Named Entities in Stock Market Tweets: A Fine-Grained and Linguistically-Motivated Annotation

Laís Piai[1,2], **Ariani Di-Felippo**[1,2], **Norton Trevisan Roman**[3]

[1] Interinstitutional Center for Computational Linguistics (NILC), São Carlos – Brazil

[2]Graduate Program in Linguistics (PPGL/UFSCar), São Carlos – Brazil

[3]School of Arts, Sciences and Humanities (EACH/USP), São Paulo – Brazil

lais.piai@estudante.ufscar.br, ariani@ufscar.br, norton@usp.br

Abstract. *This work provides a second look at the Named Entity annotation of DANTEStocks – a corpus of stock market tweets in Portuguese – offering an in-depth analysis of the linguistic decisions involved in creating a gold-standard annotation tailored to this genre and domain. Our methodology builds on the guidelines of the Second HAREM evaluation, extending and reinterpreting them to the adopted genre and domain. The article furnishes then an analysis of the linguistic phenomena that challenge to this task, proposes specific strategies for entity delimitation and classification, and presents a linguistic characterization of the corpus based on the class distribution that resulted from the annotation.*

1. Introduction

Twitter/X is a text-based social media that enables users to publicly share brief updates, opinions and news, playing a pivotal role in shaping modern discourse. Since its inception in 2006, it became a global powerhouse for real-time communication, and thus a significant source of high-valued information for diverse domains and Natural Language Processing (NLP) applications, particularly regarding sentiment analysis, emotion classification and opinion mining [Derczynski et al. 2016]. Their often non-standard content (i.e. short, noisy, and colloquial nature) makes them challenging for traditional NLP tools, requiring domain adaptation and annotated corpora for training and evaluating them. These requirements, allied to the global interest in their content, has fostered the construction of many annotated corpora in different languages worldwide [Sanguinetti et al. 2023].

A pioneer multi-layered resource of gold-standard annotations in tweets/posts written in (Brazilian) Portuguese is DANTEStocks, which comprises 4,048 tweets[1] from the stock market domain [Di Felippo and Roman 2025]. This domain is especially interesting given the alleged correlation between sentiment in tweets and stock market returns, which has been used to predict market direction [Deveikyte et al. 2022]. In DANTEStocks, tweets were preserved in their original form, i.e. they were not segmented into smaller linguistic units (e.g. sentences or phrases), and no normalization procedure was applied. The multiple layers of annotation include emotion [da Silva et al. 2020], part-of-speech (PoS) tags [Di Felippo et al. 2024c] and dependency relations [Di Felippo et al. 2024a] following the Universal Dependencies (UD) model [de Marneffe et al. 2021], with the UD-annotation layers already enabling

[1]Still under the 140-character limit.

the development of various NLP tools (e.g. [Silva et al. 2021, Di Felippo et al. 2024a, Di Felippo et al. 2024b]).

The frequent presence of named entities (NEs) such as person, location, and organization names in tweets reveals the importance of named entity recognition (NER) tools for identifying and classifying existing entities into predefined categories for further semantic interpretation [Liu et al. 2011]. In this regard, DANTEStocks has already been augmented with a stand-off layer of NEs (cf. [Zerbinati et al. 2024]), making it a suitable resource to support the development and the exploration of existing NER methods. In this layer, entities were manually annotated according to the taxonomy adopted in the Second HAREM – the Golden Collection of texts for the second evaluation campaign on NER in Portuguese [Mota and Santos 2008], being limited to the highest level of this category.

However important a resource, this limitation hinders a more in-depth analysis of existing entities, which might be of great interest when using the corpus in practical studies. Our work seeks then to fill in this gap by moving beyond category-level labelling to include entity types, enabling more fine-grained distinctions within each category. Moreover, our independent classification of these tweets helped clarify some parts an identify not linguistically accurate decisions made in the original annotation guidelines by [Zerbinati and Roman 2023], along with parts where tweet annotation should astray from the Second HAREN's guidelines. Our contribution then is not only to offer a more fine-grained classification to DANTEStocks' existing NE annotation, but also to establish guidelines for handling UGC and domain phenomena specific to stock market tweets[2].

