All-MPNet at SemEval-2024 Task 1: Application of MPNet for Evaluating Semantic Textual Relatedness

Marco Siino
Department of Electrical, Electronic
and Computer Engineering
University of Catania
Italy
marco.siino@unipa.it

Abstract

In this study, we tackle the task of automatically discerning the level of semantic relatedness between pairs of sentences. Specifically, Task 1 at SemEval-2024 involves predicting the Semantic Textual Relatedness (STR) of sentence pairs. Participants are tasked with ranking sentence pairs based on their proximity in meaning, quantified by their degree of semantic relatedness, across 14 different languages. To each sentence pair is assigned a manually determined relatedness score ranging from 0 (indicating complete lack of relation) to 1 (denoting maximum relatedness). In our submitted approach on the official test set, focusing on Task 1 (a supervised task in English and Spanish), we achieve a Spearman rank correlation coefficient of 0.808 for the English language and of 0.611 for the Spanish language.

1 Introduction

The notion of semantic relatedness between language units has been foundational in understanding meaning. Automatic determination of relatedness has wide-ranging applications, including assessing sentence representation methods, question answering, and summarization (Guarino, 1997).

Two sentences are deemed semantically similar when they exhibit a relationship of entailment or paraphrases. In contrast, relatedness encompasses a broader concept, encompassing all commonalities between two sentences: whether they pertain to the same topic, convey the same viewpoint, stem from the same temporal context, one elaborates on the other, and so forth (Hadj Taieb et al., 2020).

Historically, much of NLP research has focused on semantic similarity, predominantly in English. However, in the shared Task 1 hosted at SemEval 2024 (Ousidhoum et al., 2024b,a), the organizers extend their coverage to include the following languages: Afrikaans, Algerian Arabic, Amharic, English, Hausa, Hindi, Indonesian, Kinyarwanda, Marathi, Moroccan Arabic, Modern Standard Arabic, Punjabi, Spanish, and Telugu.

With the advancement of machine and deep learning architectures in recent years, there has been a surge of interest in NLP. Numerous efforts have been dedicated to creating algorithms capable of automatically identifying and categorizing text information available on the internet. In the literature, several strategies have already been proposed. In the last fifteen years, some of the most successful strategies have been based on SVM (Colas and Brazdil, 2006; Croce et al., 2022), on Convolutional Neural Network (CNN) (Kim, 2014; Siino et al., 2021), on Graph Neural Network (GNN) (Lomonaco et al., 2022), on ensemble models (Miri et al., 2022; Siino et al., 2022) and, recently, on Transformers (Vaswani et al., 2017; Siino et al., 2022).

The increasing adoption of Transformer-based architectures in academic research has also been bolstered by various methodologies showcased at SemEval 2024. These methodologies tackle diverse tasks and yield noteworthy findings. For instance, at the Task 2 (Jullien et al., 2024), where to address the challenge of identifying the inference relation between a plain language statement and Clinical Trial Reports is used T5 (Siino, 2024c); Task 4 (Dimitrov et al., 2024) where is employed a Mistral 7B model to detect persuasion techniques in memes (Siino, 2024b); and Task 8 (Wang et al., 2024), that utilizes a DistilBERT model to identify machine-generated text (Siino, 2024a).

For our model development, we devised a two-stage architecture. In the first stage, we utilized a Sentence Transformer specifically trained in a multilingual domain. Subsequently, we computed the cosine similarity of the generated embeddings to predict the relatedness between the analyzed sentences.

The remainder of this paper is structured as follows: Section 2 provides background information

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 379–384

on Task 1 hosted at SemEval-2024. Section 3 presents an explanation of the submitted approach. We detail the experimental setup required to reproduce our work in Section 4. The results of the formal assignment and pertinent discussions are presented in Section 5. Finally, we conclude with our findings and suggestions for future research in Section 6.

We make all the code publicly available and reusable on GitHub[1].

