VerbaNexAI Lab at SemEval-2024 Task 1: A Multilayer Artificial Intelligence Model for Semantic Relationship Detection

Anderson Morillo and **Daniel Peña**
Juan Carlos Martinez-Santos and **Edwin Puertas**
Universidad Tecnologica de Bolivar, Cartagena Colombia
epuerta@utb.edu.co

Abstract

This paper presents an artificial intelligence model designed to detect semantic relationships in natural language, addressing the challenges of SemEval 2024 Task 1. Our goal is to advance machine understanding of the subtleties of human language through semantic analysis. Using a novel combination of convolutional neural networks (CNNs), long short-term memory (LSTM) networks, and an attention mechanism, our model is trained on the STR-2022 dataset. This approach enhances its ability to detect semantic nuances in different texts. The model achieved an 81.92% effectiveness rate and ranked 24th in SemEval 2024 Task 1. These results demonstrate its robustness and adaptability in detecting semantic relationships and validate its performance in diverse linguistic contexts. Our work contributes to natural language processing by providing insights into semantic textual relatedness. It sets a benchmark for future research and promises to inspire innovations that could transform digital language processing and interaction.

1 Introduction

The analysis of semantic relationships in natural language is considered an essential pillar for understanding the inherent complexity of textual communication Wolfe et al. (2005). With the increasing application of artificial intelligence (AI) models in natural language processing, the ability to discern semantic similarity between text fragments has become a fundamental challenge due to the complexity of natural language and the diversity of meanings that words and phrases can have in different contexts Zunino (2023). In this context, Semantic Textual Relatedness (STR) is a crucial element in natural language understanding, gaining increasing significance with integrating artificial intelligence (AI) models in language processing.

This article aligns with the objectives set by SemEval 2024 Task 1, a pivotal challenge centered on predicting semantic textual relationships between sentence pairs in the English language. The task's importance lies in its profound impact on advancing contextual language understanding, a cornerstone for AI applications across diverse domains.

Our approach to tackle this challenge involves a four-layer feature extraction process. The first layer focuses on extracting lexical similarity, providing a foundation for understanding semantic connections based on word usage. Subsequently, the second layer delves into capturing knowledge-oriented similarity and incorporating domain-specific insights into the model. The third layer concentrates on Corpus-oriented features, considering the contextual influence of larger text corpora. Finally, in the fourth layer, we employ an Embedding approach. Here, we train a Long Short-Term Memory (LSTM) model, extracting sentence features from phoneme embeddings and a sentence transformer model, thereby capturing nuanced semantic nuances.

Throughout the SemEval 2024 Task 1 competition (Ousidhoum et al., 2024b), our system secured the 24th position out of 36 teams, achieving a competitive score of 0.8192. Notably, our system was designed and optimized for the English language. To foster transparency and collaboration, we have released our code, accessible at https://github.com/VerbaNexAI/SemEval2024.

2 Related Work

The analysis of semantic relations is considered fundamental for understanding the connection of meanings between words, phrases, and sentences in a text. Various relationships, such as synonymy, antonymy, hyperonymy, meronymy, and cohyponymy, can manifest in this context. Two main approaches have addressed this field: rule-based and machine learning-based approaches.

Rule-based approaches use ontologies, struc-

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1344–1350

tures that define concepts and their relationships, and semantic networks, which are graphical representations of these relationships. Lexical patterns, which are rules that describe semantic relationships based on the structure of words, have also been used. On the other hand, approaches based on machine learning have gained relevance, utilizing technologies such as convolutional neural networks (CNN), recurrent neural networks (RNN), attention models, and word embeddings.

In the literature, we can find several methods for semantic relation detection. In "Learning short-text semantic similarity with word embeddings and external knowledge sources" Nguyen et al. (2019), authors propose an approach that uses word embeddings and external knowledge to measure semantic similarity between short texts, managing to outperform traditional methods on diverse datasets.

Another significant work is "A multi-layer system for semantic relatedness evaluation" Gomaa (2019), which presents a multi-layer system for semantic relatedness evaluation between sentences, combining various similarity features and achieving promising accuracy on the SICK dataset.

