REVISTA DA **ABRALIN**

PROJECT REGISTRATION

The DANTEStocks Corpus: an analysis of the distribution of Universal Dependencies-based Part-of-Speech tags

Ariani DI FELIPPO
Federal University of São Carlos (UFSCar)

Norton Trevisan ROMAN
University of São Paulo (USP)

Thiago Alexandre Salgueiro PARDO
University of São Paulo (USP)

Lucas Panta de MOURA
University of São Paulo (USP)

OPEN ACCESS

EDITED BY
- Jorge Baptista (UAlg)
- Adriana Silvina Pagano (UFMG)
- Marta Deysiane Alves Faria Sousa (UFS)

REVIEWED BY
- Hadinei Ribeiro (UEMG)
- Amanda Rassi (UFSCar)

ABOUT THE AUTHORS
- Ariani Di Felippo
  Data Curation, Writing - Original Draft.
- Norton Trevisan Roman
  Conceptualization, Writing - Original Draft.
- Thiago Alexandre Salgueiro Pardo
  Writing - Review & Editing
- Lucas Panta de Moura
  Investigation, Formal analysis.

DATES
- Received: 20/11/2022
- Accepted: 20/12/2022
- Published: 09/09/2024

HOW TO CITE
Di Felippo, A.; Roman, N. T.; Pardo, T. A. S.; Moura, L. P. (2024). The DANTEStocks Corpus: an analysis of the distribution of Universal Dependencies-based Part-of-Speech tags. *Revista da Abralin*, v. 22, n. 2, p. 249-271, 2024.

ABSTRACT

In the research area of Natural Language Processing (NLP), Part-of-Speech (PoS) tagging is one of the first processes applied to input data (speech or written text). It is responsible for assigning a proper part-of-speech (or word class) to each word in a text. When it comes to User-Generated Content (UGC) (e.g., tweets), however, there are additional challenges that undermine current approaches to PoS tagging, and which call for NLP resources. These, however, have so far focused on UCG orthographic and lexical phenomena only (e.g., truncated word, graphical stretching, etc.), letting aside PoS itself. To help fill in this gap, in this article we characterise DANTEStocks - a corpus of stock market tweets annotated with morphosyntactic information - in terms of the distribution of the PoS tags present in it. With this effort, we intend to provide researchers a starting point for other investigations, along with a benchmark against which to compare other corpora. Specifically,

correctly characterising the corpus according to the PoS tags may support the investigation of the syntactic relations called dependencies, since some of them usually co-occur with specific PoS tags.

RESUMO

No Processamento Automático de Línguas Naturais (PLN), a etiquetação morfossintática é um dos primeiros processos aplicados aos dados de entrada (como um texto escrito). Ela é responsável por atribuir uma etiqueta a cada palavra do texto, a qual representa a sua classe gramatical adequada. Quando se trata de Conteúdo Gerado pelo Usuário (CGU) (como os *tweets*), há desafios adicionais que afetam as abordagens atuais de etiquetação e que exigem recursos de PLN. Estes, no entanto, têm focado apenas em fenômenos ortográficos e lexicais dos CGUs (como palavra truncada/quebrada, alongamento gráfico, *emoticons*, etc.), deixando de lado as categorias gramaticais. Para ajudar a preencher essa lacuna, caracterizamos o DANTEStocks - um *corpus* de *tweets* do mercado de ações anotado com informações morfossintáticas - em termos da distribuição de suas etiquetas morfossintáticas. Por distribuição, entende-se a frequência geral de cada etiqueta no *corpus* e no interior dos *tweets*. Assim, fornecemos aos pesquisadores um ponto de partida para outras investigações, juntamente com um referencial para comparação com outros *corpora*. Ademais, caracterizar corretamente o *corpus* de acordo com as etiquetas pode auxiliar na investigação das relações sintáticas chamadas dependências, uma vez que algumas delas geralmente ocorrem entre palavras de categorias gramaticais específicas.

KEYWORDS

Corpus. Tweet. Stock market. PoS tag.

PALAVRAS-CHAVE

Corpus. Tweet. Mercado financeiro. Etiqueta morfossintática.

RESUMO PARA NÃO ESPECIALISTAS

Identificar corretamente a classe gramatical (nome, adjetivo, verbo, etc.) de cada palavra é tarefa essencial para os sistemas de PLN baseados em conhecimento simbólico. Esse processo, chamado etiquetação morfossintática, associa uma etiqueta a cada palavra, representando a sua classe gramatical. Para a interpretação de um *tweet*, por exemplo, a etiquetação é mais complexa devido à linguagem não-padrão desses textos. Assim, descrever os fenômenos ortográficos e lexicais (como palavra

truncada/quebrada, alongamento gráfico, *emoticons*, etc.) tem fomentado estratégias de tratamento computacional dos *tweets*. Tal descrição, no entanto, não tem focado nas categorias gramaticais. Para ajudar a preencher essa lacuna, apresentamos uma investigação sobre a distribuição das etiquetas morfossintáticas no DANTEStocks, que é um *corpus* de *tweets* do mercado de ações em português. Por distribuição, entende-se a frequência geral de cada etiqueta no *corpus* e no interior dos *tweets*. A caracterização das categorias gramaticais no *corpus* pode ser usada em outras investigações, principalmente nas que objetivam comparar este com outros *corpora* (de outros gêneros ou domínios, por exemplo). Ademais, saber como as categorias gramaticais se distribuem em um *corpus* de *tweets* pode auxiliar na identificação das relações sintáticas chamadas dependências, uma vez que algumas delas ocorrem entre palavras de categorias gramaticais específicas.

Introduction

In Natural Language Processing (NLP), Part-of-Speech (PoS) tagging is one of the first processes applied to input data, being responsible for assigning a proper part-of-speech tag to each word in a text. Such basic information is useful for several NLP tasks and applications, as information extraction (CABRAL *et al.*, 2022) (BARBERO, 2022), semantic parsing (ANCHIÊTA; PARDO, 2022) (SENO *et al.*, 2022) and sentiment analysis (MACHADO *et al.*, 2022), just to cite a few of very recently published state of the art research for Portuguese.

PoS tagging can essentially be considered as solved for well-written texts (*i.e.* written text from traditional sources such as newspapers, books or academic papers), achieving accuracies over 97% for many languages (WU; DREDZE, 2019), including Portuguese (FONSECA *et al.*, 2015) (DE SOUZA; LOPES, 2019).

The development and availability of syntactically annotated corpora of UGC (especially containing tweets) over the last decade have generated a considerable amount of contributions on PoS tagging of UGC data, especially for the English language. A good proportion of the resources that contain syntactic (and also morphosyntactic) analyses have been annotated according to the Universal Dependencies (UD) model (NIVRE *et al.*, 2020), a dependency-based scheme which has become a popular standard reference for treebank annotation because of its adaptability to different languages, domains and genres (SANGUINETTI *et al.*, 2022).

For the Brazilian Portuguese language, efforts on PoS tagging for UGC are incipient. So far, to the best of our knowledge, Da Silva *et al.* (2021) have carried out the first and only investigation on PoS tagging for UGC. Specifically, the authors have proposed a customization of current state-of-the-art UD-based tagging strategies for Portuguese, achieving a 95% f-score. In that work, the

authors have used a corpus of tweets from the stock market domain, introduced by Da Silva *et al.* (2020) and further modified by Di Felippo *et al.* (2021), that integrates the Porttinari treebank (PARDO *et al.*, 2021). This resource, called DANTEStocks, intends to be the first (syntactically) annotated corpus of tweets (*i.e.*, a tweebank) in Portuguese. Although the mentioned accuracy is close to the state-of-the-art for standard language texts, the non-canonical language of CGU poses considerable challenges to their automatic processing.

Currently, descriptive investigations on UCG language have been focusing on systematizing common orthographic and lexical phenomena across tweebanks in different languages/domains (MELERO *et al.*, 2012) (SANGUINETTI *et al.*, 2022) or in similar data such as instant messages and blog posts (LYDDY *et al.*, 2014).