2 Background

Data for Semantic Textual Relatedness (STR) Shared Task 1[2] includes sentence pairs labeled with scores representing the degree of semantic textual relatedness between them, ranging from 0 (completely unrelated) to 1 (maximally related). These scores have been determined through manual annotation using a comparative annotation approach to mitigate biases commonly associated with traditional rating scale methods. This annotation process ensures a high reliability of the relatedness rankings.

The task involves predicting the STR of sentence pairs in 14 different languages. Participating teams were asked to submit systems for one, two, or all of the following tracks:

- Track A: Supervised — Systems trained using labeled training datasets provided. Additional publicly available datasets can be used, but teams must report the additional data and its impact on results.

- Track B: Unsupervised — Systems developed without using labeled datasets related to semantic relatedness or similarity between text units longer than two words in any language. Use of unigram or bigram relatedness datasets is permitted.

- Track C: Cross-lingual — Systems developed without using labeled semantic similarity or relatedness datasets in the target language, but with the use of labeled dataset(s) from at least one other language. Using labeled data from another track is mandatory for submissions to this track.

[1] https://github.com/marco-siino/SemEval2024/tree/main/Task%201

[2] https://semantic-textual-relatedness.github.io/

3 System Overview

The illustration of the proposed approach is provided in the Figure 1. Upon selecting a sample (i.e., a pair of sentences) from the dataset, the initial step involves encoding the first sentence using the All-MPNet embedding, thereby generating an embedding vector. Subsequently, employing an identical procedure, the second sentence from the sample is also encoded. The resulting embedding vectors are then subjected to a cosine similarity computation, facilitating the derivation of the semantic similarity prediction between two sentences.

To develop our model, we thought of a two-stage architecture. In the first stage, we used a *Sentence Transformer*. This is a Python framework for cutting-edge sentence, text, and image embeddings. The initial work is described in (Reimers and Gurevych, 2019). More than 100 languages have sentences and text embeddings that can be computed using this method. Sentences with a similar meaning can subsequently be found by comparing these embeddings, for example, using cosine-similarity. Semantic search, paraphrase mining, and semantic textual similarity can all benefit from these embeddings. The framework offers a huge selection of pre-trained models suited for different tasks and is built on PyTorch and Transformers. Moreover, fine-tuning models is also feasible.

The model used as Sentence Transformer is *all-mpnet-base-v2*, and it is available on HuggingFace[3]. The model is based on MPNet (Song et al., 2020). MPNet introduces a novel pre-training approach that combines the strengths of BERT and XLNet while addressing their respective limitations. Unlike BERT's masked language modeling (MLM), MPNet utilizes permuted language modeling (PLM) to capture dependencies among predicted tokens more effectively. Additionally, MPNet incorporates auxiliary position information as input, allowing the model to process full sentences and mitigate position discrepancy issues present in XLNet. Pre-training of MPNet is conducted on a large-scale dataset exceeding 160 GB of text corpora, followed by fine-tuning on various downstream tasks such as GLUE (Wang et al., 2018) and SQuAD (Rajpurkar et al., 2016). Experimental findings demonstrate that MPNet significantly outperforms both MLM and PLM, achieving superior results across these tasks compared to previ-

[3] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

Figure 1: Block diagram of our proposed approach. Given a sample from the dataset, the first sentence is encoded with an All-MPNet embedding for generating an embedding vector as output. Also, the second sentence in the sample is encoded in the same way. Then the two embedding vectors are compared using cosine similarity to produce the final prediction on the semantic similarity of the two sentences.

ous state-of-the-art pre-trained models like BERT, XLNet, and RoBERTa, all under the same model configuration.

The model was used to map all the words present in the text to a word embedding space. Following the embeddings, the cosine distance between two sentences was calculated. The cosine similarity between the two embedding vectors is calculated as shown in the Equation 1.

$$\cos(\theta) = \frac{A \cdot B}{\|A\|_2 \|B\|_2} \quad (1)$$

The value provided as cosine similarity was then provided as the requested prediction for the two sentences considered. Our code is available online together with the predictions generated and sent in relation to the test set.

