A Corpus of Stock Market Tweets Annotated with Named Entities

Michel Monteiro Zerbinati
EACH/USP
São Paulo – SP – Brazil
michel.zerbinati@usp.br

Norton Trevisan Roman
EACH/USP
São Paulo – SP – Brazil
norton@usp.br

Ariani Di Felippo
DLL/UFSCar
São Carlos - SP - Brazil
ariani@ufscar.br

Abstract

In this work, we present a corpus of stock market tweets written in Brazilian Portuguese and annotated with named entities according to HAREM's taxonomy. The corpus consists of 4,048 tweets and was originally built for research on emotion classification, being already annotated with it. By identifying the named entities present in the corpus, we intend it to enable new studies regarding possible correlations between named entities and emotions, along with other research on how such entities are used in this domain and linguistic genre. The annotation was manually carried out by one of the researchers and, out of the 84.397 tokens present in the corpus, 23.453 were annotated with named entities.

1 Introduction

The term Named Entity (NE) apparently emerged at the 6th Message Understanding Conference (MUC), as a task involving the identification of PERSONS, ORGANIZATIONS and LOCATIONS (denominated ENAMEX – Entity Name Expression), as well as PERCENTEAGE and MONEY (called NUMEX – Numerical Expression) (Grishman and Sundheim, 1996). Later on, LOCATION (CITY, STATE, and COUNTRY) and PERSONS (POLITICIAN, BUSINESS PERSON, and ARTIST) were further divided into subtypes (*cf.* (Fleischman, 2001; Fleischman and Hovy, 2002)), making the classifications more specialized. PERSONS, ORGANIZATIONS and LOCATIONS were the initial focus in MUC because they are well-defined and very frequent classes, which is essential for the semantic analysis of textual content.

The term's origins can, however, be traced back to the philosophical work by (Kripke, 1982), where the term "Named" was applied to entities for which a rigid designator represents only one referent, meaning that each NE represents the same referent in every possible world. Within this setup, "the automotive company founded by Henry Ford in 1903" could be referred to as "Ford" or "Ford Motor Company" in whatever context (Nadeau and Sekine, 2007). Currently, NEs are typically represented by proper names and may include certain natural terms such as biological species and substances. However, its definition may be relaxed in some cases for practical reasons (Nadeau and Sekine, 2007). For example, the entity "June" might refer to a month of an indefinite year, rather than a rigid designator like "June 2020".

The interest in Named Entity Recognition (NER) has grown in recent years, leading to the emergence of new studies on this topic. In particular, regarding NER applied in the financial domain, (Marcińczuk and Piasecki, 2015) utilized Hidden Markov Model (HMM) to recognize and classify entities of the Person and Organization classes in stock market reports in Polish, achieving 64% precision for the PERSON class and 78% for the ORGANIZATION class. Similarly, (Wang et al., 2014) employed a domain dictionary to recognize stock names in Chinese financial documents, followed by a Conditional Random Fields (CRF) classifier to classify the Organization entity, achieving 91% precision.

Along the same lines, (Khaing et al., 2019) proposed a model for detecting ORGANIZATION entities using rule-based and dictionary-based techniques, such as the names of companies listed in the S&P 500. Posts on Twitter[1] was also the subject of research by (Chen et al., 2018), who proposed a taxonomy of numerical classes for financial market tweets (e.g., VALUE, QUOTE, SELLING PRICE, BUYING PRICE, STOP LOSS, RELATIVE PERCENTAGE, ABSOLUTE PERCENTAGE), conducting experiments with Convolutional Neural Network (CNN), Long Short-Term Memory Networks (LSTM) and Bidirectional LSTMs (Bi-

[1] Now X.

LSTM), achieving better results with CNN, with a precision of 67.61%.

In Portuguese, two of the most commonly used annotated corpora for NER are the First and the Second HAREM. Some of the studies that utilized these corpora and the categories and classes employed by them include (do Amaral and Vieira, 2013), who applied the taxonomy and corpus from the Second HAREM, achieving an 48.43% F-score (F1), with Conditional Random Fields (CRF). The first HAREM was in turn used in (Souza et al., 2020), with a 78.67% F-score.

Despite these efforts, there still seems to be no example of a corpus comprising tweets from the stock market written in Portuguese and annotated with NER. This is the gap we intend to help fill in. To this end, we build on a pre-existing corpus of stock market tweets (Vieira da Silva et al., 2020), which was initially annotated with emotions according to Plutchik's wheel (Plutchik and Kellerman, 1986). Later, this corpus was enriched with morphosyntactic information (in the form of Part-of-Speech (PoS) tags, according to the Universal Dependencies model[2] (de Marneffe et al., 2021)), resulting in the DANTEStocks corpus (Di-Felippo et al., 2021).