The remainder of this article is structured as follows. Section 2 provides an overview of related work, whereas Section 3 presents the taxonomy and methodology used in this task. Section 4, in turn, describes the linguistic phenomena in the corpus that affect NE annotation, and the strategies employed to address them. In Section 5 we characterise the resulting corpus, furnishing statistics about entities and their distribution. Finally, Section 6 presents our concluding remarks.

2. Related work

Several studies have produced Twitter-based manually annotated corpora for NER in English (e.g. [Ritter et al. 2011, Finin et al. 2010, Liu et al. 2011, Derczynski et al. 2016]). These corpora share common characteristics, including a single-label approach, the use of standard categories such as PLO (i.e. PERSON, LOCATION, and ORGANIZATION), and annotation by a single expert. A recurring limitation, however, is the inconsistent treatment or outright exclusion of platform-specific elements like at-mentions (@usernames), hashtags, and URLs.

In the realm of Portuguese, Twitter-NER [Peres et al. 2017] is perhaps the earliest corpus of tweets to feature manual NE annotation. It is a general-domain resource, containing 3,968 tweets with 935 entities categorized according to PLO. The authors adhered to the guidelines proposed by [Finin et al. 2010], refraining, however, from describing how lexical idiosyncrasies of tweets were handled during the NE annotation process. Another initiative, as already mentioned, can be found within the DANTEStocks Project, in the form of a stand-off NE layer [Zerbinati et al. 2024], aiming to explore correlations between named entities and other annotated linguistic layers in the corpus. In

[2]Both resourced are freely available for download at [ANONYMISED].

this work, a single expert annotated the corpus using the Second HAREM's 10 top-level categories and guidelines [Mota and Santos 2008]. The annotation followed the BIOES scheme [Jurafsky and Martin 2025] and was context-based, though limited to one category per entity, unlike HAREM's multi-label approach.

An analysis of this application of HAREM's guidelines to DANTEStocks revealed certain annotation challenges. A key issue lies in the reliance on capitalization as the primary heuristic for NE identification. Within standard written language, this criterion is generally effective, as most entities are expressed through proper nouns. However, in social media posts, it often results in missed entities, since users in this genre frequently deviate from standard capitalization practices [Finin et al. 2010]. Another issue present in HAREM is the exclusion of URLs as NEs. Since they are very frequent in tweets, not annotating URLs may result in the loss of critical referential information, undermining the ability of NLP systems to trace sources or disambiguate entities [Liu et al. 2018].

As a further limitation, in analysing DANTEStock's annotation we observed an inconsistent handling of certain UGC phenomena (such as word truncation) and domain-specific features (e.g. informal abbreviations). A final critical observation relates to the limitation imposed by using only the 10 top-level categories of HAREM's taxonomy (i.e. ABSTRACTION, EVENT, THING, LOCATION, PRODUCTION, ORGANIZATION, PERSON, TIME, VALUE, and OTHER). While this approach enhances annotation reliability and supports the development of interoperable NLP tools, it may also oversimplify the annotation process, hindering the capture of domain-specific nuances and fine-grained entity types that are crucial for precise financial text analysis [Sekine and Nobata 2004].

3. Methodology and Taxonomy

NER was conceived as a task that involves identifying minimal, semantically stable text units that can be reliably recognized and reused across contexts and classifying them into predefined categories. In this work, the task was performed by a single annotator (degree in Linguistics), primarily based on context. Expert input and reliable online sources were also used to handle the informal language and domain-specific content of financial tweets. For instance, interpreting "CS" (abbreviation of "Credit Suisse") in "CS indicating PETR4 x PETR3" as ORGANIZATION-*company* would be difficult without extra knowledge. We took a single-label approach, where polysemous entities are assigned a unique category-type based on their contextual meaning. For instance, "vale" is categorized as ORGANIZATION-*company* in "A vale foi pro vale" ("vale went down the valley") while in "Ah minhas vale e petr4" ("Ah my vale e petr4") it is labeled as THING-*ticker*.