The recent study by (Siino et al., 2024b) highlights that preprocessing for text classification tasks lacks significant impact when employing Transformers. Specifically, the study finds that the optimal preprocessing strategies do not substantially differ from performing no preprocessing at all, particularly in the case of Transformers. Consequently, in order to maintain our system's efficiency, speed, and computational lightness, we have opted to not conduct any preprocessing on the text. This decision aligns with the findings of the study and underscores the effectiveness of Transformers in handling raw text data without the need for extensive preprocessing steps.

4 Experimental Setup

We implemented our model on Google Colab[4]. The library we used was Sentence Transformer. The library requires Python[5] (>= 3.8) and PyTorch[6] (>=1.11.0). The dataset provided for all the phases are available on the official competition page. On the basis of our preliminary experiments, we found beneficial to set the threshold value upon the cosine similarity equal to 0.5. We did not perform any additional fine-tuning on the MPNet embeddings. To run the experiment, a T4 GPU from Google has been used. After the generation of the predictions, we exported the results on the JSON format required by the organizers. As already mentioned, all of our code is available on GitHub.

5 Results

The official evaluation metric for this task is the Spearman rank correlation coefficient, which evaluates how closely the rankings predicted by the system align with human judgments. The evaluation script for this shared task is available on the GitHub page, providing a standardized method for assessing the performance of participating systems. The formula to compute the Spearman correlation coefficient is provided in the Equation 2.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (2)$$

Where d represents the pairwise distances of the ranks of the variables x_i and y_i, while n is the number of samples.

In Table 1, are reported the results obtained on the two languages we considered for our participation at the task. Considered the very low effort required to run the proposed approach and to generate the predictions, the score of 0.611 and 0.808 appears to be an interesting baseline, while still exhibiting room for improvements. It is worth noticing that the approach is a Zero-Shot one with no prior knowledge on the specific task.

In the Table 2 and 3, the results obtained by the first three teams and by the last one, as showed on the official CodaLab page, are reported. Furthermore, we reported the baselines for the two languages. Compared to the best performing models, our simple approach exhibits some room for improvements. However, it is worth notice that it

[4] https://colab.research.google.com/
[5] https://www.python.org/
[6] https://pytorch.org/

LANGUAGE	Score
English	0.808
Spanish	0.611

Table 1: The suggested method's performance on the test set. Our results are related to our participation in the Track A, for the English and for the Spanish languages only.

TEAM NAME	Score
PALI (1)	0.859
UAlberta (2)	0.853
Tübingen-CL (3)	0.850
SemRel-SemEval Baseline (*)	0.830
YNUNLP2023 (36)	0.557

Table 2: Comparing performance on the test set for the English language. In the table are shown the results obtained by the first three users and by the last one. In parentheses is reported the position in the official ranking.

required no further pre-training and the computational cost to address the task is manageable with the free online resources offered by Google Colab.

Even if our approach is simple and straightforward, here we want to qualitative analyze some results obtained with our approach to better motivate some classification mistakes. With regard to the English language, the first relevant misclassification sample is related to the sample pair: "This book is very compelling. – best book so far in the series!". Our system predicts a cosine similarity equal to 0.47 while the actual similarity is 0.73. In this case, in fact, it is very hard to assess if "being very compelling" is so semantically similar to the concept of "being the best so far in a series". Furthermore, looking at the sample: "A woman in a black coat eats dinner while her dog looks on. – A little boy is standing on the street while a man in overalls is working on a stone wall.", the prediction using our approach is equal to 0.0 (no semantic similarity) while the actual target for the provided test set is 0.29. Another interesting case — i.e., our approach predicts a high similarity of 0.92 while the actual target is 0.74 — is given by the sample: "My favorite by far is definitely Chris and I think he will win!! – My favorite and the selection for winner is Chris.". From a semantic perspective, however, both the sentences provide the same two concepts (i.e., Chris is my favorite, I think he will win). Given these and several others differences between our predictions and the actual target similarity in the provided test set, some concerns on the labelling process and on the correctness of the provided target similarity values can be raised.

