UniBuc-AE at SemEval-2025 Task 7: Training Text Embedding Models for Multilingual and Crosslingual Fact-Checked Claim Retrieval

Alexandru Enache
University of Bucharest
Faculty of Mathematics and Computer Science
alexandru.enache1@s.unibuc.ro

Abstract

This paper describes our approach to the SemEval-2025 Task 7: Multilingual and Crosslingual Fact-Checked Claim Retrieval on both the monolingual and crosslingual tracks. Our training methodology for text embedding models combines contrastive pre-training and hard negatives mining in order to fine-tune models from the E5 family. Additionally, we introduce a novel approach for merging the results from multiple models by finding the best majority vote weighted configuration for each subtask using the validation dataset. Our team ranked 6^{th} in the monolingual track scoring a 0.934 S@10 averaged over all languages and achieved a 0.79 S@10 on the crosslingual task, ranking 8^{th} in this track.

1 Introduction

Fact-checking is becoming a crucial task in today's society. Viral posts on social networks can reach an astonishing number of people in very little time, and false information tends to spread faster than true data (Vosoughi et al., 2018). Thus, automated fact-checking systems can be a useful tool in combating this problem.

The previous research done in this space was mostly focused on the English language and on the monolingual task, where the fact-check and post are in the same language, bot not on the crosslingual case. The MultiClaim dataset (Pikuliak et al., 2023) is the biggest dataset of fact-checks released to date and introduces a special section for crosslingual evaluation. This year's SemEval-2025 Task 7 (Peng et al., 2025) further enhances this dataset with modifications and augmentations for this task.

This paper proposes an automated fact-checking system based on text embedding models that allow the retrieval of relevant fact-checked claims through data vectorization. The implementation is based on the multilingual-e5-large-instruct (Wang et al., 2024b) and e5-large-v2 (Wang et al., 2022) models.

Our approach[1] creates models consistent for both the monolingual and crosslingual tasks, ranking 6^{th} in the monolingual track and 8^{th} in the crosslingual track . The full results of our approach are available in the Results section in Table 4.

2 Background

2.1 Related Work

Pikuliak et al. (2023) experimented with BM25 and various English and multilingual text embedding models on the MultiClaim dataset, achieving an S@10 score of 0.83 on the monolingual task and 0.56 on the crosslingual task using the GTR-T5-Large model. From their results, the best results were generated by first translating both the posts and fact-checks to English using out-of-the-box AI services and then running English text embedding models on the translated data.

2.2 Dataset

The dataset (Peng et al., 2025) is an enhanced version of the original MultiClaim dataset (Pikuliak et al., 2023). It contains 280K unique fact-checks and 33K unique social media posts split between train, dev and test. For the train dataset we are given 25K pairs of matching post fact-check pairs in 8 different languages: French, Spanish, English, Portuguese, Thai, German, Malay and Arabic. The test dataset introduces two other languages: Turkish and Polish. The distribution of posts and fact-checks per language in every dataset can be visualized in Figure 1. In order to evaluate our fine-tuning and find an optimal weighted configuration for the majority vote, we have created a validation dataset using a random 80-20 split on the training pairs.

[1]Our code and experiments are available at https://github.com/Alex18mai/UniBuc-AE-at-SemEval-2025-Task7