In addition, "A New Methodology for Computing Semantic Relatedness: Modified Latent Semantic Analysis by Fuzzy Formal Concept Analysis" Jain et al. (2020) proposes a hybrid methodology that combines latent semantic analysis and fuzzy formal concept analysis to compute the semantic relatedness between words and sentences, obtaining improved results compared to other baseline measures on a specific corpus.

In the field of language-specific semantic relatedness detection, "Sentence Embedding and Convolutional Neural Network for Semantic Textual Similarity Detection in Arabic Language" Mahmoud and Zrigui (2019a) proposes a deep learning-based approach to detect paraphrases in Arabic, using word2vec and a convolutional neural network to overcome traditional methods and other stylometric feature-based approaches.

Finally, "Attention-based model for predicting question relatedness on Stack Overflow" Pei et al. (2021) introduces a deep learning model called ASIM, which uses the attention mechanism to predict the semantic relationship between questions in programming question and answer websites. This model outperforms previous models in terms of performance and generalization in detecting duplicate questions and predicting the relationship between knowledge units.

Despite these advances, knowledge gaps persist. Most models focus on semantic relationships at the word or phrase level. However, we need more research in sentence- and paragraph-level relationship detection. In addition, we require more robust models to adapt to different domains and text types, ensuring a more complete and accurate understanding of semantic relations in natural language processing.

3 System Overview

This section outlines our proposed model for tackling the task presented in SemEval 2024, Track A, which involves assessing the semantic relationship between pairs of sentences. Initially, the text data undergoes preprocessing, including separating sentence pairs, followed by training a Long Short-Term Memory (LSTM) model on the training dataset. Subsequently, we extract text features based on a four-layer architecture proposed by Gomaa (2019), as illustrated in Figure 1. These layers include word embedding, syntactic relationships, corpus topics, and contextual information.

Additionally, we incorporate novel features to enhance our model's performance. These include:

Senticnet: Utilized to extract the polarity of sentences in the knowledge-oriented layer. Latent Semantic Indexing (LSI) is employed for the corpus-oriented layer to gain insights into the underlying structure of the text corpus.

Phoneme Extraction: A novel approach to capture phonetic information from the sentences.

Furthermore, we integrate an attention mechanism inspired by Vaswani et al. (2017) to effectively capture intricate dependencies within sequences. Leveraging insights from recent advancements, our model incorporates a Part-of-Speech (POS)-aware and layer ensemble transformer, further enhancing its ability to discern semantic relationships.

By drawing from diverse studies on data augmentation, ensemble learning, and transformer-based profiling, our model aims to provide a robust solution for semantic relationship detection. It showcases a comprehensive understanding of attention mechanisms and their integration with state-of-the-art techniques.

3.1 Data Description

We used the dataset STR-2022 proposed by Abdalla et al. (2021) and collected by (Ousidhoum et al., 2024a) for training the system. This dataset