We have then added an extra standoff layer to DANTEStocks[3], where we identify and classify NEs according to the taxonomy of categories defined and employed in the annotation of the Second HAREM (Mota and Santos, 2008). This taxonomy comprises ten categories, namely ABSTRACTION, EVENT, OBJECT, PLACE, PRODUCTION, ORGANIZATION, PERSON, TIME, VALUE and OTHER[4].

With the original corpus already annotated with emotions, DANTEStocks allows for a mapping between morphosyntactic and emotional information. Our contribution to the field, with this extra layer, is then to allow for a connection, however limited, to be established between syntax (limited to morphosyntax), pragmatics (limited to the tweets' emotional content) and semantics (limited to NEs). To the best of our knowledge, this is the first corpus to allow for such a link to be made.

Regarding the two essential aspects of HAREM concerning NEs (*cf.* (Mota and Santos, 2008)), we have fully adhered to the first aspect, which demands the identification and classification of a given expression as a NE to be exclusively based on its context, without being lexically restricted to any specific attributes associated with it in other linguistic resources such as dictionaries, almanacs, or ontologies. We diverge, however, from the second aspect, which allows for the association of multiple categories with a NE. Hence, in this work we have assigned only one category to each NE.

The main motivation for choosing HAREM's taxonomy was its status as a benchmark which is widely adopted by various studies on NER in Portuguese (Mota and Santos, 2008), thereby allowing for a better comparison between existing studies and ours. The rest 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 to build it, whereas Section 4 presents an analysis and a characterisation of the resulting corpus. Finally, our final remarks are presented in Section 5.

2 Related work

Many are the examples currently available of corpora annotated with NEs. One of them is GENIA (Kim et al., 2003), created to support the development and evaluation of information extraction and text mining systems for the domain of molecular biology, which contains 97,876 standardized entities across 36 categories, including PROTEIN and DNA, with 490,941 tokens in total.

Another corpus, CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003), was released as a part of CoNLL-2003 shared task: language-independent named entity recognition. The data consists of eight files covering two languages: English and German. The English data set was taken from the Reuters Corpus, consisting of news published between August 1996 and August 1997. The corpus focus on four classes of NEs: PERSONS, LOCATIONS, ORGANIZATIONS and MISCELLANEOUS (entities that do not belong any of the other groups).

Still in the realm of general domain corpora, OntoNotes 5.0 (Ralph Weischedel, 2013) stands out as a large corpus comprising various textual genres (news, conversational telephone speech, weblogs, usenet newsgroups, broadcast and talk

[2] https://universaldependencies.org/

[3] Which is freely available for download, under a Creative Commons License, at https://www.kaggle.com/datasets/michelmzerbinati/portuguese-tweet-corpus-annotated-with-ner.

[4] Originally and respectively, ABSTRAÇÃO, ACONTECIMENTO, COISA, LOCAL, OBRA, ORGANIZACÃO, PESSOA, TEMPO, VALOR and OUTRO.

shows) in three languages (English, Chinese, and Arabic). Along with NEs, the corpus also features structural information (such as syntax and predicate argument structure). NE classes are PERSON, NORP[5], FACILITY, ORGANIZATION, GPE[6], LOCATION, PRODUCT, EVENT, WORK OF ART, LAW, LANGUAGE, DATE, TIME, PERCENT, MONEY, QUANTITY, ORDINAL and CARDINAL.

When it comes to NER in Portuguese, adopted corpora are usually those constructed and annotated by the HAREM[7] initiative. During the First HAREM (Santos and Cardoso, 2006), a corpus – known as the Golden Collection (GC) – was compiled from 129 documents, comprising 92,830 words and annotated with 5,270 entities divided into 10 categories, namely ABSTRACTION, EVENT, OBJECT, PLACE, PRODUCTION, ORGANIZATION, PERSON, TIME, VALUE and MISCELLANEOUS.

In this corpus, the most frequent category is LOCATION (with 24.59%), followed by PERSON (21.1%) and ORGANIZATION (18.61%). From this Golden Collection, a smaller corpus was built, called Mini HAREM, which was based on 128 documents extracted from the same domain of the original collection, containing 62,461 words and 3,858 annotated entities (Cardoso, 2006).

Later on, in 2008, during the Second HAREM event (Mota and Santos, 2008), another Golden Collection corpus was generated, extracted from 129 documents, with 147,991 words and 7,836 annotated entities. The main difference, in terms of categories, between the Golden Collection of the First HAREM and that of the Second HAREM was the exchange of the category MISCELLANEOUS[8] for the category OTHER.

