Social Inequality in Twitter Data and Machine Readability Enhancement Using Framester's Semantic Frames

Melis Nur Verir

Vrije Universiteit Amsterdam
m.n.verir@student.vu.nl

Abstract. The analysis of social media texts is a critical area of research, and the richness of language is a key consideration in this regard. Nonetheless, the intricate nature of language poses a considerable challenge for present-day natural language processing techniques, including Large Language Models, as they may lack the capability to accurately disentangle the diverse meanings embedded within a given sentence. Therefore, semantic framework analysis plays a crucial role in enhancing the accuracy of text analysis. The present master thesis aims to enhance the analysis of diverse perspectives and political narratives in social media texts by leveraging semantic framework analysis. The analysis involves the extraction of semantic frameworks from PropBank data utilizing Fluid Construction Grammar. Afterward, a SPARQL query will be utilized to access Framester, a knowledge graph that integrates multiple linguistic resources such as Propbank, FrameNet, WordNet, and VerbNet. Based on the information obtained, a knowledge graph will be generated to present the findings.

Keywords: FrameNet · Framester · Knowledge Graph · Social Texts · SPARQL · FCG Editor.

1 Introduction

Social texts play a key role in shaping public opinion and understanding societal dynamics. Analyzing these texts is crucial for gaining insights into diverse perspectives and narratives prevalent in today's digital age. However, the complexity and richness of language pose significant challenges to accurately decipher the nuanced meanings within these texts. Contemporary advancements in natural language processing techniques, such as Large Language Models (LLMs), signify a notable transition toward data-driven methodologies. However, these techniques often encounter challenges when it comes to comprehensively capturing the intricate layers of meaning present within a singular sentence. Despite the considerable achievements of Large Language Models (LLMs), recent scholarly papers have presented empirical evidence suggesting that LLMs occasionally fall short in capturing essential aspects of linguistic meaning [17].

The effectiveness of LLMs depends on extensive training data that is accurately labeled to identify and assign semantic roles. However, these models may encounter difficulties in recognizing and labeling novel or uncommon linguistic constructions due to a lack of sufficient training data [17]. Additionally, their limited grasp of deep linguistic understanding can restrict their capacity to fully comprehend the meaning of sentences or words. Consequently, the presence of word ambiguity can potentially lead LLMs to generate erroneous frame semantics [17]. Consequently, there is a growing need for advanced methods that can enhance the accuracy of text analysis by effectively dealing with the complicatedness of language.

The phenomenon of Twitter exemplifies this challenge. The task of deriving valuable information from Twitter encounters difficulties due to the prevalence of opinionated and non-factual content within tweets [1]. Additionally, automated posts from external sources lack relevance when it comes to conducting searches. Another obstacle in extracting knowledge from Twitter is the limited effectiveness of current semantic processing and parsing models, which struggle to handle the unique characteristics of Twitter data [1].

Semantic framework analysis, which provides a structured representation of the underlying meaning, offers a promising approach to tackling the challenges associated with analyzing social texts. By utilizing semantic frameworks researchers can delve deeper into the complex relationships and contexts within a text. This framework-based analysis enables a more nuanced understanding of the diverse perspectives and narratives expressed in social media texts. This study focuses on the extraction of semantic frameworks from PropBank data by employing Fluid Construction Grammar (FCG). PropBank is a widely utilized resource that provides semantic role annotations for verbs, facilitating the identification of key events and their relationships within sentences [2]. FCG enables the effective extraction and representation of the semantic frameworks found in social texts.

The utilization of FCG Editor tackles the intricacy involved in assigning semantic roles to verbs that exhibit multiple meanings or senses, necessitating distinct sets of arguments. One such verb illustrating this complexity is *leave*, which encompasses diverse senses and assumes varying arguments depending on its contextual usage. Specifically, *leave* can denote both physically moving away from a location and the act of giving something. To accommodate these diverse interpretations, PropBank employs the differentiation of frame sets or senses for the verb, assigning unique argument labels like *leave.01* and *leave.02* to each frame set [10].

