NounBank.DS: a Lexical Repository of Nominal Frames from Stock Market Tweets in Brazilian Portuguese

Bryan K. S. Barbosa[1,3], **Ariani Di Felippo**[1,2]

[1] Núcleo Interinstitucional de Linguística Computacional (NILC)

[2]Departamento de Letras, Universidade Federal de São Carlos (UFSCar)
Caixa Postal 676 - 13565-905 - São Carlos/SP – Brazil

[3]Programa de Pós-Grad. em Linguística, Universidade Federal de São Carlos (UFSCar)
Caixa Postal 676 - 13565-905 - São Carlos/SP – Brazil

bryankhelven@ieee.org, ariani@ufscar.br

Abstract. *This paper describes NounBank.DS, a project that provides argument structure for instances of predicate nouns in DANTEStocks, a Dependency-Analyzed corpus of stock market Tweets in Portuguese. NounBank.DS is part of a larger effort to add additional semantic layers of annotation to DANTEStocks. This and other annotation projects taken together should lead to the creation of better tools for the automatic analysis of tweets on the stock market. This paper describes the NounBank.DS project in detail, including its specifications and the process involved in creating the resource.*

1. Introduction

Semantic Role Labeling (SRL) is the task of identifying a predicate and its arguments, assigning to each argument the semantic role it fulfills within the underlying argument structure [Jurafsky and Martin 2025]. By systematically answering who performed what action upon whom, as well as when, where, how, and why, SRL establishes an intermediate representation that connects surface syntax to deeper levels of interpretation in Natural Language Processing (NLP).

The PropBank project[1] [Palmer et al. 2005, Pradhan et al. 2022] has stood as the canonical reference for SRL studies by providing two key resources: a lexical repository of frame files together with manually annotated corpora. Each PropBank resource serves a distinct but complementary purpose in NLP.

The frame files enumerate every possible sense of each predicate (initially verbs) and delineate the set of semantic roles associated with each sense, furnishing the conceptual guidelines for both manual corpus annotation and guiding models in understanding the meaning and function of different sentence components. For example, the verb *share* has only one sense, *share.01*, glossed as "share, giving into co-ownership". It is described by a specific set of semantic roles along with their descriptions (known as "roleset"): Arg0 (*sharer*), Arg1 (*thing shared*), and Arg2 (*shared with, if separate from Arg0*).

The PropBank-annotated corpora consist of sentences in which every predicate is linked to a particular sense from the lexical repository, and each argument receives

[1]https://propbank.github.io/

the semantic role specified in that sense's roleset. For example, in the sentence "*John shared costs and risks with his colleagues*", the predicate *share* is annotated as *share.01* and *John* realizes Arg0, *costs and risks* realizes Arg1, and *with his colleagues* realizes Arg2. Arguments that encode general information such as time (e.g. ***Yesterday**, John shared costs and risks with his colleagues*) are outside the frames, yet may be annotated as ArgM-X (e.g. *yesterday*=ArgM-TMP). Thus, the annotated corpora furnish supervised data for training and evaluating SRL models, supporting both rule-based and data-driven approaches, and serving as benchmarks such as CoNLL-2005 and CoNLL-2012.

Advances in data-driven modeling – progressing from feature-based classifiers to deep neural networks and, more recently, Large Language Models (LLMs) – have not diminished PropBank's relevance. The current state of the art for SRL in English and Chinese, reported by [Li et al. 2025], couples retrieval-augmented generation (RAG) with an LLM that integrates external semantic knowledge extracted from PropBank frame files. Their architecture surpasses previous SRL models, including encoder-decoder and pipeline-based approaches, on standard benchmarks such as CoNLL-2005 and CoNLL-2012 (for English) and OntoNotes 5.0 (for Chinese).

Specifically, several PropBank annotated corpora have become standard benchmarks for SRL. The original corpus, based on Penn Treebank [Marcus et al. 1993, Taylor et al. 2003], consists of news texts from the Wall Street Journal (WSJ) and served as the foundation for the CoNLL-2005 shared task. Subsequent initiatives enlarged the "propbanked" resources with a coverage of predicates, genres, and languages. For instance, the NomBank[2] project [Meyers et al., 2004] was particularly responsible for annotating over 115,000 nominal predicate-argument structures in Penn Treebank based PropBank-style nominal frame files. OntoNotes[3] [Pradhan et al. 2013] covers diverse genres (broadcast news and conversation, telephone speech, weblogs, newsgroups, biblical text) and two additional languages (Chinese and Arabic), becoming the foundation of CoNLL-2012. The PropBank annotation methodology was later adapted to other languages, including Korean [Palmer, Martha et al. 2006], Hindi/Urdu [Bhatt et al. 2009], Finnish [Haverinen et al. 2015], Turkish [Şahin and Adalı 2017], Persian [Mirzaei and Moloodi 2016], Russian [Moeller et al. 2020], and Brazilian Portuguese (BP) [Duran and Aluísio 2011]. Cross-lingual generalizations continue through the Universal PropBanks initiative [Akbik et al. 2015, Jindal et al. 2022], further broadening the resource's typological reach.