In this paper, we focus on describing how UD PoS tags are distributed within DANTEStocks. In a future work, we intend to compare this distributive characterization to others, obtained from corpora built from genres other than tweets and domains other than the stock market. The comparison could be used as a starting point to the development of tools tailored to specific genres or domains, as well as to the development of new theories on language use.

Moreover, the distributive characterization of the PoS tags within the corpus may support the investigation of the dependency relations to be annotated, since some of them usually co-occur with specific PoS tags (Duran, 2022). For example, oblique nominal relation (`obl`, in UD) usually depends on verb, adjective and adverb; nominal modifier relation (`nmod`) depends on noun, and proper nouns and some numbers (compound ones) are related by the `flat` relation. By knowing the way PoS tags are distributed along the corpus, one can infer prior probabilities for such relations, thereby speeding up automatic methods that might be applied in the annotation process.

The rest of this article is organized as follows. In Section 1 we present some current initiatives on the characterization of microblogging corpora. Next, in Section 2, we give a brief description of the UD model, so as to familiarize the reader with it. DANTEStocks and its annotation are described in Section 3, along with the questions we intend to answer with our research. Our analysis is then presented next, in Section 4, with our final remarks being presented in Section 5.

1. Related Work

Twitter is arguably the most popular microblogging platform to the moment. As such, it has attracted much attention in NLP and Linguistics. Overall, the Twitter posts are notoriously characterised by a non-standard written language, which is much closer to oral language than to standard edited text. Thus, tweets often contain ungrammatical sentence and phrase structures, non-standard words and domain-specific entities.

About the linguistic characteristics of such social media data, several studies have been performed to describe orthographic and lexical phenomena of tweets or similar types of UGC (instant messages and blog posts). Melero *et al.* (2012), for example, have studied deviations from standard language norm in a

Spanish corpus of texts from blogs, consumer reviews and Twitter, classifying each deviation as: (i) capitalization, (ii) graphical accent omission, (iii) punctuation (and whitespace) omission or reduplication, (iv) informal spelling (*i.e.*, systematic shortcuts and character substitutions intentionally made by users), or (v) conventional misspellings. Besides these classes, Lyddy *et al.* (2014) have also classified non-standard spelling utterances in a corpus of text-messages (SMS) in English as: (i) accent stylization, (ii) emoticons and typographic symbols, (iv) initialisms, (v) letter/number homophones, (vi) onomatopoeic/exclamatory expressions, and (vii) semantically unrecoverable words.

Following previous works such as (EISENSTEIN, 2013) and (LIU *et al.*, 2018), Sanguinetti *et al.* (2021) recently presented the most extensive description of orthographic and lexical phenomena based on corpora partially or entirely made up of tweets in different languages. The idiosyncrasies were classified along two major dimensions: canonicallness and intentionality. "Canonicallness" refers to whether a phenomenon is also observable in standard text, and "intentionality" refers to whether its production was deliberate. These two dimensions have "types" and "subtypes". As a result, the authors have proposed a hierarchy or typology of UGC phenomena. To illustrate, "marks of expressiveness" is a type of non-canonical and intentional phenomenon, which has several subtypes, to wit: punctuation reduplication (*e.g.* "Yes!!!"), graphemic stretching (*e.g.* "Yesss!"), emoticons and smileys. This typology resulted in a set of UD-based annotation guidelines to promote consistent treatment of the phenomena found in this type of texts.

As we can see from this brief review of the literature on linguistic characteristics of UGC, there seems to be no current investigation on the morphosyntactic dimension of this type of data, which could then be used to compare different corpora, so as to try to identify possible patterns and idiosyncrasies across genres and domains. In this work, we intend to help fill in this gap, by carrying out such a study on DANTEStocks, thereby providing a starting point for future research in this direction.

2. The Universal Dependencies Model

The UD model is an international project for building consensual annotation guidelines across different languages, searching for the so called "universality" in language structuring. UD specifies a complete morphosyntactic representation in which grammatical notions may be indicated via word forms (morphologically) or via dependency relations (syntactically).

The morphological specification of a (syntactic) word in the UD scheme consists of three levels of representation: (i) a lemma representing the semantic content of the word; (ii) a part-of-speech tag representing the abstract grammatical class associated with the word, and (iii) a set of features representing lexical and grammatical properties that are associated with the particular word form. Syntactic annotation in UD consists of typed dependency relations between words represented as a

tree. In such representation, "one word is the head of the sentence, dependent on a notional ROOT and all other words are dependent on another word in the sentence"[1].

The current version of this grammatical model (UD v2) encompasses 17 part-of-speech (PoS) tags and 37 dependency relations that have been evolving through time with the collaboration of a large research community, being already adopted by more than 100 languages for the production of over 200 treebanks (NIVRE *et al.*, 2020).

This dependency model has become a popular standard reference for treebank construction in the NLP field because of its adaptability to different languages, domains and genres, earning the status of a linguistic theory (DE MARNEFFE *et al.*, 2021). As practice has shown, UD is flexible enough to accommodate, for example, the idiosyncrasies of UGC. At the morphological level, the model prescribes mechanisms (*i.e.*, features) for describing spelling variations arising from abbreviation and typos, which are two very frequent phenomena of UGC. At the PoS level, UD provides tags (such as X) to annotate tokens that are not properly recognised as usual words and, at the syntactic level, specific dependency relations (*e.g.*, *parataxis*) allow to structure sentences (or tweets) that have unconventional punctuation.

To illustrate the tweet challenges and the UD analysis style, see Figure 1, which shows an annotated tweet from DANTEStocks. Since the current version of the corpus only have PoS marks used in UD, the dependence relations in Figure 1 was manually annotated according to Duran (2022) and Sanguinett et al (2022). For the sake of brevity, the figure does not present the lexical and grammatical features, but we illustrate them with the noun *acordo* (in English, "agreement"), which has the following features-values: *Gender=Masc* and *Number=Sing*.

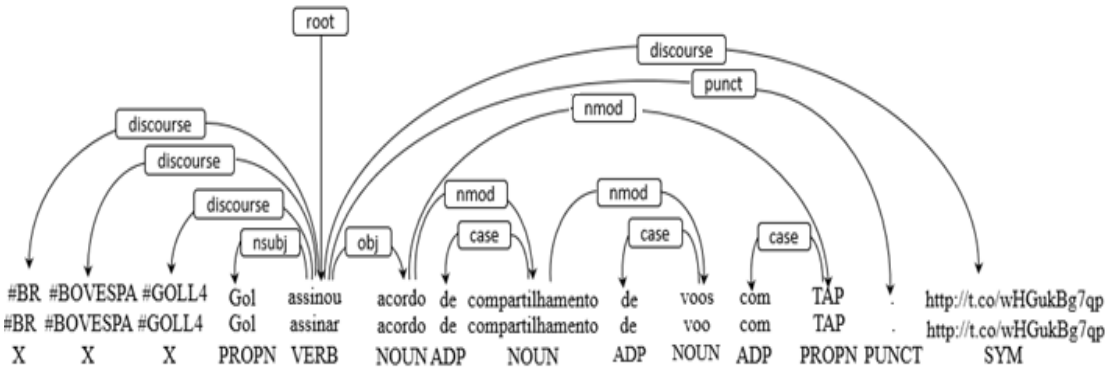

FIGURE 1 – Example of UD-annotated tweet from the DANTEStocks.
Source: authors.

The figure shows a tweet in Portuguese from DANTEStocks[2] where we can see marks of the text genre (UGC) (as *hashtags* and URL) and the complexity of the annotation that is intended. Above the

1 https://universaldependencies.org/u/overview/syntax.html

[2] Excluding the *hashatgs* and URL, it correponds "Gol signs a codeshare agreement with TAP."

words, there are the dependency relations codified by labelled arcs, and below the words, there are the PoS tags in capital letters and the lemmas of the words.