TEAM NAME	Score
AAdaM (1)	0.740
GIL-IIMAS UNAM (2)	0.731
PALI (3)	0.724
SemRel-SemEval Baseline (*)	0.7
YNUNLP2023 (25)	0.404

Table 3: Comparing performance on the test set for the Spanish language. In the table are shown the results obtained by the first three users and by the last one. Furthermore, the baseline is also provided. In parentheses is reported the position in the official ranking.

6 Conclusion

This paper introduces the utilization of an All-MPNet model embedding to tackle Task 1 at SemEval-2024. In our submission, we opted for a straightforward Zero-Shot learning approach, leveraging pre-trained Transformers that are already tailored to a multilingual-domain. Following this approach, we utilized the contextual embeddings generated by the Sentence Transformer, and we employed cosine distance to measure the similarity between pairs of sentences, thus quantifying the STR between them. Despite the effectiveness of our method, there remains room for improvement, as indicated by the final ranking. Potential alternative approaches could involve leveraging the zero-shot capabilities of models such as GPT and T5, expanding the training data size by incorporating additional datasets, or exploring different methods of integrating ontology-based domain knowledge into our approach. Furthermore, given the interesting results recently provided on a plethora of tasks, also few-shot learning (Wang et al., 2023; Maia et al., 2024; Siino et al., 2023; Meng et al., 2024) or data augmentation strategies (Muftie and Haris, 2023; Siino et al., 2024a; Tapia-Téllez and Escalante, 2020; Siino and Tinnirello, 2023) could be employed to improve the performance. Compared to the best performing models, our simple approach exhibits some room for improvements. However, our qualitative analysis raised some concerns on the labels provided for the test set. Then, we are not able to correctly assess the actual performance of our proposed approach. Eventually, it

is worth notice that thanks to our approach no further pre-training is required and the computational cost to address the task is manageable with the free online resources offered by Google Colab.

Acknowledgments

We would like to thank anonymous reviewers for their comments and suggestions that have helped to improve the presentation of the paper.

References

Fabrice Colas and Pavel Brazdil. 2006. Comparison of svm and some older classification algorithms in text classification tasks. In *IFIP International Conference on Artificial Intelligence in Theory and Practice*, pages 169–178. Springer.

Daniele Croce, Domenico Garlisi, and Marco Siino. 2022. An SVM ensemble approach to detect irony and stereotype spreaders on twitter. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022*, volume 3180 of *CEUR Workshop Proceedings*, pages 2426–2432. CEUR-WS.org.

Dimitar Dimitrov, Firoj Alam, Maram Hasanain, Abul Hasnat, Fabrizio Silvestri, Preslav Nakov, and Giovanni Da San Martino. 2024. Semeval-2024 task 4: Multilingual detection of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation*, SemEval 2024, Mexico City, Mexico.

Nicola Guarino. 1997. Semantic matching: Formal ontological distinctions for information organization, extraction, and integration. In *Information Extraction A Multidisciplinary Approach to an Emerging Information Technology: International Summer School, SCIE-97 Frascati, Italy, July 14–18, 1997*, pages 139–170. Springer.

Mohamed Ali Hadj Taieb, Torsten Zesch, and Mohamed Ben Aouicha. 2020. A survey of semantic relatedness evaluation datasets and procedures. *Artificial Intelligence Review*, 53(6):4407–4448.

Maël Jullien, Marco Valentino, and André Freitas. 2024. SemEval-2024 task 2: Safe biomedical natural language inference for clinical trials. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751. ACL.

Francesco Lomonaco, Gregor Donabauer, and Marco Siino. 2022. COURAGE at checkthat!-2022: Harmful tweet detection using graph neural networks and ELECTRA. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022*, volume 3180 of *CEUR Workshop Proceedings*, pages 573–583. CEUR-WS.org.

Beatriz Matias Santana Maia, Maria Clara Falcão Ribeiro de Assis, Leandro Muniz de Lima, Matheus Becali Rocha, Humberto Giuri Calente, Maria Luiza Armini Correa, Danielle Resende Camisasca, and Renato Antonio Krohling. 2024. Transformers, convolutional neural networks, and few-shot learning for classification of histopathological images of oral cancer. *Expert Systems with Applications*, 241:122418.