In the second Collection, the most frequent category is PERSON, followed by PLACE, TIME, and ORGANIZATION, with 27.11%, 18.15%, 15.21%, and 14.02% of all NEs, respectively (Carvalho et al., 2008). Both HAREM initiatives were the first in NER in Portuguese, which contributed and generated corpora (Mini HAREM and the Golden Collections of both First and Second HAREM) to be used by different research in NER, also building a taxonomy of categories, classes, and sub-classes for the classification of NEs.

Straying from HAREM's Golden Collections, other efforts have been carried out to annotate corpora with NEs in Portuguese. One of them is LeNEr-Br (Luz de Araujo et al., 2018), which is composed of 70 legal documents from various Brazilian courts, with 318,073 words in total.

In that work, 7,836 entities were annotated according to one of the following categories: ORGANIZATION, PERSON, TIME, PLACE, LEGISLATION, and JURISPRUDENCE. As a wider effort, the WikiNER corpus (Nothman et al., 2013) comprises texts extracted from Wikipedia documents, in 9 different languages, including Portuguese, whose NEs were annotated according to the categories PLACE, ORGANIZATION, PERSON, and OTHER.

Tweets have also been the subject of NE annotation efforts. This is the case with FinNum 1.0 (Chen et al., 2018), which comprises 707 unique tweets from the financial domain, extracted from the SemEval-2017 (Cortis et al., 2017) data set. FinNum 1.0 introduces a taxonomy that classifies numerical values into 7 categories: MONETARY, PERCENTAGE, OPTION, INDICATOR, TEMPORAL, QUANTITY and PRODUCT, with a total of 1,341 entities.

When it comes to annotated tweets in Portuguese, one finds the Portuguese (pt-br) NER Twitter corpus (Peres da Silva et al., 2017), which comprises 3,968 tweets with 935 annotated entities from a general domain. In this corpus, annotated categories are PERSON, LOCATION, and ORGANIZATION. In our work, we add to the extant body of resources by focusing on tweets written in Portuguese within the stock market domain, through the annotation of DANTEStocks (Vieira da Silva et al., 2020) according to the taxonomy adopted in the Second HAREM Golden Collection.

3 Materials and methods

As already pointed out, in this work we build on the DANTEStocks corpus, in its December 15, 2022 version[9]. This corpus comprises textual material compiled from Twitter[10] that includes tweets mentioning some of the stocks from iBOVESPA, the main Brazilian Stock Market index, collected during part of 2014. Having itself been built from

[5] Nationalities or religious or political groups

[6] GeoPolitical Entity (countries, cities, states).

[7] Avaliação de Reconhecimento de Entidades Mencionadas – Named Entity Recognition Evaluation

[8] VARIADO, in Portuguese.

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

[10] Currently, X.

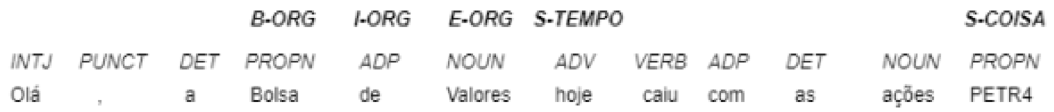

Figure 1: Example of entity category and BIOES annotation

another corpus, presented in (Vieira da Silva et al., 2020), DANTEStocks adds to its predecessor a morphosyntactic annotation layer, in the form of PoS tags, following the Universal Dependencies model (de Marneffe et al., 2021), and delivered in a stand-off manner.

In order to augment DANTEStocks with our NEs layer, we adopted the taxonomy of categories defined and adopted at the Second HAREM's Golden Collection, which includes ABSTRACTION, EVENT, OBJECT, PLACE, PRODUCTION, ORGANIZATION, PERSON, TIME, VALUE, and OTHER (Mota and Santos, 2008). We chose to classify entities only at the category level, which is the broadest level of the taxonomy.

In this work, we did not go down to classes or subclasses, which are specializations of the above mentioned categories. Given that DANTEStocks deals with the financial market domain, entities referring to financial assets, such as the stock tickers PETR4, ITUB4, VALE5, which can be considered objects of this domain, will be classified under the category OBJECT, as this category can encompass entities representing company stocks traded in the financial market. Table 1 presents some examples of tweets and their respective classifications (in bold).