The foundation of PropBank annotation lies in its lexical repository of frame files. The original release focused exclusively on verbal predicates and contained nearly 3,300 frames. NomBank later added roughly 4,700 nominal frame files, most of which aligned with their corresponding verbal frames. Initially, the projects maintained separate rolesets by part-of-speech (PoS). Later, [O'Gorman et al. 2018] introduced unified rolesets, merging etymologically related lemmas – such as verb, noun, and adjective – sharing a common meaning under single rolesets, thus expanding lexical coverage and allowing more robust conceptual generalization. Each unified roleset contains `alias` fields that specify all related lemmas along with their PoS tags (v, n, j); for instance, *share.01* lists the aliases *share-v*, *sharing-n*, and *share-n*. The roleset is named after the

[2]https://nlp.cs.nyu.edu/meyers/NomBank.html

[3]https://catalog.ldc.upenn.edu/LDC2013T19

most frequent verbal alias, or after a nominal or adjectival alias when no verb sense is attested. Following the unification, the latest release of PropBank (version 3.4) comprises more than 7,500 frame files. For BP, the counterpart resource is Verbo-Brasil[4] [Duran et al. 2013, Duran and Aluísio 2015], which contains 1,453 verb senses (1,060 lemmas), of which only 109 lack an explicit link to an English PropBank sense.

Given the pivotal role of PropBank and NomBank in SRL development and the increasing demand for semantic processing of user-generated content (UGC), this paper introduces *NounBank.DS*. It is a NomBank-like repository containing 145 unique Portuguese nouns and their predicate-argument structures in 1,756 instances extracted from DANTEStocks, a financial-domain corpus of tweets (now rebranded as "posts") [Di-Felippo et al. 2022]. The resource is expected to facilitate forthcoming manual annotation within the DANTE project, which aims to construct UGC corpora annotated according to the Universal Dependencies framework [Nivre et al. 2020] as part of the multi-genre Porttinari treebank for BP [Pardo et al. 2021]. Furthermore, NounBank.DS may contribute with the development of SRL tools tailored both to the stock-market domain and to wider multi-genre scenarios, since nominal propositions are also being annotated across other Porttinari sub-corpora.

The remainder of the article is organized as follows: Section 2 outlines the DANTEStocks corpus and its UD annotation. Section 3 recaps the linguistic investigation conducted by [Barbosa 2024] that motivated our work. Section 4 presents NounBank.DS, detailing its web interface and JSON release. Section 5 reports our results and their implications. Section 6 concludes with our final remarks and future work.

2. The DANTEStocks Corpus

DANTEStocks is a corpus of tweets (tweebank) on the stock market domain [Di-Felippo et al. 2022]. Because sentiment expressed in social media often correlates with market indices, data of this kind has been exploited in predictive finance [Deveikyte et al. 2022]. The corpus comprises 4,048 tweets (∼81k tokens) that mention the 73 stocks from Ibovespa[5]. All posts were automatically collected from Twitter in 2014 and therefore respect the former 140-character limit. No lexical normalization was applied; each tweet is treated as a single textual unit without sentence segmentation. DANTEStocks presents a combination of standard and non-standard written language, as well as domain specific vocabulary (like company names, cashtag[6], stock ticker[7] symbol, and financial jargon) and medium (Twitter/X) features (such as hashtag, at-mention, retweet marker, truncation, URL and emoticons) [Di-Felippo et al. 2021].

The grammatical annotation of the corpus follows the Universal Dependencies (UD) framework [de Marneffe et al. 2021]. UD specifies both morphological and syntactic information: each token is assigned a part-of-speech tag, a lemma, and a set of morphological features, while syntactic structure is expressed as a dependency tree whose edges carry labels ("dependency relations", or *deprels*). Currently, the model has 17 PoS

[4]http://143.107.183.175:21380/verbobrasil/

[5]The main index of the Brazilian stock exchange B3.