Zong Meng, Zhaohui Zhang, Yang Guan, Jimeng Li, Lixiao Cao, Meng Zhu, Jingjing Fan, and Fengjie Fan. 2024. A hierarchical transformer-based adaptive metric and joint-learning network for few-shot rolling bearing fault diagnosis. *Measurement Science and Technology*, 35(3).

Mohsen Miri, Mohammad Bagher Dowlatshahi, Amin Hashemi, Marjan Kuchaki Rafsanjani, Brij B Gupta, and W Alhalabi. 2022. Ensemble feature selection for multi-label text classification: An intelligent order statistics approach. *International Journal of Intelligent Systems*, 37(12):11319–11341.

Fuad Muftie and Muhammad Haris. 2023. Indobert based data augmentation for indonesian text classification. In *2023 International Conference on Information Technology Research and Innovation, ICITRI 2023*, page 128 – 132.

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. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Marco Siino. 2024a. Badrock at semeval-2024 task 8: Distilbert to detect multigenerator, multidomain and multilingual black-box machine-generated text. In *Proceedings of the 18th International Workshop on Semantic Evaluation*, SemEval 2024, Mexico City, Mexico.

Marco Siino. 2024b. Mcrock at semeval-2024 task 4: Mistral 7b for multilingual detection of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation*, SemEval 2024, Mexico City, Mexico.

Marco Siino. 2024c. T5-medical at semeval-2024 task 2: Using t5 medical embeddings for natural language inference on clinical trial data. In *Proceedings of the 18th International Workshop on Semantic Evaluation*, SemEval 2024, Mexico City, Mexico.

Marco Siino, Elisa Di Nuovo, Ilenia Tinnirello, and Marco La Cascia. 2021. Detection of hate speech spreaders using convolutional neural networks. In *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021*, volume 2936 of *CEUR Workshop Proceedings*, pages 2126–2136. CEUR-WS.org.

Marco Siino, Marco La Cascia, and Ilenia Tinnirello. 2022. Mcrock at semeval-2022 task 4: Patronizing and condescending language detection using multi-channel cnn, hybrid lstm, distilbert and xlnet. In *Proceedings of the 16th International Workshop on Semantic Evaluation, SemEval@NAACL 2022, Seattle, Washington, United States, July 14-15, 2022*, pages 409–417. Association for Computational Linguistics.

Marco Siino, Francesco Lomonaco, and Paolo Rosso. 2024a. Backtranslate what you are saying and i will tell who you are. *Expert Systems*, n/a(n/a):e13568.

Marco Siino, Maurizio Tesconi, and Ilenia Tinnirello. 2023. Profiling cryptocurrency influencers with few-shot learning using data augmentation and ELECTRA. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023*, volume 3497 of *CEUR Workshop Proceedings*, pages 2772–2781. CEUR-WS.org.

Marco Siino and Ilenia Tinnirello. 2023. Xlnet with data augmentation to profile cryptocurrency influencers. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023*, volume 3497 of *CEUR Workshop Proceedings*, pages 2763–2771. CEUR-WS.org.

Marco Siino, Ilenia Tinnirello, and Marco La Cascia. 2022. T100: A modern classic ensemble to profile irony and stereotype spreaders. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022*, volume 3180 of *CEUR Workshop Proceedings*, pages 2666–2674. CEUR-WS.org.

Marco Siino, Ilenia Tinnirello, and Marco La Cascia. 2024b. 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.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867.

José Medardo Tapia-Téllez and Hugo Jair Escalante. 2020. Data augmentation with transformers for text classification. In *Advances in Computational Intelligence*, pages 247–259, Cham. Springer International Publishing.

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.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Xixi Wang, Xiao Wang, Bo Jiang, and Bin Luo. 2023. Few-shot learning meets transformer: Unified query-support transformers for few-shot classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(12):7789–7802.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024. Semeval-2024 task 8: Multigenerator, multidomain, and multilingual black-box machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation*, SemEval 2024, Mexico, Mexico.