Along with the identification and classification of the entities, we also included BIOES tags (Jurafsky and Martin, 2020), which assists the annotation of entities that consist of multiple tokens. Within this framework, an entity's initial token is marked with a 'B' (begin), its internal tokens with 'I' (inside), and its final token with 'E' (end). Single-token entities are labeled with an 'S' (single), and tokens that are not entities are not annotated, being implicitly represented by an 'O' (outside). That way, we can tell entities that are composed of more than one token from single-token ones.

Figure 1 provides an example of a DANTEStocks tweet, with its PoS annotation, along with the integration of HAREM's taxonomy with BIOES. In this figure, the entity "Stock Exchange" (*Bolsa de Valores*), despite being composed of three tokens, can be identified as a single entity of the ORGANIZATION class, as determined by the 'B', 'I' and 'E' labels, respectively, added to 'ORG". Other entities in this example are "today" (*hoje*) and "PETR4"[13], which have only one token and are marked as 'S' (single), along with their respective classes, TIME and OBJECT.

The annotation of DANTEStocks with the Second HAREM's categories was carried out manually by one of the researchers, following the guidelines defined in the Second HAREM's annotation manual (Mota and Santos, 2008).

4 Results and Discussion

As it turns out, NEs can be found in all of the 4,048 tweets that build DANTEStocks. In total, 23.453 tokens were found to pertain to some Entity (recall that some Entities span over multiple tokens), meaning that almost 28% of all 84.397 tokens of the corpus are NEs, as illustrated in Figure 2.

The fact that 100% of all tweets present at least one NE comes hardly as a surprise, given the way the corpus was originally collected (*cf.* (Vieira da Silva et al., 2020)). In this case, the fact that tweets were fetched based on the presence of some stock market tickers, which are used to represent assets and sometimes as a surrogate to company names (*i.e.* which are themselves entities), virtually guarantees this figure, for such tickers are annotated as OBJECT. Hence, in the absence of any other entity, at least one OBJECT will be present, referring to the stock ticker.

The distribution of entities across tweets can be seen in Figure 3. In this figure, one notices that the amount of NEs found in a single tweet ranges from a single entity (found in 359 tweets, which have only the stock ticker as its NE) up

[11] Represents a set of ideas that are denoted by a proper name in Portuguese and can refer to (a) a discipline or field, a literary, scientific, artistic, religious, or ideological school; or a musical style; (b) represent a condition, especially diseases; (c) an idea; (d) a linguistic object, not the entity it designates(Mota and Santos, 2008).

[12] Represents the idea that the PETR4 stock is the biggest and most significant stock on the São Paulo Stock Exchange.

[13] Petrobras' ticker at the Brazilian Stock Exchange.

Category	Tweet
ABSTRACTION[11]	#petr4 **King Kong**[12] held it... hummmm, watching for an entry *(#petr4 **King Kong** segurou...hummmm observo p/ entrada)*
EVENT	Half Half of the traders in Brazil are trading **World Cup** stickers. That's why PETR4 isn't going up... *(Metade Metade dos traders do Brazil trocando figurinhas da **Copa**. Por isso que a PETR4 nao sobe....)*
OBJECT	Soon I'll be looking at the **#elliottwaves** of **#petr4**. *(Daqui a pouco estarei olhando as **#ondasdeelliot** de **#petr4**)*
PLACE	GOLL4 - GOL Announces Direct Flights between **Fortaleza** and **Buenos Aires** *(GOLL4 - GOL Anuncia Lançamento de Voo Direto entre **Fortaleza** e **Buenos Aires**)*
PRODUCTION	Exclusive CPI of Petrobras Petr4 - Rosa Weber took her time, but abided by the **Constitution**. Call Graça back! *(CPI exclusiva de a Petrobras Petr4 - Rosa Weber demorou mas seguiu a **Constituição**. Chama a Graça de novo!)*
ORGANIZATION	#BR #BOVESPA #ABEV3 **Ambev** will carry out a capital increase to incorporate a tax benefit. *(#BR #BOVESPA #ABEV3 **Ambev** fará aumento de capital para incorporar benefício fiscal.)*
PERSON	RTRS - **MANTEGA**: ONE SHOULD NOT ANNOUNCE A RAISE IN PETROL, ONE SHOULD DO #PETR4. *(RTRS - **MANTEGA**: NAO SE DEVE ANUNCIAR AUMENTO DA GASOLINA, SE DEVE FAZER #PETR4)*
TIME	Analysis of #Ichimoku #PETR4, #BBAS3, #GGBR4, and #ENBR3. Stock guide, trading on **Thursday, April 17th**. *(Análises #Ichimoku #PETR4, #BBAS3, #GGBR4 e #ENBR3. Guia de Ações, pregão de **quinta-feira, 17 de abril**.)*
VALUE	Itub4, daily chart. Closing at **R$ 32.73** with a **0.86%** raise. *(Itub4, gráfico diário. Fechamento em **R$ 32,73** com alta de **0,86%**)*