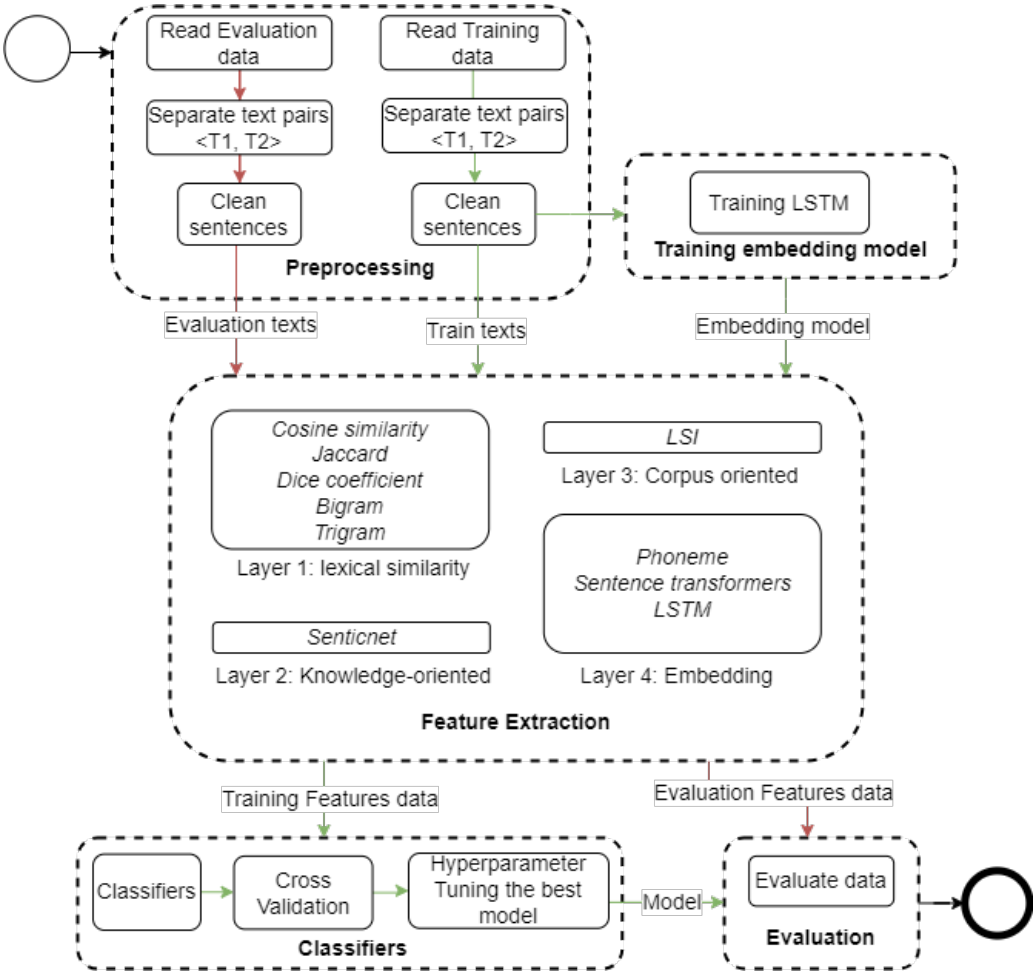

Figure 1: System General Pipeline

comprises 5,500 pairs of sentences in English. This dataset underwent meticulous curation procedures, sampling sentences from various sources, such as social media tweets, book reviews, and paraphrases, to encompass diverse linguistic characteristics and styles. Each pair is labeled with their relatedness score and distribution as shown in Table 1.

Table 1: Frequency distribution of scores intervals

Score intervals	Frequencies
0 - 0.2	502
0.2 - 0.4	1376
0.4 - 0.6	1861
0.6 - 0.8	1149
0.8 - 1	612

3.2 Data Preprocessing

During preprocessing, we separated sentences, and verb and subject decontraction were applied. Subsequently, the model was evaluated with and without lemmatization, as well as with and without stopwords, to assess differences in performance. We removed capitalization, special characters, and numbers as part of the preprocessing process. Data preprocessing is a fundamental step that significantly influences the validity and performance of text classification models, both modern transformers and traditional classifiers Siino et al. (2024). Several preprocessing decisions, such as the treatment of negation, conversion of text to lowercase, application of hyphenation, and consideration of corpus size and document length, are critical to ensure the capture of the true textual meaning and improve the reliability Hickman et al. (2022).

3.3 Training Embedding Models

This part details the training process of a Long Short-Term Memory (LSTM) neural network model for relatedness identification. It begins with data splitting into training and validation sets, followed by message tokenization and constructing a unique vocabulary. We indexed words and applied padding to standardize sequences. We converted

the data into PyTorch tensors and defined a custom dataset and data loaders to handle training batches efficiently.

We defined the model with an embedding layer, an LSTM layer, and a linear output layer. During training, the Adam optimizer and Mean Squared Error (MSE) loss function are utilized, with a loop updating model weights over multiple epochs. For monitoring, we evaluated model performance on the validation set after each epoch.

Finally, upon completion of training, the model and its parameters are saved to a file for future use, enabling its application without the need to retrain it from scratch.

3.4 Feature Extraction

In this section, we explain how the new features are built and used within the text extraction; we try to create a system that could receive the pairs of sentences and return values with consistent output shapes that can feed the model.