Following [Zerbinati et al. 2024], we delimited NEs with BIOES labels [Jurafsky and Martin 2025], whereby multi-token entities are labeled with B (beginning), I (inside), and E (end) tags whereas single-token entities are labeled as S (single). Tokens that do not correspond to entities were left unannotated, being implicitly labeled O (outside). This study builds upon version 1.0 of DANTEStocks [Di Felippo et al. 2024c], which already comes with UD PoS annotation, comprising a total of 4,048 tweets and 84,396 tokens. Our annotation was then added to this corpus' CoNLL-U file in its fifth column (cf. Figure 1) As for NE classes, we adopted the Second HAREM's 10 top-level categories along with other 43 types, facilitating comparison with related studies. This inventory of types was nevertheless

extended with 4 additional types (see Section 4) to better capture the stock market's idiosyncrasies. Table 1 presents the (9) categories and (36) types that effectively occurred in the corpus, highlighting the four newly introduced types in blue cells and providing original examples of annotated NEs.

Table 1. Categories and types of named entities in the DANTEStocks corpus.

Category	Type	Exemples
ABSTRACTION	Condition	VACA LOUCA, mal da vaca louca
	Discipline	Análise #Ichimoku, Daytrade, Streaddle
	Idea	Mercado, mão invisível
	Name	Graça
EVENT	Ephemerides	#Lavajato, Pasadenagate
	Occurrence	Reunião do Conselho de Administração, ago/e
	Organized	Copa, carnaval
THING	Class	candlesticks, Valemax, Boeing
	Memberclass	PDF
	Object	OCO, OCOI, shooting star
	Certificate	ADR, MBA, Eletrobrás ON, bbas-nm
	Indicator	IBOV, S&P500, Ibovespa/VPa,
	Ticker	ESTC3, PSSA3, VALE5, PETR4
LOCATION	Physical	Reserva da Cantareira, Bacia de Santos
	Human	Brasil, São Paulo, Refinaria Pasadena
	Virtual	@JornalOGlobo, http://t.co/LJluyesRk5
PRODUCTION	Plan	Plano de demissao voluntária, MP 627
	Reproduced	Formulário 20-F, Relatório De Análise Gerencial
ORGANIZATION	Administration	Conselho de Ministros, C.a., Conselho da Petrobras
	Company	Cielo, Bando do Brasil, Bovespa
	Institution	PT, Organização Mundial de Saúde, PMDB
PERSON	Role	CEO da Vale, Vice do Bradesco, Presidente da República
	GroupRole	ministros da Justiça
	Individual	Ministra Rosa Weber, Graça Foster
	GroupInd	governo Dilma, Gov FHC
	Member	ex-Itaú, ex-TAM
	GroupMember	BTG, Povo Brasileiro
	People	China, Rússia
	User	@Live_Trade, @PaiRico @frfontanella
TIME	Duration	20-30 anos, 7 dias
	Frequency	Todos os dias, poucas vezes
	Generic	hoje
	CalendarTime	28/04/2014, 18/03/2014 15:23, INTRADAY
VALUE	Classification	15ª, oitavo
	Quantity	+ 3,64 %, 70,2 milhões, 4
	Money	R$ 52,38, R$ 14,81, dez reais

4. Annotation Guidelines

Although HAREM's guidelines [Mota and Santos 2008] provided the foundation to this work, adaptations were made to reflect orthographic and lexical variations along with tokenization issues found in the corpus. These are discussed further in what follows.

