NLP-LISAC at SemEval-2024 Task 1: Transformer-based approaches for Determining Semantic Textual Relatedness

Abdessamad Benlahbib[1] , **Anass Fahfouh**[1], **Hamza Alami**[1] , **Achraf Boumhidi**[1]
[1] LISAC Laboratory, Faculty of Sciences Dhar EL Mehraz, USMBA, Fez, Morocco
abdessamad.benlahbib@usmba.ac.ma , anassfahfouh@gmail.com,
hamza.alami5@usmba.ac.ma, achraf.boumhidi@usmba.ac.ma

Abstract

This paper presents our system and findings for SemEval 2024 Task 1 Track A Supervised Semantic Textual Relatedness. The main objective of this task was to detect the degree of semantic relatedness between pairs of sentences. Our submitted models (ranked 6/24 in Algerian Arabic, 7/25 in Spanish, 12/23 in Moroccan Arabic, and 13/36 in English) consist of various transformer-based models including MARBERT-V2, mDeBERTa-V3-Base, DarijaBERT, and DeBERTa-V3-Large, fine-tuned using different loss functions including Huber Loss, Mean Absolute Error, and Mean Squared Error.

1 Introduction

Semantic Textual Relatedness (STR) is a natural language processing (NLP) task that focuses on measuring the degree of semantic relatedness between two pieces of text. Unlike tasks such as Semantic Textual Similarity (STS), which specifically assess the degree of similarity between texts, STR considers a broader notion of relatedness, encompassing various types of semantic relationships between words, phrases, or sentences.

The goal of STR is to quantify how closely related two pieces of text are in terms of their underlying meaning or semantic content. This relatedness can encompass a wide range of semantic relationships, including:

- **Synonymy**: Words or phrases that have similar meanings.
- **Hyponymy/Hypernymy**: Hierarchical relationships where one word is a more specific instance (hyponym) or a more general category (hypernym) of another word.
- **Meronymy/Holonymy**: Meronymy is a semantic relation between a meronym denoting a part and a holonym denoting a whole.
- **Antonymy**: Words with opposite meanings.
- **Entailment**: One statement logically implies another statement.
- **Association**: Words or concepts that are commonly associated with each other.

In the context of STR, annotators or models are typically presented with pairs of text and asked to judge the degree of relatedness based on the presence of shared concepts or semantic associations. Annotators might provide relatedness scores or labels indicating the strength of the relationship between text pairs.

In this paper, we present our findings on SemEval 2024 Task 1 Track A: Supervised Semantic Textual Relatedness (Ousidhoum et al., 2024b). Our method consists of various transformer-based approaches (Vaswani et al., 2017) fine-tuned using different loss functions including Huber Loss, Mean Absolute Error, and Mean Squared Error.

The rest of the paper is structured in the following manner: Section 2 provides the main objective of the Task. Section 3 describes our system. Section 4 details the experiments. And finally, Section 5 concludes this paper.

2 Task Description

This task aims to predict the semantic textual relatedness (STR) of pairs of sentences across 14 different languages. Participants will rank sentence pairs based on their semantic closeness, ranging from 0 (completely unrelated) to 1 (maximally related), as determined manually. Teams can submit entries for one, two, or all of the following tracks:

- **Track A: Supervised**: Participants are required to submit systems trained using provided labeled training datasets. They may utilize publicly available datasets, but must disclose additional data used and assess its impact on results.

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 213–217

- **Track B: Unsupervised**: Participants must submit systems developed without using labeled datasets on semantic relatedness or similarity between text units longer than two words in any language. However, the use of unigram or bigram relatedness datasets from any language is allowed.
- **Track C: Cross-lingual**: Participants must submit systems developed without labeled semantic similarity or relatedness datasets in the target language, but may use labeled dataset(s) from at least one other language. Note: Utilizing labeled data from another track is mandatory for submissions to this track.

3 System Description

To tackle the SemEval 2024 Task 1 Track A: Supervised Semantic Textual Relatedness, we fine-tuned several transformer-based models on an augmented training dataset and with different loss functions including Huber Loss, Mean Absolute Error, and Mean Squared Error. The different steps of our system are described as follows:

- We combined the training and development sets separately for each language in which we participated in. Besides, we duplicated the obtained datasets input, but we shifted the pairs order and we kept the same semantic relatedness score. Table 1 and 2 depict an example of augmenting the English training set.
- We replaced the newline character **\n** with **[SEP]** token in order to separate the input pairs. For example, this input: "Then, in twenty minutes, gather at the runway. \n gathering on the runway, in 20 minutes." will be converted to "Then, in twenty minutes, gather at the runway. [SEP] gathering on the runway, in 20 minutes."
- We tokenized the data using tokenizers associated with the fine-tuned transformer based models.
- We fine-tuned MARBERTv2 (Abdul-Mageed et al., 2021) on the Algerian Arabic data, DarijaBERT (Gaanoun et al., 2024) on the Moroccan Arabic data, DeBERTa-V3-Large (He et al., 2021a,b) on the English data, and mDeBERTa-V3-Base (He et al., 2021a) on the Spanish data.