3.4.1 Layer 1: String-Oriented Similarity

We based this feature on text extraction, either the characters or the words. It comprises the best features evaluated by Gomaa (2019), Cosine Similarity, Jaccard Similarity, Dice's Coefficient, Bigram, and Trigram.

This layer computes string-oriented similarity features using the following equations:

$$\text{Cosine Similarity} = \frac{A \cdot B}{\|A\|\|B\|} \tag{1}$$

$$\text{Jaccard Similarity} = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

$$\text{Dice's Coefficient} = \frac{2|A \cap B|}{|A| + |B|} \tag{3}$$

where A and B represent the sets of words in both sentences.

3.4.2 Layer 2: Corpus-Oriented Similarity

The system extracts general information using Latent Semantic Indexing (LSI) with the framework Gensim described by Řehůřek and Sojka (2010) to capture the overall thematic similarity between two sentences.

We extract the representation of the average LSI values for the primary five topics, aiming to capture the thematic similarity between the sentences.

3.4.3 Layer 3: Knowledge-Oriented Similarity

This layer is oriented to extract semantic information related to sentiment within the sentences. We used SenticNet proposed by Cambria et al. (2022), a commonsense-based Neurosymbolic framework that extracts the polarity of words. LSI was extracted by averaging the polarities of each sentence and creating a vector with it.

3.4.4 Layer 4: Sentence Embedding

We proposed three forms of word embedding, taking advantage of the good behavior of this layer to evaluate the semantic relationships within two sentences. We used the pre-trained model of sentence transformer Ni et al. (2021), LSTM, and the phonemes embedding. The phoneme embedding works by taking each letter within the sentences, extracting its representation to a phoneme, and returning its representation as a vector.

3.5 Classifiers

We compared the performance of various machine learning models proposed in (Gomaa, 2019) and evaluated the combination of different characteristic types represented by vector inputs.

Random Forest, Gradient Boosting, Multi-layer Perceptron, AdaBoost, and Support Vector Regression (SVR) were employed using the framework Sklearn proposed by Pedregosa et al. (2011), alongside an ensemble voting system combining them. This research aimed to identify the most suitable model(s) for analyzing diverse feature vector inputs.

3.6 Evaluation

The evaluating part of the code serves as a fundamental component within an empirical study focused on assessing the performance of machine learning classifiers. Its primary purpose is to automate the evaluation process, enabling the systematic comparison of various classifiers in a supervised learning context. By implementing k-fold cross-validation, the code ensures robustness and reliability in the evaluation by mitigating potential biases associated with a single train-test split.

Within the evaluating part, we compute a comprehensive suite of performance metrics for each classifier, including Spearman correlation, Mean Squared Error (MSE), R-squared (R^2), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error

(MAPE). These metrics provide a multifaceted assessment of the classifiers' predictive capabilities, accuracy, and robustness.

4 Experimental Setup

The system model utilized sentences without lemmatization and stopword removal, preserving the original form of the sentences to capture a broader range of semantic nuances. Phoneme extraction, as proposed by Del Castillo, was incorporated into the system, utilizing specific components of the provided code. This inclusion aims to enhance the model's sensitivity to phonetic features, contributing to a more comprehensive understanding of textual relationships.

We employ the ShuffleSplit method for cross-validation to assess the model's training. The dataset was split into training and validation sets using a test size 25%, and we utilized ten pairs (n_splits = 10) to ensure robust evaluation. We set the random state to 42 (random_state = 42) for reproducibility. The details of the library versions used in the implementation are provided in Table 2.

Table 2: Python Libraries and Versions

Library	Version
nltk	3.8.1
gensim	4.3.2
spacy	3.7.2
scikit-learn	1.3.2
sentence-transformers	2.2.2
senticnet	1.6
numpy	1.23.4
scipy	1.10.1
matplotlib	3.7.4
seaborn	0.13.0
torch	1.6.0
pandas	2.0.3
epitran	1.24

5 Results

We evaluated the semantic relation detection model using the training and test datasets, and the results are in Table 3. Our model secured the 24th position in the competition ranking, achieving a Spearman correlation coefficient of 0.8192 on the English language dataset. It is relevant to note that this is slightly below the baseline value of 0.83 set as baseline.