[6]Marker specifically designed to track financial instruments (e.g., $PETR4).

[7]It consists of a five- or six-character alphanumeric code used to identify a specific class of a company's stock – for example, "PETR4" refers to Petrobras' preferred shares.

tags and 37 deprels, plus a non-fixed set of morphological features. The basic representation is a tree (Figure 1), where exactly one word is the head of the utterance (root) (e.g. *assina*) ("signs"), and all other words are dependent on another word. In Figure 1, PoS tags and the lemmas are displayed below the text. The morphological features are not included in the figure, but the noun *acordo* ("agreement"), for example, has the following features Number=Sing and Gender=Masc. The UD-annotation of an utterance is encoded in a CoNLL-U file, illustrated in Figure 2.

The corpus was annotated through an iterative cycle of automatic parsing followed by manual revision, resulting in high-quality morphosyntactic (or PoS) and syntactic annotations [Di-Felippo et al. 2024a]. The UD-annotation already supported the development of various NLP tools (e.g. [Silva et al. 2021, Di-Felippo et al. 2024a, Di-Felippo et al. 2024b]) and the corpus itself was material for some linguistic studies (e.g. [Scandarolli et al. 2023, Barbosa 2024]).

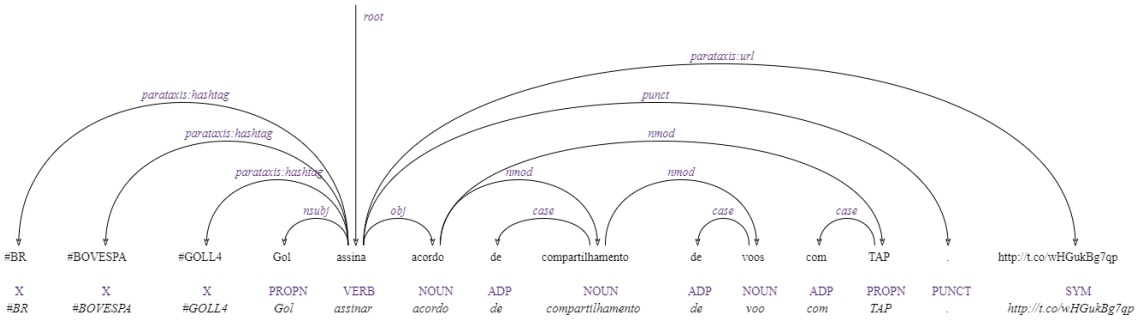

Figure 1. Dependency tree representation of UD annotation for a tweet.

Id	Form	Lemma	UPos	XPos	Feats	Head	Deprel	Deps	Misc
1	#BR	#BR	X	_	_	5	parataxis	_	_
2	#BOVESPA	#BOVESPA	X	_	_	5	parataxis	_	_
3	#GOLL4	#GOLL4	X	_	_	5	parataxis	_	_
4	Gol	Gol	PROPN	_	_	5	nsubj	_	_
5	assina	assinar	VERB	_	Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin	0	root	_	_
6	acordo	acordo	NOUN	_	Gender=Masc|Number=Sing	5	obj	_	_
7	de	de	ADP	_	_	8	case	_	_
8	compartilhamento	compartilhamento	NOUN	_	Gender=Masc|Number=Sing	6	nmod	_	_
9	de	de	ADP	_	_	10	case	_	_
10	voos	voo	NOUN	_	Gender=Masc|Number=Plur	8	nmod	_	_
11	com	com	ADP	_	_	12	case	_	_
12	TAP	TAP	PROPN	_	_	6	nmod	_	SpaceAfter=No
13	.	.	PUNCT	_	_	5	punct	_	_
14	http://t.co/wHGukBg7qp	http://t.co/wHGukBg7qp	SYM	_	_	5	parataxis	_	SpaceAfter=No

Figure 2. CoNLL-U file corresponding to the UD annotation for Figure 1 tweet.