Given the immense popularity of social media as an attractive source of data for a large number of research fields and applications and the prominence of UD, the last decade has seen numerous Twitter corpora in different languages that have been annotated according to the aforementioned model and a proposal of UD-based annotation guidelines to promote consistent treatment of the phenomena found in UGC (SANGUINETTI *et al.*, 2022).

3. Materials and Methods

In this analysis, we rely on the DANTEStocks corpus (DI FELIPPO *et al.*, 2021), which builds on a corpus of stock market tweets written in Portuguese, where each tweet was annotated according to the emotion it bears, following Plutchik's Wheel of Emotions (see Da Silva *et al.* (2020) for details). The original corpus was automatically collected during the year of 2014, by fetching tweets mentioning ticker codes of any of the 73 stocks that compose IBOVESPA, the main index at B3 -- the Brazilian Stock Exchange.

A ticker code corresponds to a five to six-character alphanumerical string, such as PETR3 for Petrobras' ordinary stocks, representing a particular kind of stock of some company. DANTEStocks contribution to the corpus presented by Da Silva *et al.* (2020) was to add an extra annotation layer to it, by tokenizing it and assigning each token a PoS tag, according to the UD model. In its current version[3], DANTEStocks comprises a total of 4,048 tweets[4], corresponding to 81,037 tokens, annotated with their corresponding PoS tags.

For the morphosyntactic annotation of DANTEStocks, annotators used version 2 of the UD guidelines, which defines 17 PoS tags[5]. This task was carried out in a semi-automatic approach due to the expensive nature of manual tagging. This means that the tweets were automatically tokenized and tagged by a version of UDPipe2 (STRAKA, 2018) that was customized for the corpus (see Da Silva *et al.* (2021) for details), and then manually revised. The revision effort was carried out by a team of senior undergraduate students of Linguistics, under the supervision of a team of linguists and computer scientists, during the year of 2020.

To help the experts in this task, an annotation manual was built with specific guidelines for the phenomena that are typical of stock market tweets (see Di Felippo *et al.* (2022)). This supporting

---

[3] The corpus is currently being annotated with UD dependencies, so it is still under constant construction.

[4] The total amount of tweets differs from the 4,517 reported by Da Silva *et al.* (2020) and Di Felippo *et al.* (2021) because, during the PoS annotation, some out of domain tweets were found, which were erroneously added to the corpus by the automatic procedure.

[5] See https://universaldependencies.org/u/pos/all.html

material was necessary since the annotation guidance described in the UD literature did not provide strategies for the proper treatment of DANTEStocks lexical idiosyncrasies. Experts could also rely on an annotation manual containing guidelines for the Portuguese language[6]. As an example of the tokenization and PoS tagging in DANTEStocks, consider the following tweet (1a):

> (1a) "#VALE5 é #VENDA? rsss #DEAL! #DEAL! #DEAL! '16 de março às 12:12' após vencto das opções podem puxar na...http://t.co/4mOMj1Om7d"[7]

which, after tokenised and PoS tagged, becomes:

> (1b) #VALE5/**PROPN** é/**AUX** #VENDA/**NOUN** ?/**PUNCT** rsss/**X** #DEAL/**NOUN** !/**PUNCT** #DEAL/**NOUN** !/**PUNCT** #DEAL/**NOUN** !/**PUNCT** `/**PUNCT** 16/**NUM** de/**ADP** março/**NOUN** a/**ADP** a/**DET** 12:12/**NUM** '/**PUNCT** após/**ADP** vencto/**NOUN** de/**ADP** as/**DET** opções/**NOUN** podem/**AUX** puxar/**VERB** em/**ADP** a/**DET** .../**PUNCT** http://t.co/4mOMj1Om7d/**SYM**

As it can be seen, the text does not comply with the standard rules for writing (especially regarding capitalization, punctuation, and abbreviations), also presenting elements that are characteristic to the platform where it was written (*e.g.* the presence of hashtags and URLs).

Tweet sizes in DANTEStocks vary from a couple up to 71 tokens in a single message (a mean of 20 tokens per tweet). It is worth mentioning that, by the time the corpus was collected, Twitter users were limited to messages no longer than 140 characters. Each message, in turn, comprises from one to a maximum of 16 different tags (found in a tweet with 31 tokens in total), with a mean value of 8.5 different tags per tweet. Tables 1 and 2 summarize these statistics. In Table 2, outliers are taken to be points below Q1 – 1.5 IQR - where Q1 is the first quartile and IQR is the interquartile range - and above Q3 + 1.5 IQR). All in all, this is currently one of the largest corpora of stock market tweets in Brazilian Portuguese annotated under the UD paradigm.

Total amount of tweets	4,048
Total amount of tokens	81,037
Tweet size	From 2 to 71 tokens (mean of 20 ± 7.8)
Different tags in a tweet	From 1 to 16 (mean of 8.5 ± 2.2)

TABLE 1 – Total, minimum, maximum, mean and standard deviation figures for DANTEStocks. Source: authors.

[6] Available at https://sites.google.com/icmc.usp.br/poetisa/publications

[7] "#VALE5 is #SALE? rsss #DEAL! #DEAL #DEAL! 'march 16 at 12:12' after option expiration you can pull it in at...http://t.co/4mOMj1Om7d"

	Q1	Median	Q3	> Q3+1.5 IQR	< Q1-1.5 IQR
Tweet size	15	20	26	10	0
Different tags in a tweet	7	9	10	2	14

TABLE 2 – Median, quartiles and number of outliers breakdown of Table 1.
Source: authors.

When analyzing DANTEStocks, we were mostly interested in determining how the PoS tags were distributed along the corpus, so as to characterize it in terms of how morphosyntactic information is used in the domain. As such, in this work we address the following questions:

1. What is the overall distribution of PoS tags along the corpus? And

2. How are the tags spread across the corpus?

To answer the first question, we determined the overall frequency of each PoS tag, by counting the amount of times each tag was used along the corpus. This frequency, in turn, will allow us to compare this corpus to others, so as to determine potential differences in their tag distributions. For the second question, we calculated the amount of tweets where each of the tags occurs, thereby determining how widespread is their use along the corpus. Under this approach, we can not only have an idea of the most popular tags in this domain, given their high overall frequency, but also build an intuition as to which tags can be considered more important, given the number of different tweets containing them.

4. Results and Discussion

Figure 2 shows the overall frequency with which each PoS tag occurs across the corpus. As it turns out, all of the seventeen tags proposed by UD can be found in DANTEStocks. Interestingly, *PUNCT* is the most frequent one, with around 16% of all tokens being assigned this tag, followed by *NOUN*, with around 15%, and *PROPN*, corresponding to approximately 14% of all the tags in the corpus. Together, these three tags add up to almost half of all tags (around 45%).

PUNCT's high frequency may be due to the fact that unconventional uses of punctuation, specially reduplication, are fairly common in DANTEStocks' tweets. To add to this characteristic, by the time DANTEStocks was created, there has been a design decision to split up repeated punctuation marks (*e.g.* "!!!!" became four "!" tokens), with each mark being annotated as a single token. These features, in turn, affect the number of tokens annotated as *PUNCT*, inflating it, and might explain this figure in DANTEStocks.