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 634–639

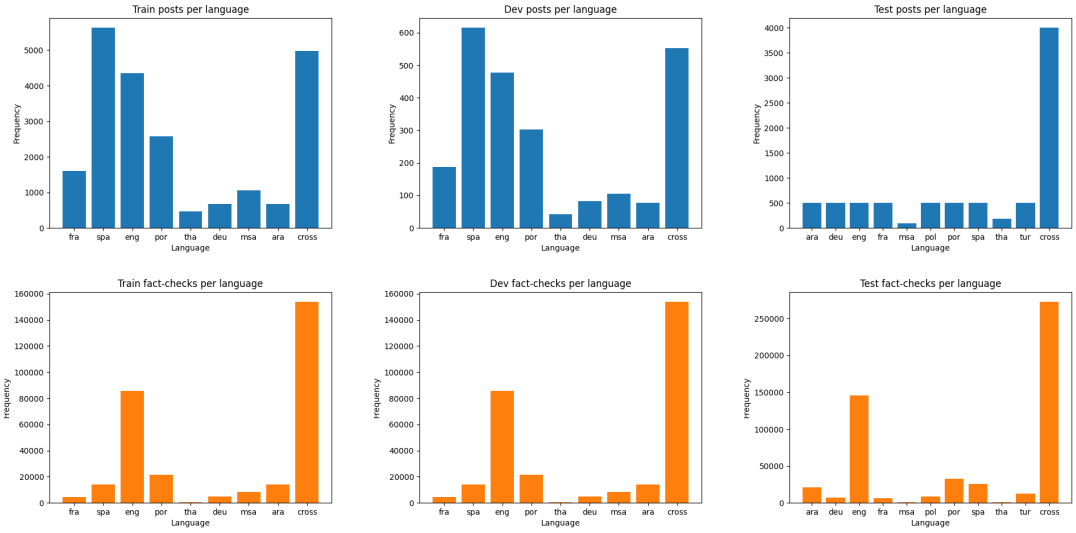

Figure 1: The language distribution for posts and fact-checks in the train, dev and test datasets.

Each social post contains the text and OCR extracted from the visual content of the post in both the original language and English translation. A parsed example of a social media post can be found in Appendix A.

Each fact-check contains the claim and title in both the original language and English translation. A parsed example of a fact-check can be found in Appendix B.

In the output submission, we need to report the top 10 retrieved fact-checks for each post.

3 System Overview

3.1 Architecture Overview

This paper presents a comprehensive approach applicable to both monolingual and crosslingual tasks, leveraging a combined training methodology that incorporates both monolingual and crosslingual pairs. Our experimental framework encompasses the utilization of multilingual text embedding models as well as English-specific text embedding models, employing the appropriate text for each scenario (original language or English translation).

The models are trained by first applying contrastive pre-training using in-batch negatives and then fine tuning using custom mined hard negatives for each post.

The final submission is generated by combining multiple models using an algorithmic approach to finding the best weighted configuration for each subtask.

3.2 Text Embedding Models

- **multilingual-e5-large-instruct** (Wang et al., 2024b) is a state of the art multilingual text embedding model initialized from xlm-roberta-large. It was trained using a dataset containing over 100 languages. Additionally, it is instruction-tuned, enhancing the quality of the embeddings by allowing custom instruction prompts to be given as context for the task at hand.

- **e5-large-v2** (Wang et al., 2022) is a powerful English text embedding model initialized from bert-large-uncased-whole-wordmasking. It was trained using the MS-MARCO, NQ and NLI datasets and has one of the best results on the MTEB Benchmark leaderboard (Muennighoff et al., 2022) for the small size of 335M parameters.

3.3 Contrastive Pre-training

The first epochs are trained using contrastive loss (van den Oord et al., 2019) with in-batch negatives and a temperature of 0.01, inspired from the E5 paper (Wang et al., 2022).

Using in-batch negatives is computationally efficient since it uses already computed embeddings as negative samples for each matching pair, but has a very low probability of having a negative pair with a high cosine similarity. Such pairs are called hard negatives since they are the negative pairs where the model fails to distance the embeddings and are very useful in the training process.