4.1. Principles of identification and delimitation

Orthographic and Lexical Variation. To handle orthographic and lexical variation in financial tweets, we apply a context-driven approach suited to the noisy and creative nature of UGC. Orthographic variants such as "Petrobras", "Petrobrás", "petro-

bras", "petrobrás", "PETROBRAS", and "PETROBRÁS", for example, are annotated uniformly. Humorous or sentiment-laden lexical variants like "PETROFUMO", "Petrobomba", "PeTebrás" are annotated when context confirms reference to Petrobras, as are informal shortenings like "Petro" and truncated instances like "Petr".

Truncated words. They occur when an entity is truncated due to platform-imposed character limits on tweets, often indicated by ellipses at the end. These NEs should be annotated when their category-type can be determined based on context, another tweet, expert judgment, or external sources. In all instances, ellipses are not considered an integral part of the entity. For instance, in "drop of nearly 5% at the moment in Bove...", the context indicates that "Bove" refers to "Bovespa", being annotated with S-ORGANIZATION-*company*. Truncated URLs, as far as they are recognized so, are S-LOCATION-*virtual*, as is the case of "htt" in "Learn more at htt...". When truncated hashtags or Twitter handles contain only a visible prefix ("#..." and "@...") and the entity cannot be confidently identified, they are left unannotated. Value changes (e.g. "+2,10%") where only the symbol is present (without an accompanying number), as in "+...", are not annotated.

Temporal and Quantifier Expressions. Following the guideline to annotate the smallest expression (text span) that preserves the core named entity, scalar modifiers, such as "mais de" in "mais de (5 meses/R$ 1 bilhão)" ("over (5 months/R$ 1 billion)") were not annotated. This is acceptable in NER because, however adding some imprecision that affects the quantitative interpretation of the expression, the modifier does not alter the entity's referential identity. We also excluded the phrase "dia de" in cases such as "dia de ontem" ("the day of yesterday") because "ontem" already convey a complete temporal reference. On the other hand, relational temporal modifiers such as "antes" ("before") in expressions like "1 ano antes" ("1 year before") were regarded as integral parts of the entity because it forms a cohesive temporal unit. Thus, the full expression was annotated as one entity.

Value Changes. Based on the UD definition of a syntactic word, changes in value (e.g. "-0.90%") were split into three tokens: symbol, numeric value, and percentage sign. Treated as a multiword entity, the index is labeled in the BIOES scheme as follows: the token "-" is marked as B, "0.90" as I, and "%" as E. Truncated indices without a percentage sign (e.g. "+1.60...") have their numeric value labeled as E. This reflects the multi-token structure of the entity and ensures accurate segmentation and classification.

Contraction. Contractions were split following UD tokenization guidelines (e.g. "neste" $\rightarrow$ "em" (ADP) "este" (DET)). If a contraction introduces a temporal expression, as shown in the top part of Figure 1, it remains unannotated, and DET is labeled as B. When a contraction occurs within a NE (e.g. "do" in the bottom of Figure 1), both contraction and its components ("de" and "o") receive the I tag.

4.2. Classification Recommendations

To enhance annotation granularity, THING types (*object*, *class*, *memberclass*, *substance* and *other*) were expanded to include *ticker*, *indicator*, and *certificate*. *Ticker* refers to alphanumeric codes used in stock negotiation, annotated uniformly regardless of hash or cashtag[3] usage (e.g., "PETR4", "#PETR4", "$PETR4"). *Indicator* includes stock indices (e.g., "Ibovespa", "S&P 500") and financial metrics such as "P/L" (i.e. "Preço/Lucro")

[3]A *cashtag* is a Twitter convention where a dollar sign precedes a stock ticker (e.g. $Petr3), linking tweets to discussions about that financial asset. It functions like a hashtag, but cashtags are finance-focused.