3. The Rise of NounBank.DS

The study conducted by [Barbosa 2024] served as the foundation for the creation of NounBank.DS. Motivated by the prominent role of predicate nouns in digital texts in the financial domain [Voskaki et al. 2016], and the language phenomena of tweets, the author analyzed the nominal valency (i.e. number, type and form of arguments) in DANTEStocks. Three principal tasks supplied the raw material for the present repository: (i) identifying predicate nouns and their instantiated arguments, (ii) describing argument structure in semantic terms, and (iii) aligning those semantic roles with the UD syntax. All outputs were stored in a structured spreadsheet that served as the seed dataset for building NounBank.DS.

a) Selection of Nominal Propositions

The identification of predicate nouns and their instances[8] (propositions) was carried out semi-automatically. Given the UD PoS annotation layer, an initial automatic search for the noun + preposition (NOUN + ADP) pattern was performed, since argument realization typically involves prepositional phrases. The results were manually reviewed to exclude irrelevant cases and validate the remaining instances. The validation was based on the DUPB[9] dictionary [Borba 2002] with the support of domain specialists. Subsequently, 1,122 instances were identified, encompassing 145 predicate noun lemmas (336 different word forms). Next, a second automatic search using the validated noun lemmas targeted instances where prepositional complements were absent and arguments appeared in alternative forms (such as adjectival modifiers), yielding a final total of 1,756 instances.

b) Argument Structure Description

The identification of the arguments in all instances was manually described following the NomBank framework, specifically the repository of 4,706 (nominal) frame files. The sense of each noun in an instance was first defined based on context, and an appropriate English equivalent was used to access the frame files. Although many nouns, such as *compartilhamento* ("sharing"), had straightforward equivalents, others posed challenges due to domain-specific meanings, such as *desova* ("rapid or mass disposal of assets"), translated to "sell-off". Next, the roleset from the frame file corresponding to the Portuguese noun sense was selected, and the arguments in all instances of the noun with this sense were identified accordingly. In addition to the rolesets, the argument identification step was also guided by the UD-dependency parse of the tweets, as illustrated in Figure 1. By anchoring semantic roles to syntactic parses, one can systematically determine the roles of dependents (like subjects and objects) across different sentence constructions.

c) Syntax-semantics mapping of arguments

Finally, the semantic arguments (Arg0-Arg5) were linked to the dependency relations they bear to the predicate noun. Whereas the previous step merely consulted the UD trees for guidance, the present step extracted the relevant DEPRELS directly from the CoNLL-U files[10]. The resulting mapping records, for each role, the distribution of syntactic realizations attested in the corpus, thereby enabling subsequent quantitative and comparative analyses of nominal predication in user-generated financial text.

4. The NounBank.DS repository

Based on Barbosa's data, NounBank.DS has been compiled and released in two parallel formats: HTML and JSON, so the resource is equally accessible to human readers and downstream applications.

4.1. From Spreadsheets to Repository Files

The raw material for NounBank.DS consisted of the `DANTEStocks.conllu` files along with a set of `.xlsx` spreadsheets produced by [Barbosa 2024], with each spreadsheet grouping together every instance whose predicate noun realized the same number of arguments (e.g., all one-argument cases in Sheet1, all two-argument cases in Sheet2,

[8] An instance is any occurrence of a predicate noun accompanied by one of its arguments.

[9] DUPB stands for Dicionário de Usos do Português do Brasil. It is a lexicographic resource that incorporates syntactic-semantic information based on valency grammar.

[10] https://universaldependencies.org/format.html

and so forth). Rows were keyed by the unique tweet identifier (`sent_ID`) and contained three obligatory columns –`Id`, `text` (the full tweet), and `Args` – plus a varying number of annotation columns that specified which arguments (Arg0-Arg5) were manifest in that instance and, when absent, marked them explicitly with "–".

Because a single spreadsheet could list several different lemmas, and because cells occasionally stored free-form observations or employed heterogeneous orthography (*compra*/*comrpa*, *exemplo*/*ex*), manual rectification was necessary. The rectification phase left every instance intact, but harmonized column headers, standardized lemma occasional spellings, and ensured that every argument column was present even when empty, thereby enabling consistent algorithmic access.

Once this uniform template had been imposed, the spreadsheets could be funneled into a semiautomatic Python workflow (pandas + openpyxl) that turned the harmonized yet still heterogeneous records into the structured lexical entries of NounBank.DS through three successive steps:

(i) *Spreadsheet ingestion and lemma partitioning*. Each cleaned spreadsheet was loaded, and its rows were partitioned by lemma so that subsequent processing could treat every predicate noun in isolation.