Furthermore, Framester, a comprehensive information repository that offers a detailed inventory of lexical units and their associated frames, will be accessed to enrich the analysis[3]. Framester serves as a valuable resource for preserving and organizing information about Frame Element inheritance, originally found in FrameNet. Furthermore, it facilitates a mapping to generic frame elements, establishing connections to a more abstract role hierarchy within the Framester framework [11]. In FrameNet, frames are utilized to describe a wide range of situations, states, or actions, with each frame containing semantic roles known as frame elements.[11] These frame elements can be evoked by Lexical Units (LUs) belonging to different parts of speech. LUs, which are linguistic predicates expressing the situation described by the frame, share semantic arguments within the same frame.

In the sentence *Mobile clinics have provided medical treatment to homeless individuals*, the role of *Supplier* is undertaken by *mobile clinics*, which are responsible for providing medical treatment. Another important Lexical Unit in the sentence is *provided*, which corresponds to the role *Theme* that the *Supplier* provides to *Recipient*. In the context of medical treatment, *provided* signifies the action of delivering or offering medical services. The term *medical treatment* represents the core activity being performed, encompassing a range of healthcare services aimed at addressing the health needs of homeless individuals. Finally, the *Recipient* refers to *homeless individuals* experiencing third-degree burns. They are the recipients of the medical treatment provided by the mobile clinics, seeking relief from their injuries and striving towards recovery.

For instance, within the frame of *Supply*, there exist various lexical units such as *provide*, *equipment*, *supplier*, and *issue*. These units share semantic arguments like Supplier, Theme, and Recipient, indicating the roles and relationships involved in the act of providing [12].

The utilization of a SPARQL query allows for the exploration of the extensive resources available in Framester, facilitating the acquisition of valuable insights that enhance the comprehension of analyzed social texts. Incorporating Framester's information into the analysis of our semantic framework contributes to a more comprehensive contextual understanding, enabling the precise identification of underlying semantic structures.

In the field of artificial intelligence (AI) research, the acquisition and representation of human knowledge play a vital role [16]. Unlike humans who possess the innate ability to comprehend and analyze their environment, AI systems necessitate additional knowledge to acquire similar capabilities and tackle intricate tasks in real-world situations. [13]

A knowledge graph serves as a standard dataset that encapsulates real-world facts and semantic relationships in the structure of triplets [16]. These triplets, when organized as a graph with entities as nodes and relations as edges, form what is commonly referred to as a knowledge graph [15]. The nodes within these knowledge graphs represent the entities of interest, while the edges depict the connections or relationships that exist between these entities [14]. The efficacy of knowledge graphs in representing intricate information has garnered significant attention from both academia and industry in recent years.

Apart from transforming data into machine-interpretable knowledge, a knowledge graph offers a range of additional advantages. One key benefit is Data Integration, which involves meaningful knowledge management that integrates real-world data and its associated context from diverse sources [16]. This integration occurs within a formal and interconnected model that supports semantic similarity, reasoning, and query expansion. Consequently, the system becomes capable of retrieving more relevant items and enhances its interpretability.

Furthermore, knowledge graph-based information retrieval yields high search efficiency by leveraging advanced item representations to significantly reduce the search space, thereby improving overall efficiency [16]. Lastly, a knowledge graph serves as an accurate visual representation of the flow of facts between data entities within the graph. The visualization capability of knowledge graphs proves highly valuable for identifying problematic areas and discovering patterns over time.

The integration of knowledge graphs in social text analysis aligns with the overarching thesis, as it enhances the understanding and interpretation of social texts, enabling more comprehensive language-structured analysis and uncovering deeper insights and nuances that would otherwise go unnoticed.

1.1 Research Question

The analysis of social texts poses a distinct technical and scientific challenge, encompassing the intricacies of language, the challenges in capturing provenance, and the subtle nuances inherent in human communication. This problem is important because social texts play a significant role in shaping opinions, attitudes, and behaviors in society. Understanding and interpreting social texts accurately is crucial for various applications, such as sentiment analysis, opinion mining, and information retrieval.