12-13	neste	_	_	_
12	em	em	ADP	_
13	este	de	DET	ENTITY =B-TIME-CALENDAR
14	mês	o	NOUN	ENTITY =E-TIME-CALENDAR
1	Banco	Banco	PROPN	ENTITY =B-ORGANIZATION-COMPANY
2-3	do	_	_	ENTITY =I-ORGANIZATION-COMPANY
2	de	de	ADP	ENTITY =I-ORGANIZATION-COMPANY
3	o	o	DET	ENTITY =I-ORGANIZATION-COMPANY
4	Brasil	Brasil	PROPN	ENTITY =E-ORGANIZATION-COMPANY

Figure 1. Annotation of named entities containing contractions.

and "Price-to-Earnings ratio". *Certificate* encompasses share types expressed without tickers, often combining company names with share classes (e.g., "Eletrobrás ON", "BBAS-NM"), listing segments (e.g., "ADR"), and financial credentials (e.g., "MBA").

According to the Second HAREM, administrative bodies, such as the board of directors, committee etc., are annotated with the same classification as their parent organization (i.e. ORGANIZATION-*company* or ORGANIZATION-*institution*). In the financial domain, distinguishing administrative bodies as ORG-administration is particularly interesting because these entities often have specific roles that directly impact investor decisions, regulatory compliance, and corporate governance. Thus, this decision allows NLP systems to capture who made the decision, which is crucial in financial analysis.

Differing from HAREM, entities such as "AGO" ("Assembleia Geral Ordinária" – "ordinary general meeting") are not treated as organizations because they typically represent specific, time-framed events rather than lasting institutional structures, not running continuously like organizations do. As such, they were annotated as EVENT-*occurrence*. URLs were annotated as LOCATION-*virtual* because they often serve as gateways to external content. In financial tweets, URLs may link to news articles, company announcements, or market reports – resources that help clarify the meaning of surrounding entities or events. Annotating URLs as LOCATION-*virtual* acknowledges their role as digital locations and ensures that this layer of meaning is preserved in corpus analysis.

Platform devices such as hashtags often serve as references to entities of different categories, such as companies (e.g. "#Petrobras"), financial instruments (e.g. "#Ibov"), people (e.g. "#Dilma"), concepts (e.g. "#daytrade") etc. They should then be analyzed within context and annotated according to the entity they represent. As for at-mentions, although current approaches block these out (e.g. [Rowe et al. 2013]) or classify them as person (e.g. [Ritter et al. 2011]), we decided to treat them as entities of any potential class. For example, if "@petrobras" refers to the Petrobras company, it should be labeled as ORGANIZATION-*company*. When the mention points to a communication channel like "Revista Epoca", it is labeled as LOCATION-*virtual*. Mentions that do not clearly correspond to organizations or media outlets should be annotated as PERSON-*user*.

5. Named Entity Characterization of DANTEStocks

As it turned out, NEs can be found in all of the 4,048 tweets that build the corpus. In total, 24,825 tokens were found to pertain to some entity (recall that some entities span

over multiple tokens), meaning that around 29% of all 84,396 tokens of the corpus belong to NEs. Among the 20,092 annotated NEs, single-token entities are the majority, with 16,898 instances compared to 3,194 multiword entities. This trend holds true for eight of the nine categories, except for PRODUCTION, where the distribution is more balanced.

Category-type distribution is highly unbalanced, following a pattern typical of real-world data, where a few entities are numerous while others appear rarely (Figure 2). As expected, there is a predominance of the THING category, accounting for nearly 40% of all entities. The VALUE category comes next, with 20.6% of the entities. In the sequence come LOCATION, ORGANIZATION, and TIME. Among the 10 original categories, only 9 appear in the corpus, as OTHER, the leftover class, was found unnecessary.