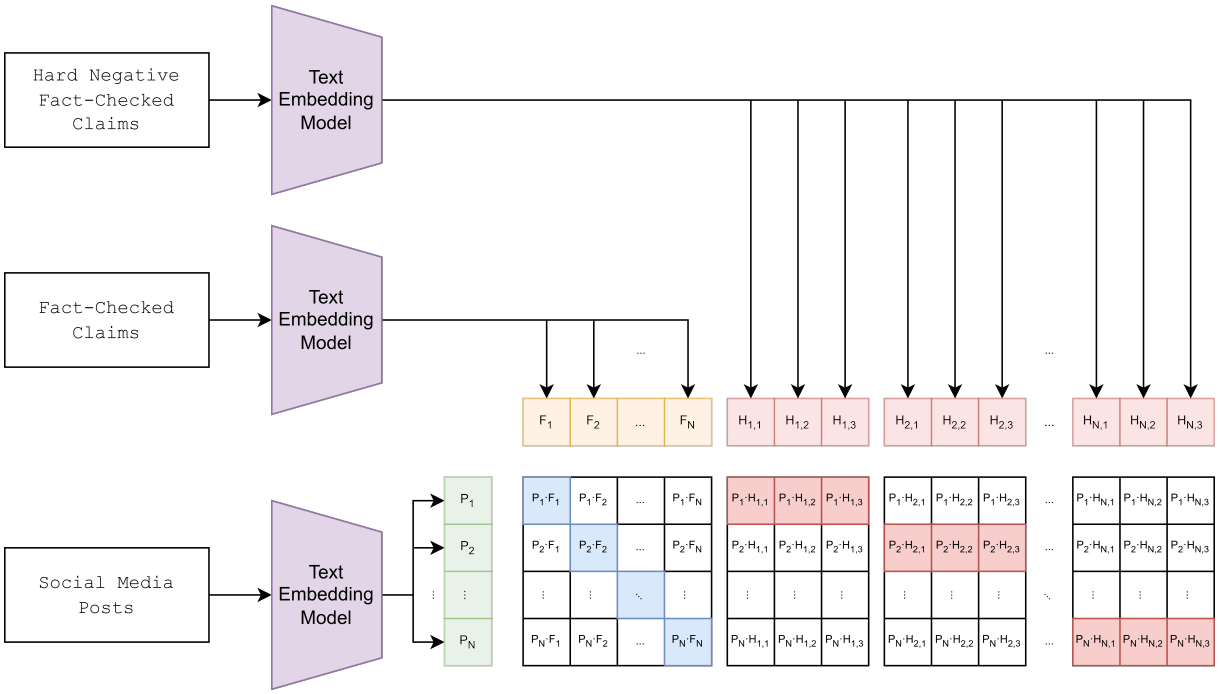

Figure 2: Contrastive training using hard negatives. For each social media post a positive and 3 hard negative fact-checked claims are sampled.

3.4 Hard Negatives Mining

After the contrastive pre-training, we use the models in order to predict the negative fact-checks with the highest similarity score to the posts from the training dataset. The hard negatives are then used in order to fine-tune the model by sampling together with the post fact-check matching pairs.

We have found that sampling any of the top 5 hard negatives for a post actually degrades the result. This can be justified by the fact that it also introduces false negatives, which are fact-checks that should have been labeled as matching or are too close to the meaning of the social media post. This is also confirmed by a recent study by Nvidia (de Souza P. Moreira et al., 2024) where they have proposed the Top-K shifted by N method. This is why we use the hard negatives ranked 8^{th} to 10^{th} (top-3 shifted by 7).

In Figure 2, we show the training process using hard negatives. For each social media post, we maximize the cosine similarity with the matching fact-checked claim (blue boxes) and minimize the cosine similarity with the rest of the fact-checked claims, including hard negatives (red boxes) and in-batch negatives (white boxes).

3.5 Weighted Majority Vote

In order to find the best weighted configuration for the majority vote on each subtask, we compute the top 1000 retrieved fact-checks with their corresponding similarity score for each post in the validation, dev and test dataset.

Formally, given N models and their retrievals, we want to find the optimal weights $w_1, w_2, ..., w_N$ ($w_1 + w_2 + ... + w_N = 1$) for each subtask so that, when attributing the score $sim_1 \cdot w_1 + sim_2 \cdot w_2 + ... + sim_N \cdot w_N$ to each fact-check we achieve the best S@10 score on the validation subtask. We denote the similarity score given by the N models for the post fact-check pair by $sim_1, sim_2, ..., sim_N$.

For simplicity, we can discretize the weights to 0.1 increments ($[0.0, 0.1, ..., 1.0]$ since smaller changes than 0.1 should not impact the results of the algorithm in a meaningful manner.

Now we can compute all the configurations of possible weights and check the validation in order to determine the best one for each subtask.

This method proved to be more reliable than the classical weighted majority vote based on the accuracy and brought a major improvement to the accuracy of our final submission.

Since the test dataset contains two languages which are not present in the train dataset, we have computed their configuration as the mean of

the monolingual configurations for the other languages.