Table 1: Examples of entity classifications and their corresponding tweets.

to 33 entities (found in five tweets), peaking at three entities, which were found in 581 different tweets. Usually, tweets with three entities follow a pattern whereby the author intends to provide more information about the stock the ticker represents, as with "$PETR3 - Petrobras (petr)" (*i.e.* the ticker for the ordinary Petrobras stock, the company's name and its code) and "$CSAN3 - Cosan (csan-nm)", followed by some other content.

In the above examples, $PETR3, $CSAN3, petr and csan-nm are all annotated as OBJECT, whereas Petrobras and Cosan are ORGANIZATION. Tweets with four and five entities usually follow the same pattern as that of tweets with three entities, but with the addition of some other entity. At the opposite end of the scale, tweets with more than 30 entities are usually composed of a stream of stock tickers and their respective values, as in "PETR4 R$ 15.42 VALE5 R$ 26.93 ...", being mostly composed of entities belonging to the VALUE and OBJECT categories.

Regarding the distribution of NEs across classes, one sees a predominance of OBJECTs (with 45.03% og all entities), as illustrated in Figure 4, with VALUE coming second (with 22.29% of the entities belonging to this class). This is something that is expected, given the nature of the tweets' content, focusing in some tickers, such as "PETR4", "ITUB4" and "VALE5", which were classified as OBJECTs, and their corresponding values. A better visualisation of the proportion of each class related to the total amount of NEs in the corpus can be seen in Figure 5.

The next three classes in Figure 4 are ORGANI-

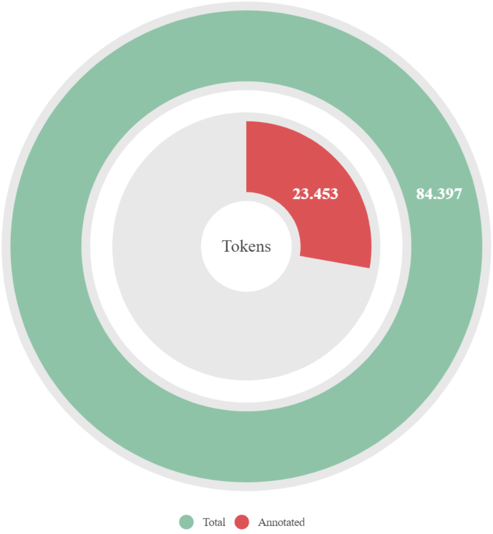

Figure 2: Number of tokens in the corpus (total and belonging to NEs).

ZATION, which is used to classify references to companies, such as “Petrobras”, “Itaú” and “Vale” for example; TIME and PERSON, which usually represents references to politicians that had somehow influenced the market. In the sequence comes PLACE, usually representing cities and some geographical locations, such as “Bacia das Almas”, “New York” and “Santos”. Categories such as ABSTRACTION, EVENT end PRODUCTION were very rare. Interestingly, the left-over category – OTHER, which was designed to group entities that do not fit in any other class, was not necessary in this corpus.

Although in our work entities may consist of multiple tokens, as in “Banco do Brasil”[14], which is a single entity of the ORGANIZATION class, spanning three annotated tokens, “Banco”, “do” and “Brasil”, this was not the rule along the corpus, as illustrated in Figure 6. In this figure, one sees the distribution of BIOES tags in each category, that is the amount of tokens at the beginning (B), ending (E) and inside (I) NEs, along with entities that correspond to a single (S) token, for each of the adopted NEs classes.

As it turns out, there is a predominance of single-token entities (the S tag in the figure) in five of the six more frequent classes. The only exception lies with VALUE which is rather balanced between single and multiple-token entities. Still, the low amount of internal tokens (I) indicates that entities with two tokens are more common that entities with three or more tokens.

This, in turn, may be explained based on the fact that stock tickers are composed of a single token. One has then to recall that the way the corpus was gathered, by fetching tweets mentioning at least one of the stocks that build up iBOVESPA, guarantees these to happen in all tweets (inline with the prevalence of OBJECT entities, illustrated in Figure 4), which in turn makes a significant impact in the imbalance observed in Figure 6. This is one of the main weaknesses of this corpus – the fact that the resulting distribution of categories was probably determined by the way the corpus was compiled, at least when it comes to OBJECTs. Still, we believe it to be a valuable resource for the comunity.