Currently, researchers have explored different methodologies and resources to enhance the comprehension and interpretation of social texts. However, there is a pressing need to investigate the effectiveness of integrating semantic frameworks, leveraging resources like Framester, and employing knowledge graphs to advance the field of social text analysis. These approaches have the potential to provide deeper insights into the semantic relationships within social texts, enabling more accurate and comprehensive language-structured analysis.

> *Incorporating semantic frameworks and utilizing Framester's resources in social text analysis will result in improved accuracy and a nuanced understanding of the complex dynamics and underlying meaning within social texts*

The first hypothesis proposes that integrating semantic frameworks and utilizing Framester's resources in social text analysis can result in enhanced accuracy and a more profound comprehension of the intricate dynamics and underlying meaning within social texts. Drawing upon Framester's comprehensive lexical database and integrating semantic frameworks enables researchers to uncover deeper insights and nuances that might otherwise go unnoticed. This hypothesis implies that the integration of these resources contributes to a more exhaustive contextual understanding of social texts.

> *The utilization of a knowledge graph generated from the semantic framework analysis and the Framester query will facilitate a coherent and visually engaging representation of complex interconnections and relationships among concepts, frames, and perspectives in social texts, leading to a more intuitive comprehension of the analyzed data.*

The hypothesis suggests that utilizing a knowledge graph created through the integration of semantic framework analysis and the Framester query allows for a visually captivating representation of intricate interconnections and relationships among concepts, frames, and perspectives in social texts. This approach enables researchers to organize and visually map the complex web of information, leading to a more intuitive understanding of the analyzed data. By employing a knowledge graph, researchers can gain an accessible and insightful understanding of the relationships and interdependencies within social texts.

To investigate these hypotheses and advance the field of social text analysis, research questions have been formulated to guide the inquiry.

> *How does the utilization of Framester, a comprehensive information repository, contribute to a more comprehensive contextual understanding of social texts?*

First, the exploration focuses on how the utilization of Framester, a comprehensive information repository, contributes to a more comprehensive contextual understanding of social texts. By leveraging Framester's resources, the potential benefits and limitations of utilizing this repository to improve the analysis and interpretation of social texts are sought.

> *What are the benefits of employing a knowledge graph to visually represent the complex interconnections and relationships among concepts, frames, and perspectives in social texts?*

Next, the investigation delves into the benefits of employing a knowledge graph to visually represent the complex interconnections and relationships among concepts, frames, and perspectives in social texts. By exploring the advantages and potential applications of knowledge graphs, the comprehension, and interpretation of social texts are aimed to be enhanced by providing a comprehensive overview of semantic relationships and contextual connections.

These research questions and hypotheses lay the foundation for the investigation into the effectiveness of semantic framework analysis, the utilization of Framester's resources, and the integration of knowledge graphs in enhancing social text analysis. Through this research, contributions are aimed to be made toward the advancement of the field, providing valuable insights that benefit various domains such as social sciences, political studies, and media analysis.

2 Related Work

In this section, we discuss relevant studies that fall under two main topics: Twitter and Semantic Frame, and Semantic Frame and Web Technology. Each topic is presented with its respective subsection, and a brief conclusion is provided at the end of each subsection, highlighting the connection to our research and the gaps addressed.

Sentilo[4] generates its output in the form of an RDF graph and strives to identify the owners and subjects on Linked Data whenever feasible. The main focus of this study revolves around sentic computing and provides a comprehensive explanation of this concept. The study aims to elevate the analysis from individual ideas and words to a higher level of concept-based opinion analysis. This approach represents a novel and interdisciplinary method in Semantic Analysis, emphasizing the significance of integrating semantic features into intuition and opinion mining. The study demonstrates that incorporating semantic features enriches the performance of the Semantic Analysis algorithm.