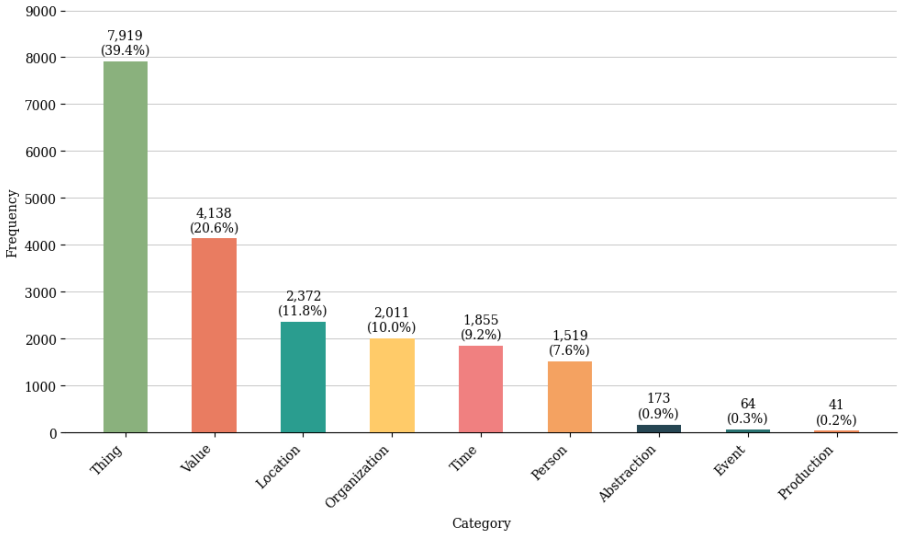

Figure 2. Frequency of entities per category.

The THING category is mainly defined by domain-specific types, led by *ticker*, with 7,076 occurrences, followed distantly by *indicator* (410) and *certificate* (332). The remaining types in this category – *class*, *classmember* and *object* – occur with minimal or single-digit frequencies and primarily represent general, rather than domain-specific, entities. As shown in the Figure 3, the most frequent type after *ticker* is *money* (VALUE) with 2,634 occurrences. This aligns with the informative nature of the tweets, which often mention *tickers* alongside their respective values. The third most frequent type, *virtual* (LOCATION), typically refers to URLs, underscoring the domain's online nature. In contrast, abstract types like *idea* (ABSTRACTION) and *duration* (TIME) are rare.

When it comes to the distribution of NEs across tweets (Figure 4), we see that the amount of NEs found in a single tweet ranges from a single entity (found in 228 tweets, having only a ticker as NE) up to 19 entities (found in 1 tweet), peaking at 3 entities, found in 794 different tweets. Tweets containing 3 entities typically follow the pattern: cashtag, company name, and stock ticker, as in "$PETR3 - Petrobras (petr)".

6. Final Remarks and Future Work

Our annotation was based on linguistically motivated decisions and on a broader notion of *entity*, encompassing not only anything that can be referred to with a proper name, but also temporal and numerical expressions, and nominal domain-specific concepts.This