In the context of semantic textual relatedness tasks, the choice of loss function plays a critical role in guiding the training process and optimizing model performance. Given the diverse nature of textual data and the wide range of semantic relationships to be captured, employing a variety of loss functions can offer several advantages.

Firstly, the Huber Loss function provides robustness to outliers by combining the advantages of Mean Absolute Error (MAE) and Mean Squared Error (MSE). MAE, which calculates the average absolute difference between predicted and target values, is less sensitive to outliers compared to MSE, which squares the differences. By behaving like MSE for large errors and like MAE for small errors, Huber Loss ensures that the training process is less influenced by outliers, thereby enhancing the model's ability to generalize to unseen data.

Secondly, Mean Absolute Error (MAE) serves as a straightforward and intuitive loss function that penalizes deviations from the target scores equally, irrespective of their direction. In tasks such as semantic textual relatedness, where the goal is to predict similarity scores between sentence pairs, MAE provides a direct measure of the magnitude of errors, facilitating easy interpretation and evaluation of model performance.

Lastly, Mean Squared Error (MSE) emphasizes the importance of accurately predicting similarity scores by penalizing larger errors more severely than smaller errors. In scenarios where precise estimation of the degree of relatedness between sentence pairs is crucial, MSE can effectively guide the training process towards minimizing the squared differences between predicted and target values, thereby optimizing model performance.

By leveraging a combination of these loss functions during the fine-tuning process, we aim to capitalize on their respective strengths and enhance the robustness and effectiveness of our models in capturing semantic relationships within textual data. This approach enables us to optimize model performance across various linguistic contexts and achieve competitive results in tasks requiring accurate assessment of semantic textual relatedness.

The decision to augment the data was validated through fine-tuning the models on concatenated train and dev sets, as well as on concatenated train and dev sets with pair shifting. Interestingly, our analysis revealed that pair shifting significantly enhanced the results on the development sets.

PairID	Text	Score
ENG-train-0047	Then, in twenty minutes, gather at the runway. \n gathering on the runway, in 20 minutes.	0.97
ENG-dev-0010	Meat is dropped into a pan. \n A woman is putting meat in a pan.	0.73

Table 1: Sample of the English training set after combining both training and development sets

PairID	Text	Score
ENG-train-0047	Then, in twenty minutes, gather at the runway. \n gathering on the runway, in 20 minutes.	0.97
ENG-dev-0010	Meat is dropped into a pan. \n A woman is putting meat in a pan.	0.73
ENG-train-0047-shifted	gathering on the runway, in 20 minutes. \n Then, in twenty minutes, gather at the runway.	0.97
ENG-dev-0010-shifted	A woman is putting meat in a pan. \n Meat is dropped into a pan.	0.73

Table 2: Sample of the English training set after combining both training and development sets and after shifting the pairs

4 Experimental Results

We experimented our model on the SemEval 2024 Task 1: Semantic Textual Relatedness (STR) test set (Ousidhoum et al., 2024a). The experiment has been conducted in Kaggle environment[1], The following libraries: Transformers - Hugging Face[2] (Wolf et al., 2020), and Keras[3] were used to train and to assess the performance of our models.

4.1 Datasets

Each instance in the training, development, and test sets (Ousidhoum et al., 2024a) is a sentence pair. The instance is labeled with a score representing the degree of semantic textual relatedness between the two sentences. The scores can range from 0 (maximally unrelated) to 1 (maximally related). Figure 1 depicts the training, dev and test sets distributions for Algerian Arabic, Moroccan Arabic, English and Spanish.

The datasets are available via GitHub[4]

4.2 Experimental Settings

We conducted numerous experiments on the development set to obtain the ideal number of epochs and identify the most effective loss function for fine-tuning each model. This paper presents the hyperparameters that yielded the best results on the development set across the target languages:

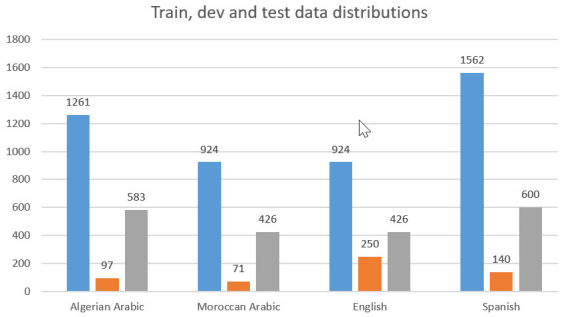

Figure 1: Train, development and test sets distributions for Algerian Arabic, Moroccan Arabic, English and Spanish

- **Algerian Arabic**: We fine-tuned MARBERTv2 using 12 epochs, a maximum sequence length of 200, and Mean Absolute Error as the loss function.