Our work shares a similarity with this article in terms of utilizing RDF graphs as a formal representation of ideas, employing an ontology that defines the core concepts and relationships associated with social texts. However, while the article solely focuses on social texts from a Sentiment Analysis perspective and employs graphics for analysis, our study enriches all social texts with a semantic frame, expanding the scope beyond sentiment analysis.

The second article[5] examines the detection of figurative language in social media, with a specific focus on employing semantic features to identify irony and sarcasm. Similar to our study, this article utilizes semantic features to enhance tweet representations by incorporating event information, frames, and lexical meanings in addition to lexical units. The article's main objective is to identify tweets that contain figurative language expressions by analyzing a dataset consisting of both regular and sarcastic tweets. The findings highlight the significance of frames as crucial indicators for determining the subjectivity of a text, particularly in the context of figurative language. Furthermore, the use of semantic features demonstrates improved classification accuracy across various combinations, except when solely relying on semantic frames.

In the article[1], a comprehensive investigation is conducted using four distinct methodologies to extract factual information concerning 60 assets from the Freebase database. The study evaluates these approaches across various dimensions and reveals that by employing accurate syntactic analysis, precise and pertinent information can be extracted, surpassing the existing knowledge base. Nevertheless, the evaluation reveals that, in the majority of systems, two out of three triads were deemed nearly incomprehensible due to deficiencies in part-of-speech (POS) tagging and dependency parsing. The research findings highlight that the utilization of frame semantics yields more resilient outcomes and significantly reduces errors caused by ambiguity, surpassing a 20% improvement. Notably, a noteworthy distinction between this study and ours is the employment of a semantic role tagger provided by MATE-TOOLS, considering solely the ARG0, ARG1, and ARG2 relationships in the extraction process.

In this article[6], the authors develop a tool called Apollo, which performs sentiment analysis and frame detection on Twitter feeds related to the COVID-19 pandemic. They employ Word Sense Disambiguation (WSD) and connect tweets to various linked data sources. While our study diverges in focus, their utilization of semantic techniques and sentiment analysis provides valuable insights for our research.

This work[7] combines Semantic Web technologies with FrameNet, emphasizing the benefits of applying Semantic Web technologies in the context of FrameNet. It aims to enhance the documentation process, enable frame-based searches, and provide tools for assisting framework documents and sentence annotations. Our research complements this work by incorporating semantic frames into the extraction of knowledge from Twitter.

This article[8] introduces a double-graph framework for frame semantic parsing, creating a heterogeneous graph that encompasses both frames and Framework Elements (FEs). Their proposed framework knowledge graph facilitates incremental graphing, strengthening interactions between subtasks and relationships between arguments. Although our research focuses on knowledge extraction from Twitter, we can draw inspiration from their graph-based approach.

This research[9] focuses on TakeFive, a semantic role-tagging technique used to transform text into a frame-based infographic, while also conducting dependency parsing. The article acknowledges a limitation of knowledge graphs, namely the absence of contextual and situational information, which poses challenges to achieving interoperability. The study's key finding suggests that NLP evaluation settings may prove insufficient in accurately measuring the absolute performance of intricate semantic tasks, such as Semantic Role Labeling. As a result, the research proposes the establishment of criteria and raters for evaluating knowledge graphs, along with a reevaluation of the design of role ontologies, particularly in the context of Semantic Role Labeling. Our research aligns with this study by considering semantic role labeling for knowledge extraction, and we aim to address similar challenges in incorporating contextual information in our analysis.

3 Methods

3.1 Data

Data Collection The dataset used in this thesis consists of a collection of textual data sourced from Twitter, publicly shared by users. It is important to acknowledge that the specific selection criteria and associated limitations regarding these sources are subject to further investigation.

The dataset utilized in this study was obtained from the project titled "Narrative networks for understanding tweets about inequality," implemented by Meaning and Understanding in Human-centric AI (MUHAI). This project received funding from the European Union's Horizon 2020 research and innovation program under the grant agreement numbered 951846. The dataset was acquired through a CURL request to Twitter's API V2.