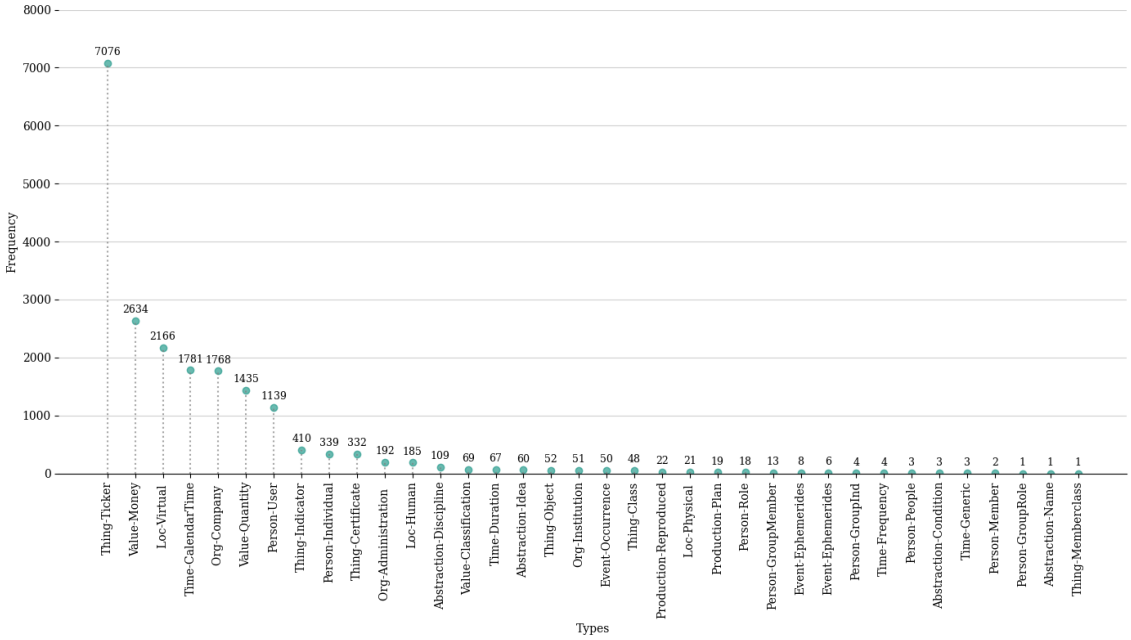

Figure 3. Frequency of each type.

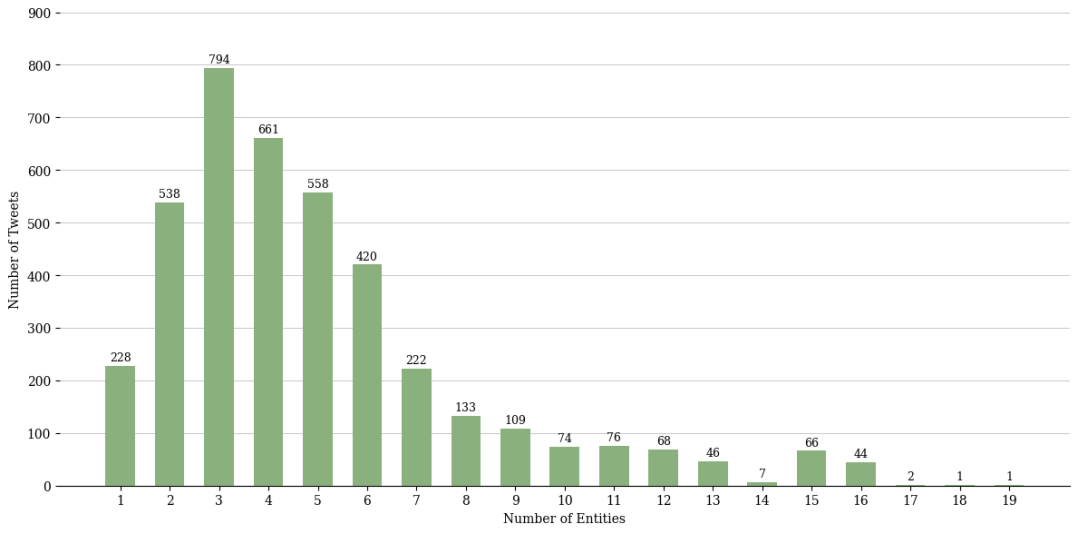

Figure 4. Entity count per tweet and corresponding tweet frequency.

broader view led to the creation of fine-grained types (i.e *ticker*, *indicator*, *certificate* and *user*) that enhance corpus characterization by addressing the linguistic and informational needs of the financial domain. As future research, a comparison of our approach with [Zerbinati et al. 2024] can highlight their similarities and differences.

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 S˜ao Paulo Research Foundation (FAPESP 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, with in the scope of PPI-SOFTEX, coordinated by Softex and published as Residence in TIC13, DOU01245.010222/2022-44.

