of Law N. 8,248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published as Residence in TIC 13, DOU 01245.010222/2022-44.

[11] https://sites.google.com/icmc.usp.br/poetisa/resources-and-tools

References

Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

Gabriel Ceregatto and Ariani Di-Felippo. 2025. Dantestocks-amr em construção: Avanços e desafios na anotação semântica de tweets financeiros. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 608–617, Porto Alegre, RS, Brasil. SBC.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. Issues and perspectives from 10,000 annotated financial social media data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6106–6110, Marseille, France. European Language Resources Association.

Keith Cortis, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. 2017. SemEval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 519–535, Vancouver, Canada. Association for Computational Linguistics.

Eugenio Coseriu. 1952. Sistema, norma y habla. *Revista de la Facultad de Humanidades y Ciencias, Universidad de la República*, (9):113–181.

Fred J. Damerau. 1964. A technique for computer detection and correction of spelling errors. *Commun. ACM*, 7(3):171–176.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Justina Deveikyte, Helyette Geman, Carlo Piccari, and Alessandro Provetti. 2022. A sentiment analysis approach to the prediction of market volatility. *Frontiers in Artificial Intelligence*, 5.

Ariani Di-Felippo, Maria das Graças Volpe Nunes, and Bryan K. da S. Barbosa. 2024. A dependency treebank of tweets in brazilian portuguese: Syntactic annotation issues and approach. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 192–201, Porto Alegre, RS, Brasil. SBC.

Ariani Di-Felippo and Norton Trevisan Roman. 2025. DANTEStocks: a multi-layered annotated corpus of stock market tweets for Brazilian Portuguese. *Brazilian Journal of Applied Linguistics*, Corpus Linguistics: Studies and Applications:1–23. To appear.

Paul Gaskell, Frank McGroarty, and Thanassis Tiropanis. 2013. An investigation into correlations between financial sentiment and prices in financial markets. In *Proceedings of the 5th Annual ACM Web Science Conference*, WebSci '13, page 99–108, New York, NY, USA. Association for Computing Machinery.

Priscila Gimenes, Norton Roman, and Ariadne Carvalho. 2015. Spelling error patterns in brazilian portuguese. *Computational Linguistics*, 41:175–183.

Nathan Hartmann, Erick Fonseca, Christopher Shulby, Marcos Treviso, Juliana Rodrigues, and Sandra Aluisio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Proceedings of the Symposium in Information and Human Language Technology (STIL)*.

Huina Mao, Scott Counts, and Johan Bollen. 2011. Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data. Papers n. 1112.1051, arXiv.org.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR), Workshop Track*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Yulong Pei, Amarachi Madu, Akshat Gupta, Salwa Alamir, Hanxuan Lin, Xiaomo Liu, and Sameena Shah. 2022. Tweetfinsent: A dataset of stock sentiments on twitter. pages 37–47.

Laís Piai, Ariani Di-Felippo, and Norton Roman. 2025. Named entities in stock market tweets: A fine-grained and linguistically-motivated annotation. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 654–663, Porto Alegre, RS, Brasil. SBC.

Norton Trevisan Roman, Paul Piwek, Ariadne Maria Brito Rizzoni Carvalho, and Alexandre Rossi Alvares. 2013. Introducing a corpus of human-authored dialogue summaries in portuguese. In *RANLP*, pages 692–701.

Manuela Sanguinetti, Cristina Bosco, Lauren Cassidy, Özlen Çetinoğlu, Alessandra Teresa Cignarella, Teresa Lynn, Ines Rehbein, Josef Ruppenhofer, Djamé Seddah, and Amir Zeldes. 2020a. Treebanking user-generated content: a proposal for a unified representation in universal dependencies. In *Proceedings of the 12th International Language Resources and Evaluation Conference*, pages 5240–5250.

Manuela Sanguinetti, Cristina Bosco, Lauren Cassidy, Özlen Çetinoğlu, Alessandra Teresa Cignarella, Teresa Lynn, Ines Rehbein, Josef Ruppenhofer, Djamé Seddah, and Amir Zeldes. 2022. Treebanking user-generated content: a ud based overview of guidelines, corpora and unified recommendations. *Language Resources & Evaluation*, 57:493–544.

Manuela Sanguinetti, Lauren Cassidy, Cristina Bosco, Özlem Çetinoğlu, Alessandra Teresa Cignarella, Teresa Lynn, Ines Rehbein, Josef Ruppenhofer, Djamé Seddah, and Amir Zeldes. 2020b. Treebanking user-generated content: a ud based overview of guidelines, corpora and unified recommendations. *Preprint*, arXiv:2011.02063.

Clarissa Lenina Scandarolli, Ariani Di-Felippo, Norton Trevisan Roman, and Thiago A. S. Pardo. 2023. Tipologia de fenômenos ortográficos e lexicais em cgu: o caso dos tweets do mercado financeiro. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL 2023)*, STIL 2023. Sociedade Brasileira de Computação.

F. J. V. Silva, N. T. Roman, and A. M. B. R. Carvalho. 2020. Stock market tweets annotated with emotions. *Corpora*, 15(3):343–354.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: Pretrained bert models for brazilian portuguese. In *Proceedings of the Brazilian Conference on Intelligent Systems (BRACIS)*.

Michel Monteiro Zerbinati, Norton Trevisan Roman, and Ariani Di Felippo. 2024. A corpus of stock market tweets annotated with named entities. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese (PROPOR 2024)*, pages 276–284, Santiago de Compostela, Galicia/Spain.

Xue Zhang, Hauke Fuehres, and Peter A. Gloor. 2011. Predicting stock market indicators through twitter "i hope it is not as bad as i fear". *Procedia - Social and Behavioral Sciences*, 26:55–62. The 2nd Collaborative Innovation Networks Conference - COINs2010.