5 Conclusion

In this work we introduced a corpus annotated with NEs following the terminology adopted in the Second HAREM (Mota and Santos, 2008). Being previously annotated with morphosyntactic information (PoS tags, following the Universal Dependencies model) along with emotions (according to Plutchik’s Wheel of Emotions), this corpus represents an opportunity to link all this information, thereby providing researchers with a valuable tool to study[15] phenomena related to these dimensions.

Additionally, and to the best of our knowledge, this is the first corpus to allow for such a cross-dimensional analysis in Portuguese, and perhaps in any other language, specially in the domain of tweets from the financial market. Among other possibilities, this corpus can be used to study the relation of NEs and the tweets’ associated emotion, perhaps correlating them to stock market price moves.

Regarding weaknesses of our research, one has to recall that the current version of the corpus was manually annotated by a single person only. Although this effort was carried out in a systematic way, resulting in an annotation guide to be used by others, results are still bound to reflect this annotator’s opinion. We, however, intend to remedy this in the near future, by having other annotators deal with the corpus, in accordance do the guidelines built during the current research.

As for future research, a deeper exploration can be conducted to determine the specific classes and

[14] Bank of Brazil

[15] Which is freely available for download, under a Creative Commons License, at https://www.kaggle.com/datasets/michelmzerbinati/portuguese-tweet-corpus-annotated-with-ner.

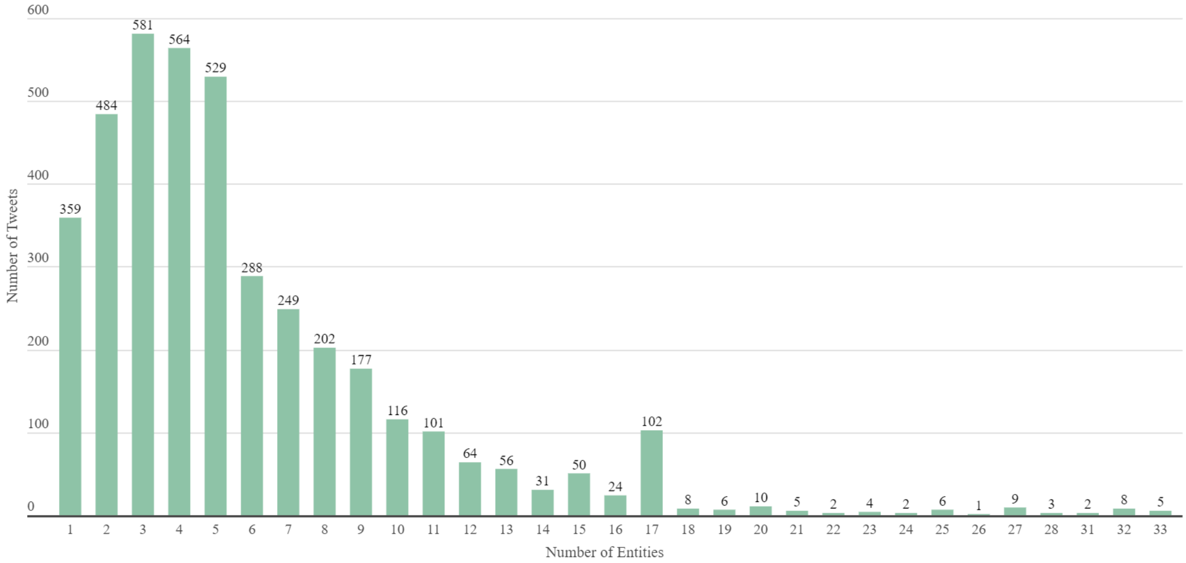

Figure 3: Number of Entities in each tweet and amount of tweets with that amount of Entities.