- **Moroccan Arabic**: We fine-tuned DarijaBERT using 5 epochs, a maximum sequence length of 200, and Huber loss as the loss function.

- **English**: We fine-tuned DeBERTa-V3-Large using 5 epochs, a maximum sequence length of 150, and Huber loss as the loss function.

[1] https://www.kaggle.com/
[2] https://huggingface.co/docs/transformers/index
[3] https://keras.io/
[4] https://github.com/semantic-textual-relatedness/Semantic_Relatedness_SemEval2024

- **Spanish**: We fine-tuned mDeBERTa-V3-Base using 12 epochs, a maximum sequence length of 200, and Mean Squared Error as the loss function.

The same parameters were utilized during the final submission phase. Additionally, Table 3 displays the additional hyperparameter settings employed during the fine-tuning process for all models.

Hyperparameters	Settings
Learning rate	10^{-5}
Batch size	4
Optimizer	Adam (Kingma and Ba, 2015)

Table 3: Hyperparameters settings for the model in the experiments

4.3 System Performance

Table 4 depicts the results of our proposed approaches on SemEval 2024 Task 1 Track A Supervised Semantic Textual Relatedness. The official evaluation metric for this task is the Spearman's rank correlation coefficient, which captures how well the system-predicted rankings of test instances align with human judgments.

Language	Score (Spearman)	Ranking
Algerian Arabic	0.6035781253	6
Spanish	0.7171198162	7
Moroccan Arabic	0.7893667707	12
English	0.8345843316	13

Table 4: Results of our proposed models on SemEval 2024 Task 1 Track A : Supervised Semantic Textual Relatedness test set

Based on the experimental results, our approaches for SemEval 2024 Task 1 Track A: Supervised Semantic Textual Relatedness demonstrated competitive performance across multiple languages. Here's a summary of our findings:

- **Algerian Arabic** : Our model achieved a score of 0.6035781253, ranking 6th out of 24 submissions. This indicates that our approach effectively captured the semantic relatedness between sentence pairs in Algerian Arabic, outperforming a significant portion of the participating systems.

- **Spanish** : In Spanish, our model achieved a score of 0.7171198162, securing the 7th position out of 25 submissions. This suggests that our approach successfully captured semantic relationships in Spanish text, performing competitively compared to other systems.

- **Moroccan Arabic** : Our model attained a score of 0.7893667707, ranking 12th out of 23 submissions in Moroccan Arabic. While our performance in this language was slightly lower compared to others, our approach still demonstrated notable effectiveness in capturing semantic relatedness in Moroccan Arabic text.

- **English** : For English, our model achieved a score of 0.8345843316, placing 13th out of 36 submissions. Despite the larger number of submissions in English, our approach still showcased strong performance, indicating its capability to accurately assess semantic relatedness in English sentence pairs.

Overall, our experimental results highlight the robustness and effectiveness of our proposed approaches across different languages in capturing semantic textual relatedness. These findings underscore the potential of transformer-based models fine-tuned with appropriate hyperparameters and loss functions to excel in tasks requiring semantic understanding of textual data. Additionally, the competitive rankings across multiple languages signify the versatility and generalizability of our approach, further validating its suitability for real-world applications requiring accurate assessment of semantic relatedness in diverse linguistic contexts.

5 Conclusion

In conclusion, this paper has presented our system and findings for SemEval 2024 Task 1 Track A: Supervised Semantic Textual Relatedness. The primary objective of this task was to detect the degree of semantic relatedness between pairs of sentences across multiple languages. Our submitted models, leveraging various transformer-based architectures including MARBERT-V2, mDeBERTa-V3-Base, DarijaBERT, and DeBERTa-V3-Large, fine-tuned with different loss functions such as Huber Loss, Mean Absolute Error, and Mean Squared Error, achieved competitive rankings across different language tracks.

Our approach highlights the effectiveness of leveraging advanced transformer-based models and fine-tuning techniques to capture intricate semantic relationships within textual data. By incorporating diverse loss functions during the training process, we aimed to optimize the models' performance and enhance their robustness across various linguistic contexts.

Moving forward, further research in semantic textual relatedness should focus on refining existing methodologies, exploring novel architectures, and addressing cross-lingual challenges. Additionally, efforts to incorporate additional linguistic features and develop more comprehensive evaluation metrics can contribute to advancing the state-of-the-art in this field.

Overall, our contributions underscore the significance of semantic textual relatedness in natural language processing tasks and pave the way for the development of more sophisticated and context-aware systems capable of understanding and interpreting textual data with greater precision and accuracy.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Kamel Gaanoun, Abdou Mohamed Naira, Anass Allak, and Imade Benelallam. 2024. Darijabert: a step forward in nlp for the written moroccan dialect. *International Journal of Data Science and Analytics*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.