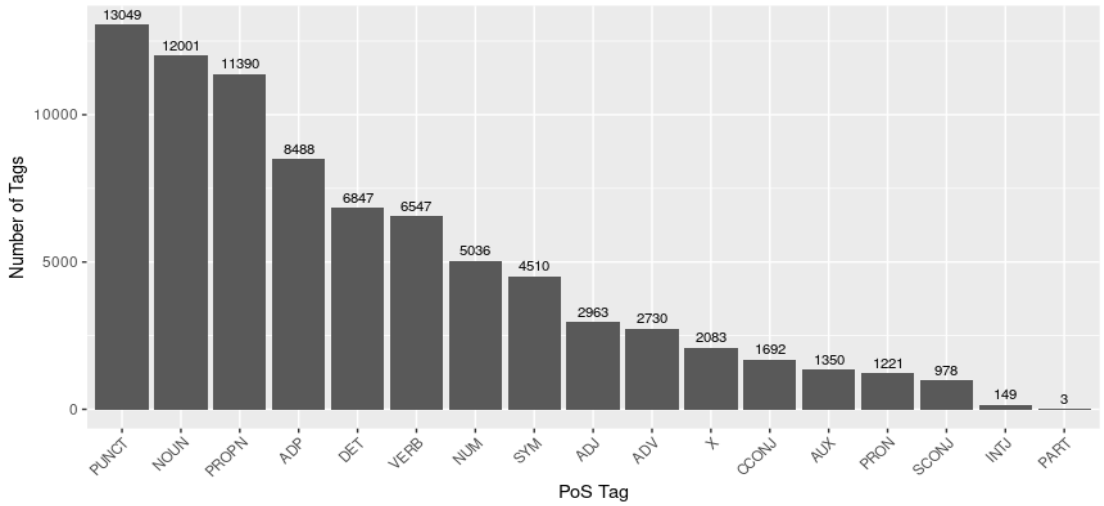

FIGURE 2 – Overall PoS tags distribution.
Source: authors.

The high frequency of NOUN tags is actually inline with current research in Linguistics, since predicative nouns seem to be used more frequently in UGC about financial market than other types of predicates, such as verbs, for example (see Voskaki *et al.*, (2016)). Finally, the prominence of PROPN reveals another characteristic of the corpus. Specifically, the high frequency of this tag results from the fact that the tickers used to compile the tweets were annotated as PROPN, since they are commonly used as a surrogate for their company names.

At the other end of the scale, we find PART, with only three occurrences in the corpus (*i.e.* 0.004%). About this tag, it is important to mention that, according the general UD guidelines, PART covers function words that must be associated with another word or phrase to impart meaning and that do not satisfy definitions of other universal parts of speech. In DANTEStocks, PART was only applied to the prefixes "*des-*" and "*pré-*" used as free forms in two occurrences of the same neologism (*e.g.* "*(des) Graça Foster*") and in a case of wrongly split word (*i.e.* "*pré abertura*"[8] instead of "*pré-abertura*").

Next to PART, we find INTJ, which corresponds to 0.18% of the total number of tags. The low frequency of INTJ may be explained by the fact that this tag should be associated with a word that is used most often to express an emotional reaction. In a corpus of UGC (like DANTEStocks), however, users commonly apply other marks or linguistic devices to express emotions, such as duplicated punctuation, graphemic stretching, emoticons and smileys (LIU *et al.*, 2018, SANGUINETTI *et al.*, 2021).

Regarding the low frequency of SCONJ, PRON, AUX and CCONJ, we believe that all these cases might be caused by the size limit imposed to tweets by the time the corpus was built (i.e., 2014). For PRON, its frequency might also be related to the domain, since stock market tweets tend to be very factual, containing no significant occurrence of certain types of pronouns such as personal, reflexive,

[8] Such typos are not normalised in the corpus to keep the tokens as they originally occurred.

interrogative, or demonstrative. In fact, we believe that most of *PRON* tags were assigned to relative pronouns, but this still needs to be checked.

About *SCONJ* and *CCONJ*, it is worth mentioning that these (and mainly *SCONJ*) are tags that should be used in tokens that link constructions (or clauses). Thus, given the tweets length limit, it is not surprising to find these two tags with low frequencies. We can say something similar about *AUX*. Once auxiliaries are used to compose passive voice constructions, which are longer than constructions in the active voice, users might be choosing the later in tweets. Finally, and as a design decision in DANTEStocks, X was assigned to hashtags and cashtags (*e.g.* #petr4 and $petr4, respectively) where these are used for indexing purposes only (*i.e.* with no syntactic function), which are not frequent.

In the middle of the scale, we find SYM, which is more frequent than ADJ and ADV, for example. One possible reason for this is the fact that the frequent unconventional use of punctuation, especially strings of repeated punctuation marks (e.g., !!!!), was split so that each punctuation mark was considered an individual token (e.g., "!" + "!" + "!" + "!").

Regarding the frequency of each tag inside the tweets, figures vary from no occurrence at all (*i.e.* the tag does not show up in the tweet) up to 32 times in a single tweet, with *NUM* (Figure 3). Although this may come up as no surprise, since these are tweets from the stock market domain, the tweets where this happened had, in fact, a political content. In one of them, the author only repeated the number of a political party over and over, blaming it for the profit results of a state company, whereas the other retweeted this content, adding a little more to it. As it seems, not only the domain plays a role in the PoS tags one finds in tweets, but the political moment also has something to contribute to it.

These values can be seen in more detail in Figure 3, which summarizes some statistics on the frequency of each tag within the tweets where they can be found. As it turns out, the smallest frequency of each tag corresponds to a single occurrence[9], meaning that no tag starts with more than one occurrence inside a tweet. At the other side of the scale, two tags stand out: *NUM*, with 32 occurrences in two tweets, and PUNCT, which can be found 30 times in one tweet and 29 in another. Interestingly, these are the same tweets with the 32 occurrences of *NUM*. This, however, must not be taken as a widespread phenomenon, since both examples are outliers, as shown in the figure.

[9] Recall that these are tweets in which the tag occurs.

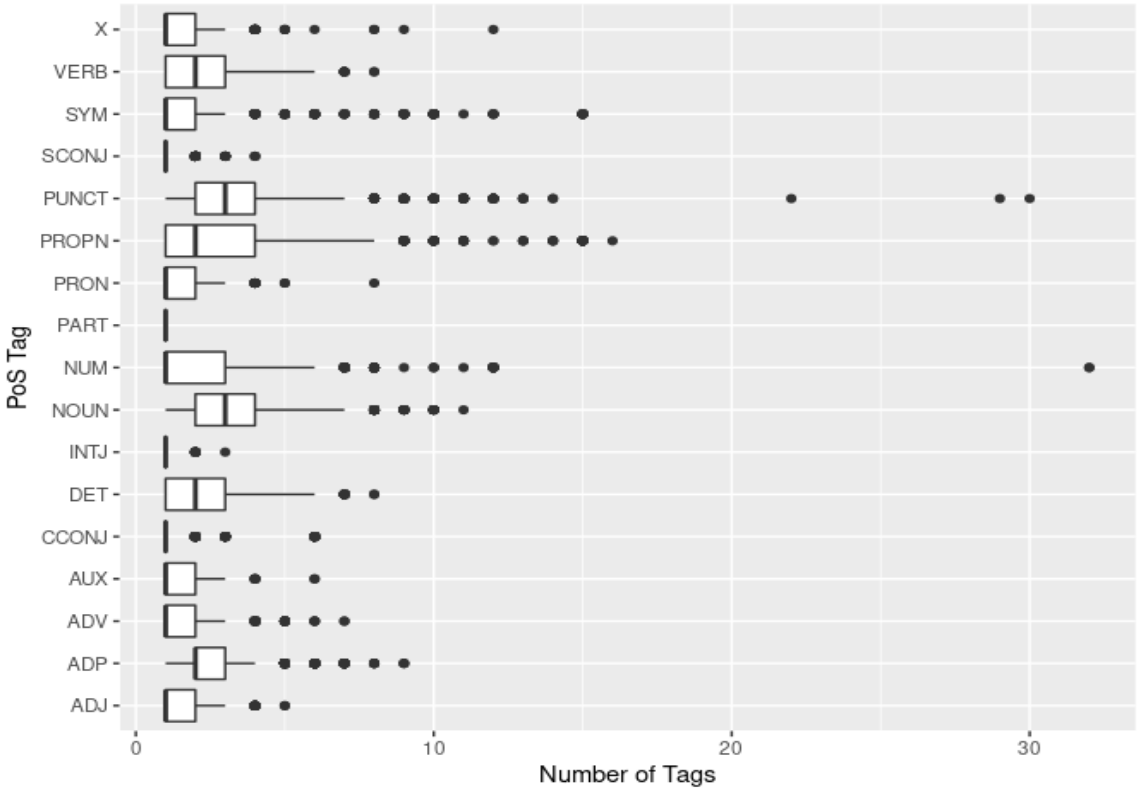

FIGURE 3 – PoS tags frequency inside the tweets where they occur.
Source: authors.