References

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

de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Derczynski, L., Bontcheva, K., and Roberts, I. (2016). Broad Twitter corpus: A diverse named entity recognition resource. In Matsumoto, Y. and Prasad, R., editors, *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1169–1179, Osaka, Japan. The COLING 2016 Organizing Committee.

Deveikyte, J., Geman, H., Piccari, C., and Provetti, A. (2022). A sentiment analysis approach to the prediction of market volatility. *Frontiers in Artificial Intelligence*, 5.

Di Felippo, A., Nunes, M. d. G. V., and Barbosa, B. K. d. S. (2024a). A dependency treebank of tweets in Brazilian Portuguese: Syntactic annotation issues and approach. In *Proceedings of the XV Symposium in Information and Human Language Technology*, pages 192–201, Porto Alegre, RS, Brasil. SBC.

Di Felippo, A., Roman, N., Barbosa, B., and Pardo, T. (2024b). Genipapo - a multigenre dependency parser for Brazilian Portuguese. In *Proceedings of the XV Symposium in Information and Human Language Technology*, pages 257–266, Porto Alegre, RS, Brasil. SBC.

Di Felippo, A., Roman, N., Pardo, T., and Moura, L. (2024c). The DANTEStocks corpus: an analysis of the distribution of Universal Dependencies-based Part-of-Speech tags. *Revista da Abralin*, 22:493–544.

Di Felippo, A. and Roman, N. T. (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.

Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., and Dredze, M. (2010). Annotating named entities in twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 80–88.

Jurafsky, D. and Martin, J. H. (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall, 3rd (draft) edition.

Liu, X., Zhang, S., Wei, F., and Zhou, M. (2011). Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 359–367, Portland, Oregon, USA. Association for Computational Linguistics.

Liu, Y., Zhu, Y., Che, W., Qin, B., Schneider, N., and Smith, N. A. (2018). Parsing tweets into Universal Dependencies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 965–975, New Orleans, Louisiana. Association for Computational Linguistics.

Mota, C. and Santos, D. (2008). *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca.

Peres, R., Esteves, D., and Maheshwari, G. (2017). Bidirectional lstm with a context input window for named entity recognition in tweets. In *Proceedings of the 9th Knowledge Capture Conference*, K-CAP '17, New York, NY, USA. Association for Computing Machinery.

Ritter, A., Clark, S., Mausam, and Etzioni, O. (2011). Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Rowe, M., Stankovic, M., Dadzie, A.-S., Nunes, B. P., and Cano, A. E. (2013). Making sense of microposts (#msm2013): Big things come in small packages. In *Proceedings of the 22nd International Conference Companion on World Wide Web*. ACM. Workshop on Making Sense of Microposts.

Sanguinetti, M., Bosco, C., Cassidy, L., and et al. (2023). Treebanking user-generated content: a ud based overview of guidelines, corpora and unified recommendations. *Language Resources & Evaluation*, 57:493–544.

Sekine, S. and Nobata, C. (2004). Definition, dictionaries and tagger for extended named entity hierarchy. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Silva, E. H., Pardo, T. A. S., Roman, N. T., and Di-Felippo, A. (2021). Universal Dependencies for tweets in Brazilian Portuguese: tokenization and part of speech tagging. In *Proceedings of the 18th National Meeting on Artificial and Computational Intelligence*, pages 1–12.

Zerbinati, M. M. and Roman, N. T. (2023). Manual de anotação de entidades nomeadas do dantestocks utilizando categorias do segundo harem. Technical Report PPgSI-000/2023, PPgSI-EACH-USP, São Paulo, SP.

Zerbinati, M. M., Roman, N. T., and Di-Felippo, A. (2024). A corpus of stock market tweets annotated with named entities. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 276–284, Santiago de Compostela, Galicia/Espanha. Association for Computational Linguistics.