Table 3: Ranking of results in framing detection classification

Lang	Spearman Correlation
EN	0.8192

In addition, Table 4 summarizes the performance of the voting system's configuration, and we also present its performance compared to the best individual model.

While the results obtained are below the established baseline, we recognize opportunities for improvement. Abstraction analysis could reveal the specific contributions of each system component to this performance and guide future improvements. It is also crucial to consider the nature of the baseline and the inherent complexity of the task at hand.

6 Limitations

While presenting our approach for evaluating semantic relations between sentences, it's crucial to acknowledge certain limitations that may impact the interpretation and applicability of our proposed model. We outline these limitations below:

- Dataset Representativeness: The STR-2022 dataset, comprising 5,500 English sentence pairs, may not fully capture linguistic diversity and semantic nuances across different languages, limiting the model's generalization to diverse linguistic contexts.
- Preprocessing Impact: Decisions done during preprocessing (such as removing capital letters, special characters, and numbers) could significantly affect semantic representations. When modifying these preprocessing steps, careful consideration is needed to avoid potential bias or information loss.
- Hyperparameter Sensitivity: The model's performance is sensitive to hyperparameter choices, like the number of LSTM layers or the learning rate of the Adam optimizer. Fine-tuning is crucial for optimizing the model's ability to capture semantic relationships effectively.

7 Ethical Considerations

We linked the text similarity field to the detection of paraphrasing (Mahmoud and Zrigui, 2019b), which

Table 4: Correlation of Spearman's Rank between Various Text Preprocessing Methods and Machine Learning Models

Preprocessing	Machine Learning Model	Spearman's Correlation Coefficient
No lemmatized, no stopwords	AdaBoost	0.82
Lemmatized, no stopwords	AdaBoost	0.82
No lemmatized, no stopwords	Gradient Boosting	0.82
Lemmatized, no stopwords	Gradient Boosting	0.82
No lemmatized, no stopwords	Multi-layer Perceptron	0.82
Lemmatized, no stopwords	Multi-layer Perceptron	0.82
No lemmatized, no stopwords	Voting	0.81
Lemmatized, no stopwords	Voting	0.81
No lemmatized, stopwords	Multi-layer Perceptron	0.77
Lemmatized, stopwords	AdaBoost	0.76

can pose an ethical problem when using an author's work without proper citation. Our solution addresses the bias by extracting various text features, from word information to context and vector representation. This way, we can avoid some limitations from training the model with insufficient features.

8 Conclusions

This paper presented a comprehensive approach to evaluating semantic relations between sentences, addressing the challenges posed by SemEval 2024 Task 1. Our model employs a sophisticated four-layered feature extraction technique, encompassing lexical similarity, knowledge orientation, corpus orientation, and embedding layers.

Despite achieving a notable 24th place in the competition, we acknowledge certain limitations, including concerns about dataset representativeness, preprocessing decisions, and hyperparameter sensitivity. These insights serve as valuable lessons for future enhancements in our approach.

While the Spearman correlation of 0.8192 places our model slightly below the established baseline of 0.83, this outcome provides an invaluable learning experience. Moving forward, we plan to conduct ablation studies to dissect the impact of individual components, explore alternative models and preprocessing strategies, and conduct a detailed error analysis to address specific shortcomings.

Ultimately, this work contributes to a deeper understanding of semantic relations and provides a competitive model for SemEval 2024. We are committed to advancing semantic understanding and improving AI systems for natural language processing. The journey from this competition is a stepping stone toward more refined and practical solutions in semantic relationship detection.

Acknowledgments

To the SemEval contest, sponsored by the SIGLEX Special Interest Group on the Lexicon of the Association for Computational Linguistics. To the master's degree scholarship program in engineering at the Universidad Tecnológica de Bolivar (UTB) in Cartagena, Colombia.

We would like to express our gratitude to the team at the VerbaNex AI Lab [1] for their dedication, collaboration, and ongoing support of our research endeavors.

References

Mohamed Abdalla, Krishnapriya Vishnubhotla, and Saif M. Mohammad. 2021. What makes sentences semantically related: A textual relatedness dataset and empirical study. *CoRR*, abs/2110.04845.