(ii) *Enrichment with syntactic data*. For each spreadsheet row, the script located the corresponding tweet in the CoNLL-U file by its `sent_ID`, located the predicate noun token, and extracted (a) every dependency relation in which the noun functioned as *head* and (b) any relation chain that projected a subject from the clause root to the noun (cf. the examples in Section 3). The enriched record was serialized as a minimalist JSON object with six fields: SENT_ID (text identifier), TEXT (raw tweet text), ROLESET_ID (unified sense label, like *sharing.01*), REALIZATION (dictionary that maps `Arg0-Arg5` to their surface spans ("–" if unexpressed)), SYNTAX (parallel dictionary that maps the same arguments to the *dominant* UD relation governing each span (e.g. `nmod`, `nsubj`), and STATS (empty dictionary initialized at instance level. During lemma-level aggregation, this field is populated with counts of every `<role, deprel>` pairing; the populated version drives the frequency table rendered in the interface (see the lower panel of Figure 4)).

(iii) *Aggregation and export*. After all rows for a lemma had been processed, the script incremented the STATS counters to obtain the frequency of every (role, *deprel*) pairing, merged the instance-level objects into a lemma-level JSON entry, and wrote:
(a) a master JSON file holding the full inventory of lemmas;
(b) one lemma-specific JSON file per noun; and
(c) a JavaScript bundle that embeds the same information as a global variable, later consumed by the static site generator for the HTML pages.

Before control passed to the static-site generator, a brief manual post-editing stage completed each lexical entry. The JSON objects produced by step (iii) already contained every datum needed for programmatic access – roleset identifier, argument spans, UD relations, and their aggregate counts – yet they did not specify which instances should be showcased, nor how those instances ought to be visually highlighted, or even guaranteed that the English roleset link and role inventory were correct.

Consequently, for every lemma a human annotator (i) selected up to three representative tweets per core role (favouring short, unambiguous spans) and then flagged them inside the JSON under `examples`; (ii) inserted lightweight HTML markup (color spans, line breaks) that would later be rendered exactly as in Figure 4; (iii) verified the English mapping link of the `roleset_id` and cross-checked that the list of roles in both the JSON and the `Roles` section matched the original NomBank specification.

All edits were written back to the lemma-level JSON and duplicated in a companion Markdown file; at build time the site generator merges these sources so that HTML pages and downloadable JSON stay perfectly aligned.

4.2. Interface of the Web Application

The HTML version of NounBank.DS[11] is delivered through a lightweight single-page web application. Figure 3 presents the main page: a clean design for easy navigation and exploration of predicate nouns from the DANTEStocks corpus. At the top, a navigation bar provides access to the core sections – HOME, ABOUT, and CONTACT – while the central area displays the repository title ("NounBank.DS") and a brief description in Portuguese. Beneath the header, an alphabetical menu allows users to browse noun entries by initial letter, and a scrollable list displays the corresponding lemmas.

Figure 3. NounBank.DS main page.

The HTML interface presents each noun entry in a user-friendly view. Figure 4 illustrates the lexical entry of the noun *compartilhamento*. The header states the Portuguese lemma and sense identifier (*compartilhamento.01*), followed by the indication of the corresponding NomBank roleset – also referred to as the roleset ID (*sharing.01*) –, together with a hyperlink to the aligned English PropBank sense (*verb-share.01*). The roleset inventory, inherited verbatim from NomBank, is reproduced with its semantic glosses. To illustrate each role, the interface shows example tweets with the arguments described (argument spans are visually distinguished (e.g. colour shading) for quick inspection). Next, it describes the syntactic realization of arguments by UD-dependency relations, with a frequency table rendered for easy visualization of the counts. The layout resembles a NomBank frame file entry, but is enhanced with a compact overview of the distribution of syntactic patterns for the predicate's arguments, similar to the lexical entries in FrameNet[12].

[11] `https://bryankhelven.github.io/NounBank.DS/`
[12] https://framenet.icsi.berkeley.edu/luIndex

Figure 4. Lexical entry of "compartilhamento" in the NounBank.DS web interface.