Interestingly, of the 17 tags that can be found in the corpus, in 11 of them the median lies at the minimum frequency, *i.e.* in at least half of the tweets where these tags occur, they happen only once. This is the case with X, SYM, SCONJ, PRON, PART, NUM, INTJ, CCONJ, AUX, ADV and ADJ. Moreover, in only five tags (VERB, PUNCT, PROPN, NOUN and DET) the median lies inside the interquartile range (*i.e.* above the first quartile), with one of them (ADP) lying at the first quartile's border.

This, in turn, along with the fact that outliers[10] can be found at the upper side of the scale only, indicates the distribution of tags inside tweets to be right skewed. As it seems, the limitation on text size imposed by Twitter, which naturally leads to smaller texts, may also lead people to be as short as possible when it comes to the syntactic choices they make, thereby preferring structures with a small number of different morphosyntactic categories over more diversified ones. This, however, is still to be determined.

Regarding tag coverage in DANTEStocks, Figure 4 shows the number of different tweets where each tag happens at least once. Not surprisingly, the top three most frequent tags in the corpus (*i.e.* PUNCT, NOUN and PROPN) are also the ones that can be found in the highest number of tweets, although in a different order of frequency, being found in over 90% of all tweets. The most commonly

[10] Values above 1.5 times the interquartile range.

found tag, in this case, is *PROPN*, which can be found in 98% of the tweets. In the sequence, we find NOUN, present in 94.3% of all tweets, and *PUNCT*, which covers 91.3% of the corpus.

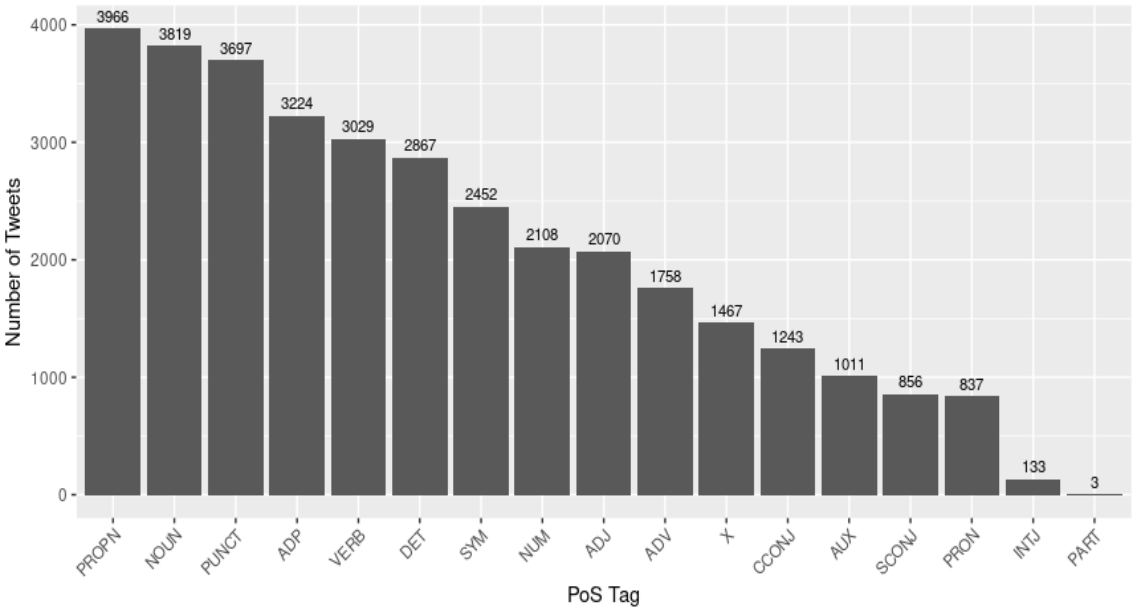

FIGURE 4 – Number of distinct tweets where each PoS tag occurs.
Source: authors.

Of the 17 tags present in DANTEStocks, nine (*PROPN*, *NOUN*, *PUNCT*, *ADP*, *VERB*, *DET*, *SYM*, *NUM* and *ADJ*) can be found in more than 50% of the tweets (*i.e.* more than 2,024 tweets). That means around half of the tags can be found in over half of the tweets in the corpus. With a mean value of 8.5 different tags per tweet, and recalling that the top three tags are present in over 90% of all tweets, these nine tags dominate a good deal of the corpus.

At the bottom line, we find *PART*, present in only three tweets (~ 0.07% of the corpus), and INTJ, with 133 occurrences (~ 3.3%). Once again, these are also the same tags with the smallest overall frequency in the corpus. Possible reasons for this behavior have already been discussed earlier in this section. Finally, it is interesting that *PRON*, which stands mostly for all types of pronouns, can be found in 20.7% of the corpus only. As we pointed out earlier, it might be related to the factual nature of the tweets from DANTEStocks and to the Twitter's message length limit.

Regarding specifically the PoS tags PUNCT, NOUN and PROPN, which are the most frequent ones in the entire corpus, occurring in more than 90% of the tweets, we present some observations on their impact on the annotation of dependency relationships (deprels). This, by the way, is the next step in DANTEStocks development.

As noted, the PUNCT tag is the most frequent in the corpus. With that, we know in advance that the DANTESocks annotation at the syntactic level will involve the identification of many punct relations, since this dependency relation occurs between a PUNCT tag (i.e., a punctuation symbol) and a word of content from its neighborhood. In this sense, the following examples (2-4) illustrate the main phenomena related

to punctuation, which are sequences of repeated punctuation marks (2) and long lists of items, separated by vertical bars (3) or commas (4). As every punctuation mark will depend on a punct deprel, these examples show that the corpus will be strongly marked by this deprel.

(2) vale petr4 hrt Explode cotação!!!!!!!!!!!!!![11]

(3) Maiores Altas: LLXL3 +7,79% | OIBR4 +6,14% | JBSS3 +6,11% | RSID3 +5,67% | ELET3 +4,83%.[12]

(4) Maiores Baixas: MRVE3 -12,50% R$7,35, DASA3 -9,67% R$15,13, CMIG4 -5,69% R$12,94, GFSA3 -4,76% R$3,00, ELPL4 -4,03% R$7,62.[13]

Regarding the PROPN tag, considering that the corpus is on the stock market domain, a portion of the tokens annotated as PROPN make up multiword expressions that usually designate companies, such as "CLWP Brasil III Participações" (5) and "Tim Part S/A" (6). Since the tokens that make up these expressions (such as "Tim", "Part" and "S/A" in (6)) are somehow related, but not syntactically related, there is no naturally identifiable head and dependent. Instead, we relate them with the flat relation. On the other hand, some multiword expressions do have a syntactic relationship between their tokens, such as in "Companhia Energética de Minas Gerais" (6), which must be annotated with normal relationships. In these cases, the deprels that will probably predominate are amod (between "Companhia" and "Energia", for example) and nmod (between "Companhia" and "Minas", for example). Furthermore, it was observed that DANTEStocks is characterized by the occurrence of tweets composed of a noun phrase, such as "Maiores Altas" (3) and "Maiores Baixas" (4), followed by a colon and a list of PROPN tags. The occurrence of this type of construction highlights the probable prominence of the appos deprel, which, in this case, is established between the nominal to the left of the colon (head) and the first PROPN in the list (dependent). Among the elements of this type of list, in particular, we believe that the most appropriate relationship is list, since the last item in the list is not introduced by an “and”. By the way, lists like the one exemplified in (8) -- very common in the corpus -- also show that we will often see the nummod deprel between the PROPN tags and a numerical modifier (NUM PoS).