Erik Cambria, Qian Liu, Sergio Decherchi, Frank Xing, and Kenneth Kwok. 2022. SenticNet 7: A commonsense-based neurosymbolic AI framework for explainable sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3829–3839, Marseille, France. European Language Resources Association.

Edwin Alexander Puertas Del Castillo. 2023. *Análisis de elementos fonéticos y elementos emocionales para predecir la polaridad en fuentes de microblogging*. Ph.D. thesis, Pontificia Universidad Javeriana, Colombia 9.

WAEL HASSAN Gomaa. 2019. A multi-layer system for semantic relatedness evaluation. *Journal*

[1] https://github.com/VerbaNexAI

of Theoretical and Applied Information Technology, 97(23):3536–3544.

Louis Hickman, Stuti Thapa, Louis Tay, Mengyang Cao, and Padmini Srinivasan. 2022. Text preprocessing for text mining in organizational research: Review and recommendations. *Organizational Research Methods*, 25(1):114–146.

Shivani Jain, KR Seeja, and Rajni Jindal. 2020. A new methodology for computing semantic relatedness: modified latent semantic analysis by fuzzy formal concept analysis. *Procedia Computer Science*, 167:1102–1109.

Adnen Mahmoud and Mounir Zrigui. 2019a. Sentence embedding and convolutional neural network for semantic textual similarity detection in arabic language. *Arabian Journal for Science and Engineering*, 44:9263–9274.

Adnen Mahmoud and Mounir Zrigui. 2019b. Sentence embedding and convolutional neural network for semantic textual similarity detection in arabic language. *Arabian Journal for Science and Engineering*, 44.

Hien T Nguyen, Phuc H Duong, and Erik Cambria. 2019. Learning short-text semantic similarity with word embeddings and external knowledge sources. *Knowledge-Based Systems*, 182:104842.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y. Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. 2021. Large dual encoders are generalizable retrievers.

Nedjma Ousidhoum, Shamsuddeen Hassan Muhammad, Mohamed Abdalla, Idris Abdulmumin, Ibrahim Said Ahmad, Sanchit Ahuja, Alham Fikri Aji, Vladimir Araujo, Abinew Ali Ayele, Pavan Baswani, Meriem Beloucif, Chris Biemann, Sofia Bourhim, Christine De Kock, Genet Shanko Dekebo, Oumaima Hourrane, Gopichand Kanumolu, Lokesh Madasu, Samuel Rutunda, Manish Shrivastava, Thamar Solorio, Nirmal Surange, Hailegnaw Getaneh Tilaye, Krishnapriya Vishnubhotla, Genta Winata, Seid Muhie Yimam, and Saif M. Mohammad. 2024a. Semrel2024: A collection of semantic textual relatedness datasets for 14 languages.

Nedjma Ousidhoum, Shamsuddeen Hassan Muhammad, Mohamed Abdalla, Idris Abdulmumin, Ibrahim Said Ahmad, Sanchit Ahuja, Alham Fikri Aji, Vladimir Araujo, Meriem Beloucif, Christine De Kock, Oumaima Hourrane, Manish Shrivastava, Thamar Solorio, Nirmal Surange, Krishnapriya Vishnubhotla, Seid Muhie Yimam, and Saif M. Mohammad. 2024b. SemEval-2024 task 1: Semantic textual relatedness for african and asian languages. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, D. Passos, A. andCournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jiayan Pei, Yimin Wu, Zishan Qin, Yao Cong, and Jingtao Guan. 2021. Attention-based model for predicting question relatedness on stack overflow. In *2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR)*, pages 97–107. IEEE.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Marco Siino, Ilenia Tinnirello, and Marco La Cascia. 2024. Is text preprocessing still worth the time? a comparative survey on the influence of popular preprocessing methods on transformers and traditional classifiers. *Information Systems*, 121:102342.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Michael BW Wolfe, Joseph P Magliano, and Benjamin Larsen. 2005. Causal and semantic relatedness in discourse understanding and representation. *Discourse Processes*, 39(2-3):165–187.

Gabriela Mariel Zunino. 2023. Comprender lo desconocido: expectativas, relaciones semánticas y causalidad por defecto revisitada. *Lenguaje*, 51(1):156–186.