4.3. Machine-readable format

The machine-readable release is a single JSON file whose top-level array stores one object for each predicate noun. Inside every object the information is organized in four coherent blocks that parallel the sections displayed on the website. First, a compact `header` records the lexical unit through three scalar keys: the Portuguese lemma (e.g. *compartilhamento*), its unified sense label (*compartilhamento.01*), and the aligned English roleset inherited from NomBank (*sharing.01*). This header is followed by a `roles` array that lists the core semantic roles – Arg0, Arg1 and Arg2 – in this case-each accompanied by the NomBank gloss (*sharer*, *thing shared*, *shared with*). A third key, `examples`, holds the illustrative instances shown in the interface: for *compartilhamento* two tweets are preserved, and for every tweet the file records both the surface span of each realized argument (in a sub-object called `realization`) and the dominant UD relation that links that span to the noun (in a parallel sub-object called `syntax`). Arguments not attested in the sentence are encoded with a null symbol, keeping the alignment across roles intact. Finally, a dictionary named `syntactic_profile` aggregates all observations for the lemma, counting how often each role is realized by each dependency relation – here Arg0 appears twice as *nsubj*, Arg1 twice as *nmod*, and Arg2 does not occur. Because the JSON reproduces exactly the predicate-argument linkages, illustrative sentences, and UD relations displayed in the interface, it can be queried programmatically, used as input to SRL pipelines, or fuel quantitative studies without any loss of information.

```
{
  "lemma": "compartilhamento",
  "roleset_id": "compartilhamento.01",
  "english_roleset": "sharing.01",
  "roles": [
    { "id": "Arg0", "desc": "sharer" },
    { "id": "Arg1", "desc": "thing shared" },
    { "id": "Arg2", "desc": "shared with, if separate from Arg0" }
  ],
  "examples": [
    {
    "sent_ID": "dante_01_4546077433921781771",
      "text": "#BR #BOVESPA #GOLL4 Gol assina acordo de compartilhamento de voos com TAP.",
      "realization": { "Arg0": "Gol",       "Arg1": "de voos",  "Arg2": null },
      "syntax":      { "Arg0": "nsubj",     "Arg1": "nmod",     "Arg2": null }
    },
    {
      "sent_ID": "dante_02_452825022586307589y",
      "text": "$GOLL4 - GOL e TAP assinam acordo para compartilhamento de voos.",
      "realization": { "Arg0": "GOL e TAP",  "Arg1": "de voos",  "Arg2": null },
      "syntax":      { "Arg0": "nsubj",     "Arg1": "nmod",     "Arg2": null }
    }
  ],
  "syntactic_profile": {
    "Arg0": { "nsubj": 2 },
    "Arg1": { "nmod": 2 },
    "Arg2": {}
  }
}
```


Figure 5. JSON file of the "compartilhamento" entry from NounBank.DS.

5. Results and Discussion

The repository covers 145 distinct predicate nouns and 336 different word forms, including inflectional and orthographic variants. Since DANTEStocks contains 4,048 tweets and the 1,756 nominal instances cover 1,212 unique tweets, this means that the repository includes about 30% of the tweets from DANTEStocks. On average, each noun has about 12 instances, though the distribution is highly skewed: a few frequent nouns (e.g. *oferta* ("offer"), *compra* ("purchase"), *venda* ("sale"), *alta* ("rise/increase") each occur dozens of times, while many others appear only once or a handful of times. Most tweets (63%) contain a single instance, but 37% include two or more (e.g. *acordo* and *compartilhamento* in (1)).

In general, every lexical entry is linked to at least one NomBank roleset; in total the repository contains 145 distinct frames, where true polysemy is rare: only *acordo* activates more than one roleset, being mapped to *agreement.01*, *accord.01* and *accordance.01*. Synonymy is also rare, with only two rolesets being shared by pairs of lexemes – *rateio* ("apportionment") and *locação* ("leasing/allocation") both instantiate *allocation.01*, while *fusão* ("merger") and *incorporação* "incorporation" realise *fusion.01*. Apart from these two clusters (four nouns altogether), every lemma maps one-to-one to its roleset, yielding a clean inventory that is easy to align with external resources.

Although based on NomBank, the semantic role description differs from it in several respects. The first concerns the fact that arguments were identified using the tweet as the basic unit of analysis, rather than the propositional NP, also called "markable" noun phrase (NP) by [Meyers 2007]. The second difference lies in the fact that, due to the scope of Barbosa's project, only core arguments - those defined in the rolesets - were considered, as they are essential to the predicate's meaning and structure. Consequently, modified arguments were not addressed in the instances. Finally, since the tweet is the unit of analysis, subject arguments that occur outside the NP were also considered nominal arguments whenever possible. This approach reflects the goal of capturing the full argument structure of predicate nouns within the context of entire tweets, even when those arguments are syntactically external to the NP (cf. Figure 1 and Figure 4).