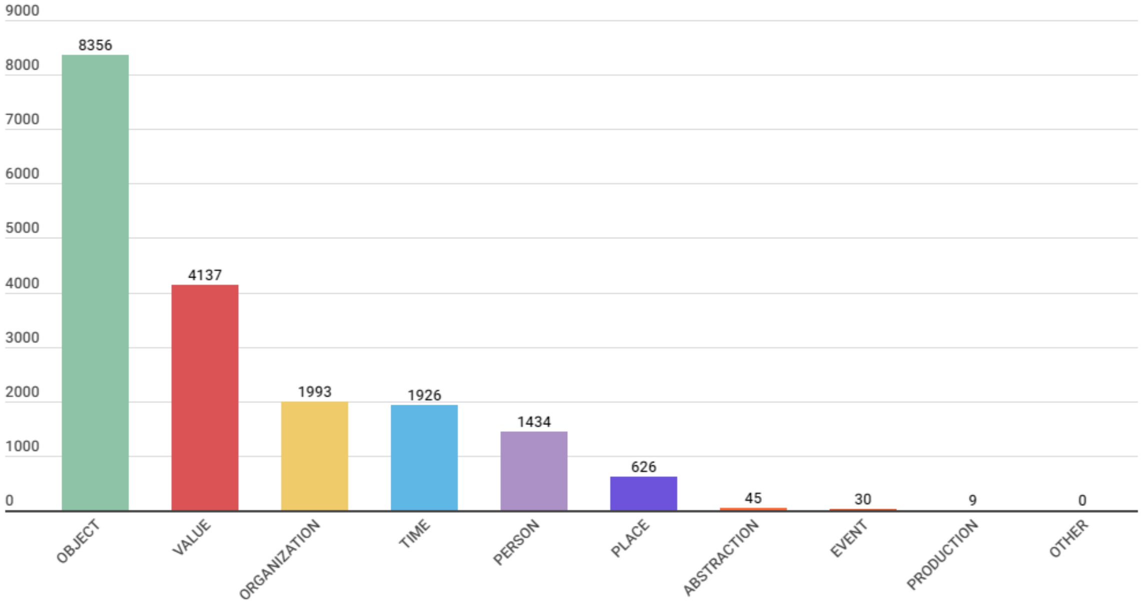

Figure 4: Amount of entities in each class.

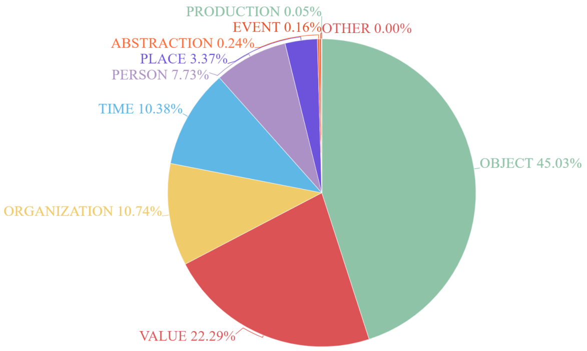

Figure 5: Percentage of entities by category

subclasses to which financial assets can be assigned, rather than keeping them only at the category level. It is also our intention to increase the size of the corpus, through the annotation of a larger collection of stock market tweets.

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.

References

Nuno Cardoso. 2006. Harem e miniharem: Uma análise comparativa. In *Encontro do HAREM (Porto, Portugal, 15 de Julho de 2006)*.

Paula Carvalho, Hugo Gonçalo Oliveira, Diana Santos, Cláudia Freitas, and Cristina Mota. 2008. Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O segundo harem.

Chung-Chi Chen, Hen-Hsen Huang, Yow-Ting Shiue, and Hsin-Hsi Chen. 2018. Numeral understanding in financial tweets for fine-grained crowd-based forecasting. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 136–143.

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.

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

Ariani Di-Felippo, Caroline Postali, Gabriel Ceregatto, Laura Gazana, Emanuel Silva, Norton Roman, and Thiago Pardo. 2021. Descrição preliminar do corpus dantestocks: Diretrizes de segmentação para anotação segundo universal dependencies. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 335–343, Porto Alegre, RS, Brasil. SBC.

Daniela O. F. do Amaral and Renata Vieira. 2013. O reconhecimento de entidades nomeadas por meio de conditional random fields para a língua portuguesa (named entity recognition with conditional random fields for the Portuguese language) [in Portuguese]. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*.

Michael Fleischman. 2001. Automated subcategorization of named entities. pages 25–30.

Michael Fleischman and Eduard Hovy. 2002. Fine grained classification of named entities. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Dan Jurafsky and James H. Martin. 2020. *Speech and Language Processing*, pages 280–281. Taylor Graham Publishing, GBR.

Ei Thwe Khaing, Myint Myint Thein, and Myint Myint Lwin. 2019. Stock trend extraction using rule-based and syntactic feature-based relationships between named entities. In *2019 International Conference on Advanced Information Technologies (ICAIT)*, pages 78–83.

J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. 2003. GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl$_1$) : $i180 - -i182$.

Saul Kripke. 1982. *Naming and Necessity*. Boston: Harvard University Press.

Pedro Henrique Luz de Araujo, Teofilo de Campos, Renato Oliveira, Matheus Stauffer, Samuel Couto, and Paulo De Souza Bermejo. 2018. *LeNER-Br: A Dataset for Named Entity Recognition in Brazilian Legal Text: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings*, pages 313–323.