---

[11] Vale petr4 hrt Heart blowing!!!!!!!!!!!!!

[12] Top Highs: LLXL3 +7,79% | OIBR4 +6,14% | JBSS3 +6,11% | RSID3 +5,67% | ELET3 +4,83%.

[13] Top Lows: MRVE3 -12,50% R$7,35, DASA3 -9,67% R$15,13, CMIG4 -5,69% R$12,94, GFSA3 -4,76% R$3,00, ELPL4 -4,03% R$7,62.

(5) $TBLE3 - TRACTEBEL (TBLE3) adquire totalidade do capital social da CLWP Brasil III Participações http://t.co/U6Ye8zmmjk[14]

(6) $TIMP3 - Tim Part S/a (timp-nm) - Arquivamento Do Formulario Anual 20-f http://t.co/OLZrEpFUcJ[15]

(7) $CMIG4 - Aviso de Vídeo Webcast: A Companhia Energética de Minas Gerais - CEMIG anuncia a Vídeo Webcast e http://t.co/bU05WM0FUN[16]

(8) Maiores Altas: ALLL3 +7,16% | CRUZ3 +6,71% | PRML3 4,35% | CSAN3 +2,98% | ELET3 +2,64%.[17]

Finally, regarding the NOUN tag, we believe that many of the corresponding tokens must be predicate names, since, as mentioned, these tend to be frequent in the stock market/financial/economic domain, especially in digital genres. Thus, names such as "totalidade" and "capital" in (5), "arquivamento" and "formulário" in (6) and others (e.g., "venda", "compra", "valorização", etc.) may occur with complements. In such cases, the head is a nominal and the dependent is a modifier, which can commonly be nmod (between "totalidade" and "capital", for example) or amod (between "capital" and "social", for example).

Conclusion

In this article, we reported on some statistics regarding UD PoS tags distribution in DANTEStocks - a corpus of stock market tweets annotated with morphosyntactic information. Amongt our main contributions, we highlight the characterization of the corpus, thereby providing researchers a way to compare DANTEStocks to other corpora, possibly from different genres and domains. To this end, we present two different distributions, to wit, the overall distribution of tags across the corpus, and the coverage of each tag, in terms of the number of different tweets where they can be found.

---

[14] $TBLE3 - TRACTEBEL (TBLE3) acquires the totality of CLWP Brasil III Participações' share capital http://t.co/U6Ye8zmmjk

[15] $TIMP3 - Tim Part S/a (timp-nm) - 20-f Annual Form Filing http://t.co/OLZrEpFUcJ

[16] $CMIG4 - Vídeo Webcast Notice: The Companhia Energética de Minas Gerais - CEMIG announces Vídeo Webcast and http://t.co/bU05WM0FUN

[17] Top Highs: ALLL3 +7,16% | CRUZ3 +6,71% | PRML3 4,35% | CSAN3 +2,98% | ELET3 +2,64%.

These figures could in turn be used to compare this corpus to others in the same genre and domain, or even as a first step in determining how apart corpora from different genres and domains can lie. Given enough research in this direction, this information could be used as a way to build some sort of corpora proximity measure based of UD PoS tags distribution.

To the best of our knowledge, this is one of the few, and possibly the only research so far to focus on how morphosyntactic information is used in a corpus of UGC and, more specifically, stock market tweets. As a consequence, our results can serve as a starting point for other investigations and, as mentioned, comparisons with other efforts on the same direction, so as to advance our knowledge not only in terms of the linguistic tools used in different genres and domains, but also on how automatic techniques can be tailored to them.

As an additional contribution, we also presented some possible reasons for the observed phenomena, such as the small number of each tag occurrence within tweets, for example, which leads to a right skewed distribution of these tags inside the tweets (*i.e.* their concentration at the smallest side of the scale). In this case, people are actually using a higher number of different tags, but a smaller amount of times, *i.e.* they might be preferring diversity over quantity. To what extent this is a DANTEStocks feature, a domain feature, or even a feature of the microblogging genre is still to be determined by future investigation.

Acknowledgements

The authors would like to thank the Center for Artificial Intelligence (C4AI-USP).

Additional information

Evaluation and authors' response

Evaluation: https://doi.org/10.25189/rabralin.v21i2.2119.R

PUBLISHERS

Jorge Manuel Evangelista Baptista
Affiliation: University of Algarve
ORCID: https://orcid.org/0000-0003-4603-4364

Adriana Silvina Pagano
Affiliation: Federal University of Minas Gerais

ORCID: https://orcid.org/0000-0002-3150-3503

Marta Deysiane Alves Faria Sousa
Affiliation: Federal University of Sergipe
ORCID: https://orcid.org/0000-0002-0480-0422

EVALUATION ROUNDS

Evaluator 1: Hadinei Ribeiro Batista
Affiliation: State University of Minas Gerais
ORCID: https://orcid.org/0000-0002-3157-6366

Evaluator 2: Amanda Pontes Rassi
Affiliation: Federal University of São Carlos
ORCID: https://orcid.org/0000-0001-5314-1868

EVALUATOR 1

O artigo apresenta movimentos retóricos previstos no processo de textualização de uma pesquisa. Tem como objetivo caracterizar o corpus DANTEStocks em termos de sua distribuição morfossintática. Metodologicamente, os autores informam que as etiquetas foram inseridas de forma semiautomática e revisadas por especialistas da computação e da linguística. Os resultados apresentados são compostos de uma lista de frequência dessas etiquetas. Do ponto de vista linguístico, não se observa qualquer detalhamento da tomada de decisões que levaram à etiquetagem. As afirmações são vagas, como 'frases curtas', 'textos menores', 'categorias morfossintáticas simples e complexas', etc. Para além da contagem de palavras e etiquetas (nomeadas como morfossintáticas), não se observa nenhuma análise linguística acurada. Nesse âmbito, os resultados e conclusões, superficialmente, apresentam afirmações com modalizadores (como *might be*) ou crenças dos autores, haja vista o emprego de verbos como '*to believe*' e '*to tend*'. Recomenda-se o desenvolvimento de uma argumentação consolidada do tratamento linguístico do estudo.

EVALUATOR 2
TÍTULO

O título é adequado e expressa exatamente aquilo que o artigo faz, que é prover uma análise da distribuição dos POS em um corpus chamado Dantestocks, seguindo diretrizes da Universal Dependencies.

RESUMO

O resumo está claro e compreensível, apresentando de forma sucinta o contexto, a lacuna e o objetivo da pesquisa. Faço apenas 2 ressalvas:

1. No Resumo (em português), onde se lê "a qual representa a sua categoria gramatical (parte do discurso) adequada", sugiro substituir "parte do discurso" por "classe de palavra" ou "classe gramatical", que são termos mais adequados para representar "part-of-speech"; e
2. No Lay Summary, onde se lê "Para os computadores interpretarem adequadamente um texto, identificar corretamente a categoria gramatical (nome, adjetivo, verbo, etc.) de cada palavra do texto é essencial." Dadas as novas abordagens, isso não é mais essencial, então sugiro excluir essa afirmação ou reescrevê-la de forma mais modalizada.

INTRODUÇÃO

A introdução é clara, apresentando o contexto do PLN e da tarefa de etiquetagem de POS dentro do PLN. Também apresenta o estado da arte e acurácia da anotação em textos bem escritos (97%), comparando-o com a acurácia conseguida pelo corpus Dantestocks, que é user-generated-content (95%). Fico me perguntando se essa diferença de 2% é relevante o suficiente para ser considerada a principal justificativa do trabalho...

RELEVÂNCIA

A relevância do trabalho é justificada a partir das possibilidades de desdobramento da pesquisa, a saber:

1. subsidiar anotação de relações de dependência; e/ou
2. poder ser usado como ponto de partida para o desenvolvimento de ferramentas e novas teorias de linguagem, especialmente para gêneros e domínios específicos.