Quality control of the semantic layer relied on an independent three-judge evaluation rather than a single "second pass". Baseline annotations were created by the first author and, after four weeks of guided training, two additional linguistics students annotated 180 unseen instances (approximately 10% of the corpus, drawn from 42 predicate nouns of varying valency). For each instance they received the pre-selected NomBank roleset and had to decide, for each argument, whether it was overtly realized and, if so, which span it covered. Agreement was measured with Fleiss's Kappa κ [Fleiss 1981], appropriate for three annotators. The overall coefficient was $K = 0.80$, classified as "excellent" by Fleiss's scale. Broken down by role, scores were $K_{\text{Arg0}} = 0.69$, $K_{\text{Arg1}} = 0.74$ (both "good"), and $K_{\text{Arg2}} = 0.80$, $K_{\text{Arg3}} = 0.86$, $K_{\text{Arg4}} = 0.91$ (all "excellent"). These figures-obtained on a deliberately challenging sample-attest to the consistency of the role-identification guidelines and confirm that the lexical-semantic information distributed with NounBank.DS is reliable for downstream use.

Integrating UD-dependencies into a classic NomBank-style frame file represents a modern approach to the syntax-semantics interface: predicate-argument information is anchored in the same dependency formalism employed by state-of-the-art parsers, allowing NounBank.DS to slot directly into any UD-based pipeline. At the same time, the use of UD labels standardizes the recording of argument realization and yields a cross-linguistically comparable inventory of patterns-whereas the original NomBank, tied to Penn Treebank constituency labels, could not offer this level of framework or language portability.

6. Final Remarks and Future Work

This resource will support future manual semantic annotation efforts within the DANTE[13] project (Dependency-ANalyzed corpora of TwEets), which develops corpora annotated according to the UD framework, as part of the Porttinari[14] multigenre treebank for BP. It also enables the integration of a semantic role layer into nominal propositions in the DANTEStocks corpus, following the NomBank framework, and can support LLM-based pipelines (e.g., as a RAG layer or weak supervision). Furthermore, it will contribute to the development of SRL tools tailored both to the stock market domain and to broader multigenre contexts, since nominal propositions in other Porttinari corpora are also being annotated; a cross-walk to PT resources (e.g., Verbo-Brasil, FrameNet Brasil, OpenWordNet-PT) is planned.

More details about this work may also be found at the POeTiSA project web page at https://sites.google.com/icmc.usp.br/poetisa/.

Acknowledgements

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.

[13] `https://sites.google.com/icmc.usp.br/poetisa/resources-and-tools`
[14] `https://sites.google.com/icmc.usp.br/poetisa/porttinari-2-0`

References

Akbik, A., Chiticariu, L., Danilevsky, M., Li, Y., Vaithyanathan, S., and Zhu, H. (2015). Generating high quality proposition Banks for multilingual semantic role labeling. In Zong, C. and Strube, M., editors, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 397–407, Beijing, China. Association for Computational Linguistics.

Barbosa, B. K. d. S. (2024). Descrição sintático-semântica de nomes predicadores em tweets do mercado financeiro em português. Msc dissertation, Universidade Federal de São Carlos (UFSCar), São Carlos, Brazil.

Bhatt, R., Narasimhan, B., Palmer, M., Rambow, O., Sharma, D., and Xia, F. (2009). A multi-representational and multi-layered treebank for Hindi/Urdu. In Stede, M., Huang, C.-R., Ide, N., and Meyers, A., editors, *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 186–189, Suntec, Singapore. Association for Computational Linguistics.

Borba, F. d. S. (2002). *Dicionário de usos do português do Brasil*.

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

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., das Graças Nunes, M., and Barbosa, B. (2024a). A dependency treebank of tweets in brazilian portuguese: Syntactic annotation issues and approach. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 192–201, Porto Alegre, RS, Brasil. SBC.

Di-Felippo, A., Postali, C., Ceregatto, G., Gazana, L., Silva, E., Roman, N., and Pardo, T. (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.

Di-Felippo, A., Roman, N., Barbosa, B., and Pardo, T. (2024b). Genipapo - a multigenre dependency parser for brazilian portuguese. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 257–266, Porto Alegre, RS, Brasil. SBC.

Di-Felippo, A., Roman, N. T., Pardo, T. A. S., and Panta de Moura, L. (2022). The dantestocks corpus: An analysis of the distribution of universal dependencies-based part of speech tags.

Duran, M. S. and Aluísio, S. (2015). Automatic generation of a lexical resource to support semantic role labeling in Portuguese. In Palmer, M., Boleda, G., and Rosso, P., editors, *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 216–221, Denver, Colorado. Association for Computational Linguistics.