4 Experimental Setup

The social media posts are converted to a single string by concatenating the text and all the OCR lines, generating one string in the original language and one in English using the translated text. A similar process is applied to the fact-checks by concatenating the claim and title.

As evaluation metrics we use the Success@10 which measures whether at least one relevant item is present in the top 10 recommendations.

$$\text{S@10} = \begin{cases} 1 & \text{if } \sum_{i=1}^{10} \text{rel}_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

,where rel_i represents the relevance score of the i-th item in the list.

The instruction prompt used for the multilingual-e5-large-instruct model is : "Instruct: Given a social media post, retrieve relevant fact-checked claims for the post".

All the experiments[2] were conducted using AdamW with a learning rate of 10^{-5} with a linear scheduler of 100 warmup steps and a weight decay of 10^{-5} .

Due to hardware limitations, the contrastive pre-training is done using a batch size of 8 pairs of matching post fact-check pairs and the hard negatives fine-tuning is done using a batch size of 4 matching pairs and 3 hard negatives for each post.

We note the configurations used in the experiments in the following way:

- **Eng** : e5-large-v2 model trained with English texts using **5** epochs of contrastive pre-training.
- **Orig** : multilingual-e5-large-instruct model trained with the original texts using **5** epochs of contrastive pre-training.
- **EngHard** : Eng further fine-tuned for **3** epochs using hard negatives generated.
- **OrigHard** : Orig further fine-tuned for **3** epochs using hard negatives generated.
- **MV** : The weighted majority vote using the best configuration for combining Eng, Orig, EngHard and OrigHard.

[2]The environment used for the experiments is available at https://github.com/Alex18mai/UniBuc-AE-at-SemEval-2025-Task7/blob/main/requirements.txt

Subtask	Eng	Orig	Eng Hard	Orig Hard	MV
eng	0.901	0.900	**0.951**	0.943	0.964
fra	0.931	0.937	**0.956**	0.946	0.962
deu	0.835	0.873	0.932	**0.940**	0.955
por	0.928	0.932	**0.959**	0.957	0.976
spa	0.939	0.937	**0.965**	0.958	0.973
tha	0.957	0.957	**0.989**	0.978	0.989
msa	0.939	0.924	**0.971**	0.957	0.990
ara	0.897	0.904	0.926	**0.941**	0.963
monolingual	0.916	0.920	**0.956**	0.953	0.971
crosslingual	0.793	0.869	0.889	**0.905**	0.934

Table 1: S@10 results on the validation dataset. The monolingual results represent the average of the monolingual subtasks. The best individual results (without MV) are in bold.

Subtask	Eng	Orig	Eng Hard	Orig Hard
eng	0.0	0.3	0.7	0.0
fra	0.0	0.0	0.8	0.2
deu	0.0	0.3	0.1	0.6
por	0.3	0.0	0.3	0.4
spa	0.3	0.3	0.4	0.0
tha	0.0	0.0	0.4	0.6
msa	0.0	0.0	0.4	0.6
ara	0.0	0.1	0.4	0.5
crosslingual	0.0	0.2	0.4	0.4

Table 2: Best weighted configuration for majority vote found by our algorithm.

From Table 1, we can conclude that fine-tuning using hard negatives has brought a significant improvement of 3.5% for monolingual and 6% for crosslingual. In addition, by implementing the algorithm for finding the optimal weighted configuration for majority vote, we have further improved by 3% the results on both tasks.

We discovered that adding the Eng and Orig models to the majority vote helped to improve our result even if they had worse accuracies than EngHard and OrigHard. As it can be seen from Table 2, the algorithm attributed them small weights.

The interesting part when looking at the algorithm's results is that the configuration is not always directly proportional to the models' accuracies. For example, in the msa task the EngHard model is a clear winner, but it is weighted less that

the OrigHard model by the algorithm.

For the new languages added in the test dataset we have used the average of the monolingual configurations : $[0.075, 0.125, 0.4375, 0.3625]$.

5 Results

Subtask	S@10
eng	0.941
fra	0.957
deu	0.987
por	0.960
spa	0.972
tha	1.000
msa	0.971
ara	0.948
monolingual	0.967
crosslingual	0.943

Table 3: Results on the dev dataset.