Mas qual a relevância da pesquisa em si? Qual a relevância em apresentar a distribuição das etiquetas morfossintáticas em um corpus? Entendo que essa relevância ficaria mais explícita se os autores tivessem feito uma comparação com a distribuição das mesmas etiquetas em outros corpora de outros gêneros (por exemplo, jornalístico ou wikipedia) ou de outros domínios. Se apresentassem um comparativo com gêneros e domínios mais "comuns" de textos bem-escritos, o leitor conseguiria ter um baseline para saber se a distribuição dos POS é parecida ou em quais etiquetas um corpus de tweets de ações se difere de outros corpora de linguagem mais usual. Isso sim poderia contribuir com a caracterização do gênero tweet e do domínio de mercado de ações.

LACUNA

Na seção de trabalhos relacionados, assim como na Introdução e no Resumo, os autores justificam a anotação de POS em tweets por conta de uma lacuna: outros trabalhos que pesquisaram tweets focaram em investigar fenômenos lexicais e ortográficos do gênero, em vez de POS. AFAIK, ambas as abordagens (distribuição de POS e análise de fenômenos lexicais/ortográficos) são dependentes de língua. Portanto, não faz muito sentido indicar trabalhos de espanhol e inglês que abordam fenômenos lexicais e ortográficos dos tweets como uma lacuna para uma abordagem de POS em português. Em outras palavras, o que o presente artigo faz (anotar POS e apresentar sua distribuição em um corpus em português) não preenche a lacuna de que os estudos anteriores só focaram em

fenômenos lexicais e ortográficos, pois aqueles estudos são de outras línguas e as duas abordagens são language-dependent.

MÉTODOS

A amostragem dos dados é adequada. Ressalte-se que foi anotada uma amostra de 4.048 tweets, que corresponde a 81.037 tokens, usando as guidelines de anotação da UD versão 2, que contém 17 PoS tags.

A seção de materiais e métodos está bem explicada e os experimentos podem ser reproduzidos, tanto a anotação de mais dados seguindo as mesmas diretrizes, quanto os cálculos e estatísticas da distribuição das etiquetas no corpus. Três informações muito importantes dessa seção não estão contidas no próprio artigo, mas estão indicadas no artigo e disponíveis na web:

1. o manual de anotação construído pelos autores e já divulgado em outra publicação (Relatório técnico do ICMC/USP em Di Felippo et al, 2022);
2. as diretrizes da Universal Dependencies v2, cujo link está em nota de rodapé; e
3. o corpus Dantestocks com as anotações, disponível para Download em um link do Drive.

Faço uma ressalva em relação a uma informação faltante nesta seção. Na página 7, onde se lê "the tweets were automatically tokenised and tagged (see Da Silva et al. (2021) for details), and then manually revised", entendo que os detalhes podem ser conferidos nessa outra publicação, porém é necessário indicar neste artigo pelo menos qual ferramenta (tagger) e qual modelo de pt foram usados para etiquetar automaticamente. Mesmo que não mencionem os detalhes, essa informação principal precisa estar explícia neste artigo.

Ademais, a seção "Materiais e métodos" indica algumas medidas estatísticas na Tabela 1, como quantidade absoluta de tokens e de tags (mínimo, máximo e média). Naquele momento, senti falta também de outras medidas tais como mediana, desvio padrão, quartil, outliers, etc. Seguindo a leitura, percebi que os autores explicitam algumas delas na seção de resultados e discussão. Para fins de sistematização e clareza, sugiro que os autores apresentem as técnicas e medidas estatísticas na seção "Materiais e métodos" e indiquem que algumas delas serão discutidas posteriormente na seção "Resultados e discussão". A tabela 1 também pode ser complementada com essas informações estruturadas.

RESULTADO

Dado o objetivo principal do trabalho, que é analisar a distribuição de POS tags no corpus Dantestocks, considero que a seção de "Resultados e discussão" seja o ponto-chave do artigo. As figuras apresentam os dados de forma resumida e os textos que se seguem às figuras apresentam exemplos e alguma informação a mais, além da informação da figura. Mas sinto que faltou aprofundamento em algumas discussões.

Por exemplo, quando se fala da quantidade de PROPN no Dantestocks e se justifica sua alta frequência por conta da quantidade de tickers (para nomes das empresas), eu gostaria de comparar essa frequência com a de outros corpora. Em outros textos, a frequência de PROPN costuma ser baixa ou igualmente alta? E se forem textos jornalísticos do domínio de ações, os quais também usam tickers para nomes de empresas? Isso depende mais do gênero tweet ou do domínio de mercado de ações?

Quando se fala de PRON, parece ter uma ambiguidade ou falta clareza. Nas p.10 e 11, a baixa frequência de PRON parece estar relacionada ao domínio do mercado de ações ("For PRON, its frequency might also be related to the domain, since stock market tweets tend to be very factual"). Já na p. 13, parece estar relacionada ao gênero tweet ("which might come up as a result from Twitter's message length limits"). Não está claro se a baixa frequência de PRON é devida ao gênero tweet ou ao domínio de mercado de ações. Esse tipo de avaliação/análise é necessária, por exemplo, para quem trabalha com resolução de anáfora em tweets ou posts de redes sociais.

A alta frequência da tag SYM é algo que carece de explicação, pois isso provavelmente se difere da distribuição de outros corpora. Nas Figuras 2 e 4, a distribuição de SYM é maior que de ADJ e ADV, que são tags extremamente frequentes em pt. O que justifica essa alta frequência? Tem a ver com o uso de emoticons para substituir interjeições? (Referência: p.10: "social media users apply duplicated punctuation, graphemic stretching, emoticons and smileys as marks of expressiveness") Ou tem alguma outra explicação?

No mesmo sentido, imagino que a quantidade de INTJ também deve ter uma diferença relevante em relação a outros domínios. Mesmo aparecendo pouco no Dantestocks, deve aparecer ainda menos em corpora de linguagem mais comum, certo? Para entender a "baixa frequência" de SCONJ, PRON, AUX and CCONJ, também seria interessante comparar com outros corpora.

Se os autores tivessem comparado essas estatísticas com as de outros corpora, principalmente os de linguagem mais corriqueira ou de textos bem-escritos, ficaria mais claro para o leitor o que está sendo considerada alta, média e baixa frequência para algumas etiquetas. Também ficaria mais claro se essa distribuição é específica de tweets de mercado de ações ou segue uma distribuição parecida com outros gêneros e domínios da língua portuguesa.

GENERALIDADES

De maneira geral, o artigo está bem escrito, claro, coerente e apresenta todas as seções relevantes para um artigo científico da Revista em questão. Elenco a seguir alguns pontos específicos a serem revisados:

1. (Obrigação) Nas referências, está faltando CHIARA (2022)
2. (Recomendação) Não me sinto confiante em avaliar a escrita em Língua Inglesa, mas algumas estruturas me chamaram atenção por serem incomuns em artigos científicos, por ex:
    1. p.13: Of the 17 tags [...] nine [...] can be found;
    2. p.13: With that, [...];
    3. p.2 "scientific articles" instead of "academic papers". Sugiro uma revisão mais detalhada por algum especialista.

1. (Obrigação) Existe uma citação literal na página 5 que deveria estar entre aspas e com indicação da fonte: https://universaldependencies.org/u/overview/syntax.html. O trecho é: "one word is the head of the sentence, dependent on a notional ROOT and all other words are dependent on another word in the sentence."
2. (Recomendação) Ao falar do modelo de Universal Dependencies, eu não entendi o motivo de usarem o exemplo da Figura 1.