Duran, M. S. and Aluísio, S. M. (2011). Propbank-br: a Brazilian Portuguese corpus annotated with semantic role labels. In *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*.

Duran, M. S., Martins, J. P., and Aluísio, S. M. (2013). Um repositório de verbos para a anotação de papéis semânticos disponível na web (a verb repository for semantic role labeling available in the web) [in Portuguese]. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*.

Fleiss, J. L. (1981). The measurement of interrater agreement. In *Statistical Methods for Rates and Proportions*, pages 212–236. John Wiley, New York, 2nd edition.

Haverinen, K., Kanerva, J., Kohonen, S., Missilä, A., Ojala, S., Viljanen, T., Laippala, V., and Ginter, F. (2015). The finnish proposition bank. *Language Resources and Evaluation*, 49(4):907–926.

Jindal, I., Rademaker, A., Ulewicz, M., Linh, H., Nguyen, H., Tran, K.-N., Zhu, H., and Li, Y. (2022). Universal Proposition Bank 2.0. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1700–1711, Marseille, France. European Language Resources Association.

Jurafsky, D. and Martin, J. H. (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd edition. Online manuscript released August 20, 2024.

Li, X., Chen, H., Liu, C., Li, J., Zhang, M., Yu, J., and Zhang, M. (2025). Llms can also do well! breaking barriers in semantic role labeling via large language models.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Meyers, A. (2007). Annotation guidelines for nombank - noun argument structure for propbank. Technical report, Tech Report – New York University.

Mirzaei, A. and Moloodi, A. (2016). Persian Proposition Bank. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3828–3835, Portorož, Slovenia. European Language Resources Association (ELRA).

Moeller, S., Wagner, I., Palmer, M., Conger, K., and Myers, S. (2020). The Russian PropBank. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5995–6002, Marseille, France. European Language Resources Association.

Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., and Zeman, D. (2020). Universal Dependencies v2: An evergrowing multilingual treebank collection. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

O'Gorman, T., Pradhan, S., Palmer, M., Bonn, J., Conger, K., and Gung, J. (2018). The new Propbank: Aligning Propbank with AMR through POS unification. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Palmer, Martha, Ryu, Shijong, Choi, Jinyoung, Yoon, Sinwon, and Jeon, Yeongmi (2006). Korean propbank.

Pardo, T., Duran, M., Lopes, L., Felippo, A., Roman, N., and Nunes, M. (2021). Porttinari - a large multi-genre treebank for brazilian portuguese. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 1–10, Porto Alegre, RS, Brasil. SBC.

Pradhan, S., Bonn, J., Myers, S., Conger, K., O'gorman, T., Gung, J., Wright-bettner, K., and Palmer, M. (2022). PropBank comes of Age—Larger, smarter, and more diverse. In Nastase, V., Pavlick, E., Pilehvar, M. T., Camacho-Collados, J., and Raganato, A., editors, *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 278–288, Seattle, Washington. Association for Computational Linguistics.

Pradhan, S., Moschitti, A., Xue, N., Ng, H. T., Björkelund, A., Uryupina, O., Zhang, Y., and Zhong, Z. (2013). Towards robust linguistic analysis using OntoNotes. In Hockenmaier, J. and Riedel, S., editors, *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

Scandarolli, C. L., Di-Felippo, A., Roman, N. T., and Pardo, T. A. S. (2023). Tipologia de fenômenos ortográficos e lexicais em cgu: o caso dos tweets do mercado financeiro. In *Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana - STIL*. SBC.

Silva, E., Pardo, T., Roman, N., and Di-Fellipo, A. (2021). Universal dependencies for tweets in brazilian portuguese: Tokenization and part of speech tagging. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 434–445, Porto Alegre, RS, Brasil. SBC.

Taylor, A., Marcus, M. P., and Santorini, B. (2003). The penn treebank: An overview. In Abeillé, A., editor, *Treebanks: Building and Using Parsed Corpora*, pages 5–22. Springer, Dordrecht.

Voskaki, R., Tziafa, E., and Annidou, K. (2016). Description of predicative nouns in a modern greek financial corpus. In *Selected Papers of the 21st International Symposium on Theoretical and Applied Linguistics (ISTAL)*, pages 488–503.

Şahin, G. G. and Adalı, E. (2017). Annotation of semantic roles for the turkish proposition bank. *Language Resources and Evaluation*, 52(3):673–706.