The selection and retrieval process of inequality-related tweets were carried out programmatically by querying the Twitter Full Archive and Stream API

(V2) endpoints. This involved using inequality-related keywords and phrases defined in the multilingual inequality dictionary to programmatically download large volumes of tweets.

Data Preprocessing and Limitations The dataset comprises a total of 448,227 rows, encompassing a rich collection of data. It consists of 16 attributes that provide valuable insights for the analysis of the tweets under investigation.

One crucial attribute is the ”entities” attribute, which contains mentions and dictionaries. These dictionaries include essential details such as the name, ID number, hashtag, and URL of the mentioned users. The ”text” attribute follows, housing the complete text of each tweet, and providing the primary content for analysis. The ”id” attribute serves as a unique identification number assigned to each tweet, facilitating individual tweet referencing and tracking. Alongside, the ”created_at” attribute denotes the specific date when each tweet was posted, offering a temporal dimension for analysis. To gauge engagement and interaction, the ”public_metrics” attribute records important counts such as retweets, replies, likes, and quotes as a dictionary, shedding light on the level of audience engagement with the tweets. The ”author_id” attribute contains the unique user identification number associated with each tweet, enabling author-specific analysis. Understanding the source of the tweets is vital, and thus, the ”Source” attribute indicates the platform or application from which each tweet was sent. Moreover, the ”edit_history_tweet_ids” attribute denotes the edit dates of the tweets, providing insights into the evolving nature of the content.

The dataset also includes attributes such as ”referenced_tweets” and ”conversation_id,” which aid in tracking the tweets within a conversation and understanding the context of the communication. The ”lang” attribute indicates the language in which the tweets are written, facilitating language-specific analysis. The ”in_reply_to_user_id” attribute contains the user identification number of the tweets sent as replies to other users, enabling the examination of user interactions. The presence of URLs within the tweets is captured by the ”url” attribute, allowing for analysis related to shared links. Considering the sensitivity of some content, the ”possibly_sensitive” attribute highlights whether a tweet may contain media or content classified as sensitive. The ”geo” attribute provides coordinates if available, potentially enabling spatial analysis of the tweets. Media content attached to the tweets is represented by the ”attachments” attribute, while the ”withheld” attribute provides information about any content barriers associated with the tweets.

Collectively, these attributes contribute to a comprehensive dataset, providing a foundation for in-depth analysis and exploration within the scope of this academic thesis.

The utilization of Twitter data in this academic thesis is accompanied by several inherent limitations that should be acknowledged to ensure a comprehensive understanding of the findings. These limitations encompass the following aspects:

Firstly, it is important to recognize that the dataset used for this study comprises a selection of publicly shared tweets on Twitter. Consequently, the findings may not fully capture the entire Twitter user population, potentially leading to sampling bias. The dataset represents only those users who have chosen to make their tweets publicly available, which may not accurately represent the perspectives and behaviors of private or inactive users. Secondly, the dataset includes tweets from a diverse range of users, including individuals, and organizations. Each category of users may exhibit distinct behaviors, intentions, and characteristics, introducing complexity and requiring caution when making generalizations based on the data. Thirdly, it is crucial to note that the dataset constitutes a snapshot of tweets collected during a specific time frame. Consequently, it does not encompass the complete history of interactions or account activities. Moreover, any tweets that have been deleted or made private subsequent to the data collection phase will not be accounted for. Additionally, modifications or edits made to the tweets after data collection are not reflected in the dataset.

Moreover, it is crucial to acknowledge that the dataset may contain biased or false information. Twitter has been recognized as a platform susceptible to the dissemination of misinformation, rumors, and biased content. Lastly, ethical considerations regarding privacy and consent should be taken into account. While the dataset comprises publicly available tweets, it is essential to handle users' personal information and identifiable details in compliance with relevant data protection regulations and ethical guidelines.