3. Por que usar um exemplo em italiano em vez de um em português (ou inglês)?
4. Por que não usar um exemplo real do corpus Dantestocks ou pelo menos algum exemplo de sentença do domínio de mercado de ações?
5. O exemplo usado contém 2 sentenças (1. Chi mi fa un po' di compagnia a quest'ora? e 2. Mi sento sola.) e uma única root ligando ambas por uma deprel de parataxis. Acho que isso pode confundir o leitor, então sugiro usarem um exemplo com 1 única sentença.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Link to Preprint

The preprint was submitted to the TechRxiv repositor. However, the preprint will be publicly available at TechRxiv (https://www.techrxiv.org/) upon the successful completion of our moderation checks.

Research Preregistration and Standards

The authors reviewed the standards and any of them was relevant for the research application.

Data Accessibility Statement

The data that support the results of this study was derived from the following resources available in the public domain: https://drive.google.com/file/d/1MeG1KV3sPNAMJ1LlHim-YyR_bi8M_GXij/view?usp=sharing.

Disclosure of Funding Sources

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.

REFERENCES

ANCHIÊTA, R. T.; PARDO, T. A. S. Análise Semântica com base em AMR para o Português. **LinguaMÁTICA**, v. 14, n. 1, p. 33-48. 2022.

BARBERO, C. CQL Grammars for Lexical and Semantic Information Extraction for Portuguese and Italian. *In:* INTERNATIONAL CONFERENCE ON COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE, 15., 2022, Fortaleza/Brazil. **Anais** [...]. 2022. p. 376-386.

CABRAL, B.; SOUZA, M.; CLARO, D. B. PortNOIE: A neural framework for open information extraction for the Portuguese language. *In:* INTERNATIONAL CONFERENCE ON COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE, 15., 2022, Fortaleza/Brazil. **Anais** [...]. 2022. p. 243-255.

DA SILVA, E. H.; PARDO, T. A. S. ROMAN, N. T.; DI FELLIPO, A. Universal Dependencies for Tweets in Brazilian Portuguese: Tokenization and Part of Speech Tagging. In: ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL E COMPUTACIONAL (ENIAC), 18., 2021, Evento Online. **Anais** [...]. 2021. p. 434-445.

DA 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.

DE MARNEFFE, M. C.; MANNING, C. D.; NIVRE, J.; ZEMAN, D. Universal dependencies. **Computational Linguistics**, v. 47, n. 2, p. 255-308. 2021.

DE SOUZA, R. C. C.; LOPES, H. Portuguese POS Tagging Using BLSTM Without Handcrafted Features. *In:* IBEROAMERICAN CONGRESS ON PROGRESS IN PATTERN RECOGNITION, IMAGE ANALYSIS, COMPUTER VISION, AND APPLICATIONS, 24., 2019, Havana/Cuba. **Anais** [...]. 2019. p. 120-130.

DI FELIPPO, A.; POSTALI, C.; CEREGATTO, G.; GAZANA, L. S.; DA SILVA, E. H.; ROMAN, N. T.; PARDO, T. A. S. Descrição Preliminar do Corpus DANTEStocks: Diretrizes de Segmentação para Anotação segundo Universal Dependencies. *In:* SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA, 13., 2021, Evento Online. **Anais** [...]. 2021. p. 335-343.

DI FELIPPO, A.; POSTALI, C.; CEREGATTO, G.; GAZANA, L. S.; ROMAN, N. T. **Diretrizes de anotação de POS Tags em tweets do mercado financeiro:** Orientações para anotação em língua portuguesa segundo a abordagem Universal Dependencies (UD**)**. 2022. Relatório Técnico do ICMC - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2022.

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). 2022. Relatório Técnico do ICMC - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2022.

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, 2019, Atlanta/USA. **Anais** [...]. 2019. p. 359-369.

FONSECA, E. R.; ROSA, J. L. G.; ALUÍSIO, S. M. Evaluating word embeddings and a revised corpus for part-of-speech tagging in Portuguese. **Journal of the Brazilian Computer Society**, v. 21, n. 2, p. 1-14. 2015.

LIU, Y.; ZHU, Y.; CHE, W.; QIN, B.; SCHNEIDER, N.; SMITH, N. A. Parsing tweets into universal dependencies. *In:* ANNUAL CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES, 16., 2018, New Orleans/USA. **Anais** [...]. 2018. p. 965-975.

LYDDY, F.; FARINA, F.; HANNEY, J.; FARRELL, L.; O'NEILL, N. K. An analysis of language in university students' text messages. **Journal of Computer-Mediated Communication**, v. 19, n. 3, p. 546-561. 2014.

MACHADO, M. T.; PARDO, T. A. S.; RUIZ, E. E. S.; DI FELIPPO, A.; VARGAS, F. Implicit opinion aspect clues in Portuguese texts: analysis and categorization. *In:* INTERNATIONAL CONFERENCE ON COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE, 15., 2022, Fortaleza/Brazil. **Anais** [...]. 2022. p. 68-78.

MELERO, M.; COSTA-JUSSÀ, M. R.; DOMINGO, J.; MARQUINA, M.; QUIXAL, M. Holaaa!! writin like u talk is kewl but kinda hard 4 nlp. *In:* INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 8., 2012, Istanbul/Turkey. **Anais** [...]. 2012. p. 3794-3800.

NIVRE, J.; DE MARNEFFE, M.; GINTER, F.; HAJIC, J.; MANNING, C. D.; PYYSALO, S.; TYERS, S. S. F. M.; ZEMAN, D. Universal dependencies v2: An evergrowing multilingual treebank collection. *In:* INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 12., 2020, Marseille/France. **Anais** [...], 2020. p. 4034-4043.

PARDO, T. A. S.; DURAN, M. S.; LOPES, L.; DI FELIPPO, A.; ROMAN, N. T.; NUNES, M. G. V. Porttinari - a Large Multi-genre Treebank for Brazilian Portuguese. *In:* SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA, 13., 2021, Evento Online. **Anais** [...]. 2021. p. 1-10.

SANGUINETTI, M.; BOSCO, C.; CASSIDY, L.; ÇETINOGLU, Ö.; CIGNARELLA, A. T.; LYNN, T.; REHBEIN, I.; RUPPENHOFER, J.; SEDDAH, D.; ZELDES, A. Treebanking user-generated content: A proposal for a unified representation in universal dependencies. *In:* INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 12., 2020, Marseille/France. **Anais** [...], 2020. p. 5240-5250.

SANGUINETTI, M.; BOSCO, C.; CASSIDY, L.; ÇETINOGLU, Ö.; CIGNARELLA, A. T.; LYNN, T., REHBEIN, I.; RUPPENHOFER, J.; SEDDAH, D.; ZELDES, A. Treebanking user-generated content: a UD based overview of guidelines, corpora and unified recommendations. **Language Resources & Evaluation**. 2022.

SANGUINETTI, M.; BOSCO, C.; LAVELLI, A.; MAZZEI, A.; ANTONELLI, O.; TAMBURINI, F. PoSTWITA-UD: an Italian Twitter treebank in Universal Dependencies. *In:* INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 11., 2018, Miyazaki/Japan. **Anais** [...], 2018. p. 1768-1775.

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/Belgium. **Proceeding** [...], 2018, p. 197–207.

SENO, E.; CASELI, H.; INÁCIO, M.; ANCHIÊTA, R.; RAMISCH, R. XPTA: um parser AMR para o Português baseado em uma abordagem entre línguas. **LinguaMÁTICA**, v. 14, n. 1, p. 49-68. 2022.

VOSKAKI, R.; TZIAFA, E.; IOANNIDOU, K. Description of predicative nouns in a modern greek financial corpus. In: INTERNATIONAL SYMPOSIUM ON THEORETICAL AND APPLIED LINGUISTICS, 21. 2016, New Orleans/USA. **Anais** [...], 2016. p. 488-503.

WU, S.; DREDZE, M. Beto, Bentz, Becas: The Surprising Cross-lingual Effectiveness of BERT. *In:* CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING AND THE INTERNATIONAL JOINT CONFERENCE ON NATURAL LANGUAGE PROCESSING, 2019, Hong Kong/China. **Anais** [...], 2019. p. 833-844.