Subtask	S@10
eng	0.886
fra	0.920
deu	0.932
por	0.880
spa	0.962
tha	1.000
msa	1.000
ara	0.970
tur	0.910
pol	0.876
monolingual	0.934
crosslingual	0.790

Table 4: Results on the test dataset.

The dev results are very similar to the predicted results on the validation set, even achieving a perfect score for the Thai language.

The test results seem to have some differences, experiencing quite a large drop on the crosslingual accuracy. We believe that this is caused by the difference in distribution between the test dataset and the train and dev datasets.

Our team ranked 6^{th} in the monolingual track and 8^{th} in the crosslingual track.

6 Conclusion

Our training methodology accurately creates text embedding models with strong retrieval capabilities for fact-checking multilingual tasks for both monolingual and crosslingual scenarios. Fine-tuning using hard negatives greatly improves the accuracy of the models, while the algorithm for finding optimal weighted configurations for majority vote consistently outperforms naive approaches.

Our future work will revolve around testing the same training methodology on state of the art English text embedding models such as NV-Embed-v2 (Lee et al., 2025) (de Souza P. Moreira et al., 2024) or E5-mistral-7b-instruct (Wang et al., 2024a). In addition, we can also use such models in knowledge distillation processes in order to improve the accuracies of the smaller models used in this paper.

7 Acknowledgments

We would like to especially thank our professor Dumitru-Bogdan Alexe (University of Bucharest), for offering their guidance, expertize and advice in the writing process of this paper.

References

Gabriel de Souza P. Moreira, Radek Osmulski, Mengyao Xu, Ronay Ak, Benedikt Schifferer, and Even Oldridge. 2024. Nv-retriever: Improving text embedding models with effective hard-negative mining. *Preprint*, arXiv:2407.15831.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025. Nv-embed: Improved techniques for training llms as generalist embedding models. *Preprint*, arXiv:2405.17428.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Qiwei Peng, Robert Moro, Michal Gregor, Ivan Srba, Simon Ostermann, Marian Simko, Juraj Podroužek, Matúš Mesarčík, Jaroslav Kopčan, and Anders Søgaard. 2025. Semeval-2025 task 7: Multilingual and crosslingual fact-checked claim retrieval. In *Proceedings of the 19th International Workshop on Semantic Evaluation*, SemEval 2025, Vienna, Austria.

Matúš Pikuliak, Ivan Srba, Robert Moro, Timo Hromadka, Timotej Smoleň, Martin Melišek, Ivan Vykopal, Jakub Simko, Juraj Podroužek, and Maria Bielikova. 2023. Multilingual previously fact-checked claim retrieval. In *Proceedings of the 2023*

Conference on Empirical Methods in Natural Language Processing, page 16477–16500. Association for Computational Linguistics.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding. *Preprint*, arXiv:1807.03748.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Improving text embeddings with large language models. *Preprint*, arXiv:2401.00368.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

A Social media post example


```
"text_original" : 'El Senor Jesus
    nos dejo en Mateo 24: "Mirad que
     nadie os engane"',
"text_english" : 'The Lord Jesus
    left us in Matthew 24: "Beware
    lest anyone deceive you"',
"ocr_original" : ['lud Alni lud Anh
    #NoBajemosLaGuardia Gobierno del
     Por type', 'PEN) Gobierno del
    Perd NO HAY AGULA'],
"ocr_english" : ['crazy others crazy
     Anh #NoBajemosLaGuardia
    Gobierno del Por type', 'PEN)
    Government of Peru THERE IS NO
    AGULA']
```


B Fact-checked claim example


```
"claim_original" : '"Branca de Neve
    ". Disney vai excluir anoes da
    historia "para evitar ofensas a
    pessoas com nanismo"?',
"claim_english" : '"Snow White".
    Will Disney Exclude Dwarves From
     The Story "To Avoid Offense To
    People With Dwarfism"?',
"title_original" : '"Branca de Neve
    ". Disney vai excluir anoes da
    historia "para evitar ofensas a
    pessoas com nanismo"?',
"title_english" : '"Snow White".
    Will Disney Exclude Dwarves From
     The Story "To Avoid Offense To
    People With Dwarfism"?'
```