Methodology The methodology proposed for this study aims to optimize the analysis of diverse perspectives and narratives in social texts by leveraging semantic framework analysis. The first step involves the collection of a representative dataset of social texts that encompass a wide range of topics and include an adequate number of texts for analysis. Afterward, steps such as Topic Labeling and Sentiment Analysis are applied to enrich the text on the collected data and represent its relations with each other. The purpose of these applications is to increase the machine readability of the data enriched with semantic frames.

To enrich the analysis further, access to Framester, a lexical-semantic resource containing a vast repository of annotated frames and their semantic information, will be gained through a SPARQL query. This query will retrieve relevant frame information based on the extracted semantic frameworks, enhancing the understanding of the meaning and context within the tweets.

The next step involves the generation of a knowledge graph that visually represents the relationships between the extracted semantic frameworks and the associated frames from the Framester. The knowledge graph will provide a structured and intuitive representation of the findings, with nodes representing semantic frameworks and edges depicting the connections between the frameworks and relevant frames. This knowledge graph will serve as a valuable tool for analyzing and interpreting the results.

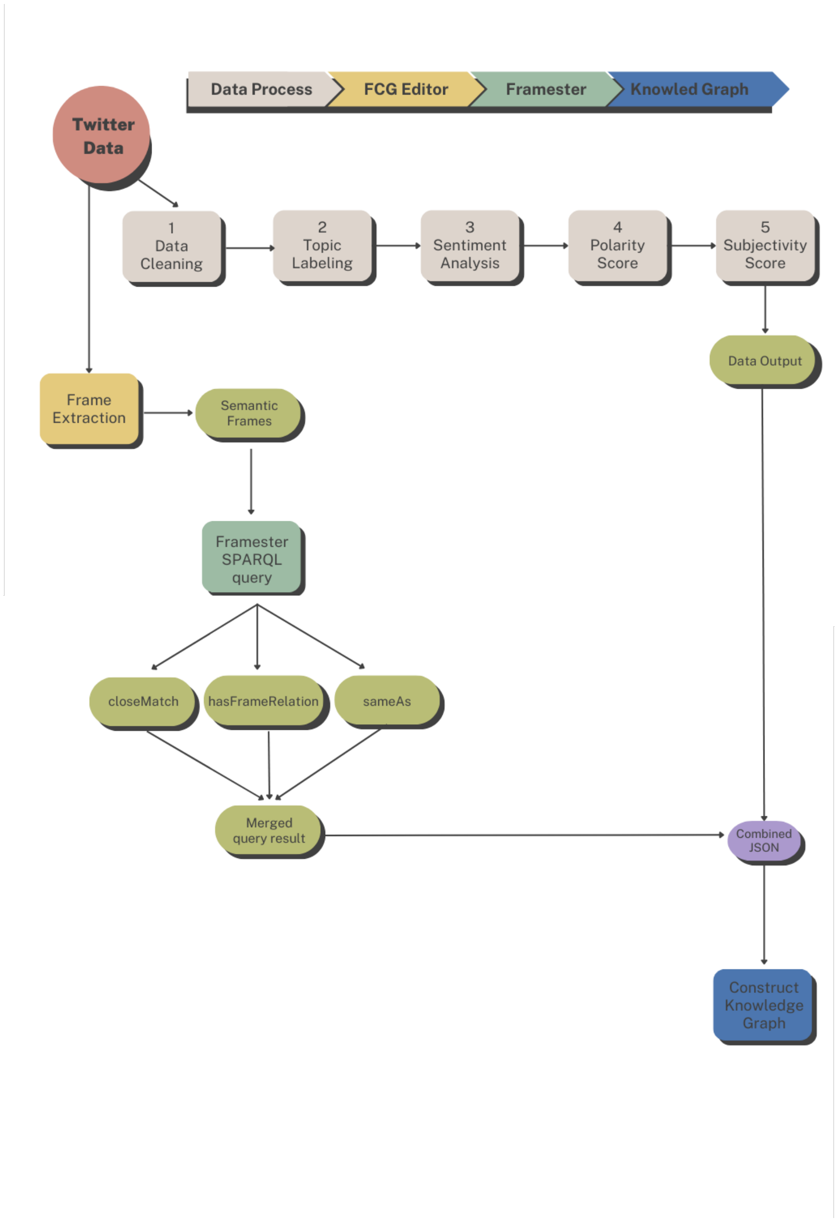

Fig. 1. Thesis plan

4 Results

5 Conclusion

6 Discussion

References

1. Søgaard, A., Plank, B., Alonso, H. (2015). Using Frame Semantics for Knowledge Extraction from Twitter. Proceedings of the AAAI Conference on Artificial Intelligence, 29(1) https://doi.org/10.1609/aaai.v29i1.9524
2. Palmer, M., Gildea, D., Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. Computational Linguistics, 31(1), 71-106.
3. Baker, C., Fillmore, C., Lowe, J. (1998). The Berkeley FrameNet Project. In 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1 (pp. 86–90). Association for Computational Linguistics.
4. Reforgiato Recupero, D., Presutti, V., Consoli, S., Gangemi, A., Nuzzolese, A. G. (2015). Sentilo: frame-based sentiment analysis. Cognitive Computation, 7, 211-225.