Michał Marcińczuk and Maciej Piasecki. 2015. Named entity recognition in the domain of polish stock exchange reports.

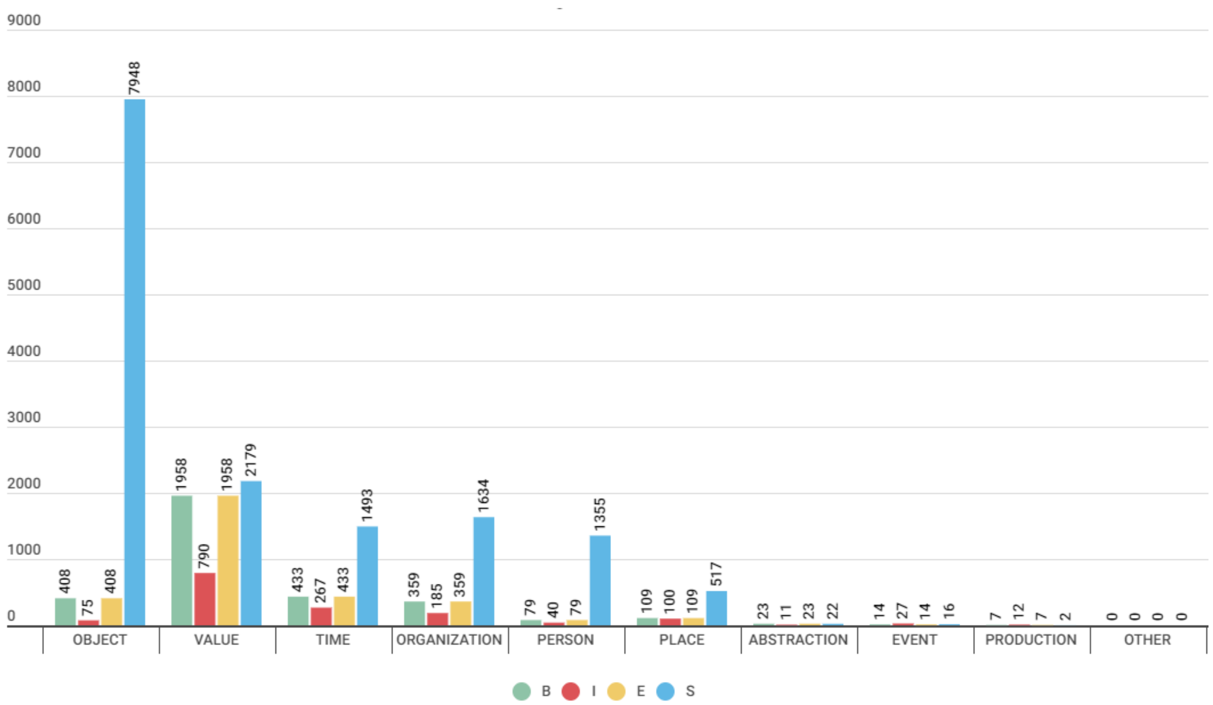

Figure 6: Frequency - Category x BIOES

Cristina Mota and Diana Santos. 2008. Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O segundo harem.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30:3–26.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175. Artificial Intelligence, Wikipedia and Semi-Structured Resources.

Rafael Peres da Silva, Diego Esteves, and Gaurav Maheshwari. 2017. Bidirectional lstm with a context input window for named entity recognition in tweets. pages 1–4.

R Plutchik and H Kellerman. 1986. Emotion - theory, research, and experience, vol 3, biological foundations of emotion.

Mitchell Marcus Eduard Hovy Sameer Pradhan Lance Ramshaw Nianwen Xue Ann Taylor Jeff Kaufman Michelle Franchini Mohammed El-Bachouti Robert Belvin Ann Houston Ralph Weischedel, Martha Palmer. 2013. *OntoNotes Release 5.0*. Philadelphia: Linguistic Data Consortium.

Diana Santos and Nuno Cardoso. 2006. A golden resource for named entity recognition in portuguese. In *Computational Processing of the Portuguese Language*, pages 69–79, Berlin, Heidelberg. Springer Berlin Heidelberg.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Fernando J. Vieira da Silva, Norton T. Roman, and Ariadne M.B.R. Carvalho. 2020. Stock market tweets annotated with emotions. *Corpora*, 15(3):343–354.

Shuwei Wang, Ruifeng Xu, Bin Liu, Lin Gui, and Yu Zhou. 2014. Financial named entity recognition based on conditional random fields and information entropy. In *2014 International Conference on Machine Learning and Cybernetics*, volume 2, pages 838–843.