5. Recupero, D. R., Alam, M., Buscaldi, D., Grezka, A., Tavazoee, F. (2019). Frame-based detection of figurative language in tweets [application notes]. IEEE Computational Intelligence Magazine, 14(4), 77-88.
6. Alam, M., Kaschura, M., Sack, H. (2020). Apollo: Twitter Stream Analyzer of Trending Hashtags: A case-study of# COVID-19. In ISWC (Demos/Industry) (pp. 64-69).
7. Hauck, P., Villela, R. M. M. B., Campos, F., Torrent, T., Matos, E. E., David, J. M. N. (2015). Supporting FrameNet Project with Semantic Web Technologies. In ONTOBRAS.
8. Zheng, C., Chen, X., Xu, R., Chang, B. (2022). A Double-Graph Based Framework for Frame Semantic Parsing. arXiv preprint arXiv:2206.09158.
9. Alam, M., Gangemi, A., Presutti, V., Reforgiato Recupero, D. (2021). Semantic role labeling for knowledge graph extraction from text. Progress in Artificial Intelligence, 10, 309-320.
10. Babko-Malaya, Olga. "Guidelines for Propbank framers." Unpublished manual, September (2005).
11. Gangemi, A., Alam, M., Asprino, L., Presutti, V. and Recupero, D.R., 2016. Framester: A wide coverage linguistic linked data hub. In Knowledge Engineering and Knowledge Management: 20th International Conference, EKAW 2016, Bologna, Italy, November 19-23, 2016, Proceedings 20 (pp. 239-254). Springer International Publishing.
12. Pennacchiotti, M., De Cao, D., Basili, R., Croce, D. and Roth, M., 2008, October. Automatic induction of FrameNet lexical units. In Proceedings of the 2008 conference on empirical methods in natural language processing (pp. 457-465).
13. Ji, S., Pan, S., Cambria, E., Marttinen, P. and Philip, S.Y., 2021. A survey on knowledge graphs: Representation, acquisition, and applications. IEEE transactions on neural networks and learning systems, 33(2), pp.494-514.
14. Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G.D., Gutierrez, C., Kirrane, S., Gayo, J.E.L., Navigli, R., Neumaier, S. and Ngomo, A.C.N., 2021. Knowledge graphs. ACM Computing Surveys (Csur), 54(4), pp.1-37.

15. Cheng, D., Yang, F., Xiang, S. and Liu, J., 2022. Financial time series forecasting with multi-modality graph neural network. Pattern Recognition, 121, p.108218.
16. Peng, C., Xia, F., Naseriparsa, M. and Osborne, F., 2023. Knowledge graphs: Opportunities and challenges. Artificial Intelligence Review, pp.1-32.
17. Asher, N., Bhar, S., Chaturvedi, A., Hunter, J. and Paul, S., 2023. Limits for Learning with Language Models. arXiv preprint arXiv:2306.12213.