UMUTeam at SemEval-2025 Task 7: Multilingual Fact-Checked Claim Retrieval with XLM-RoBERTa and Self-Alignment Pretraining Strategy

Ronghao Pan, Tomás Bernal-Beltrán, José Antonio García-Díaz, Rafael Valencia-García
Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100 Murcia, Spain
{ronghao.pan, tomas.bernalb, joseantonio.garcia8, valencia}@um.es

Abstract

In today's digital age, the rapid dissemination of information through social networks poses significant challenges in verifying the veracity of shared content. The proliferation of misinformation can have serious consequences, influencing public opinion, policy decisions, and social dynamics. Fact-checking plays a critical role in countering misinformation; however, the manual verification process is time-consuming, especially in multilingual contexts. This paper presents our participation in the Multilingual and Crosslingual Fact-Checked Claim Retrieval task (SemEval 2025), which aims to identify previously fact-checked claims relevant to social media posts. We propose a retrieval approach based on XLM-RoBERTa, a multilingual Transformer model, combined with metric learning, hard negative mining, and a Multi-Similarity Loss function to optimize cross-lingual semantic representations. Our system uses a single multilingual encoder to handle all languages, offering a scalable and efficient solution without requiring language-specific adaptations. Although our final ranking (25th place) reflects modest performance compared to top systems, our model achieved consistent results across languages, with over 50% hit rate in most cases. These results highlight both the potential and current limitations of general-purpose multilingual models for fact-checking retrieval.

1 Introduction

The massive and continuous dissemination of information on social networks makes it increasingly difficult to distinguish between accurate content and misinformation (Bartolomé, 2021). This challenge is exacerbated by the fact that fake news can spread rapidly and influence public opinion, social behavior, and even political processes. Ensuring the veracity of online content is therefore critical to safeguarding societal well-being.

Manual fact-checking is a slow and resource-intensive task, especially in a multilingual context where claims and their verifications may appear in different languages. To address this challenge, recent research has proposed decomposing the process into subtasks that can be partially automated using Natural Language Processing (NLP), such as identifying check-worthy claims, retrieving relevant evidence, and assessing veracity (Guo et al., 2022; Pikuliak et al., 2023). Despite progress, automated fact-checking remains difficult due to linguistic ambiguity, implicit or biased wording, and the challenge of assessing the reliability of sources. Nevertheless, automating this process is essential to combat misinformation at scale and support a more informed and resilient society.

The Multilingual and Crosslingual Fact-Checked Claim Retrieval shared task (SemEval 2025) (Peng et al., 2025) focuses on the efficient identification of claims that have already been fact-checked in a multilingual and cross-lingual context. Given a social media post, the goal is to determine the most relevant corresponding to fact-check for the post. It is divided into two "subtasks", that is, the task could be done following two different setups: (1) **Setup A: Monolingual**. In which both posts and fact-checks are in the same language; and (2) **Setup B: Crosslingual**. In which posts and fact-checks may be in the same language or in different languages.

For this task, we propose an approach based on XLM-RoBERTa, a multilingual Transformer-based model, to generate contextual embeddings for both social media posts and fact-checked claims, enabling semantic comparisons across languages. Our approach employs a metric learning pipeline where the model is trained to optimize the cosine similarity between embeddings of relevant claim-post pairs, facilitating effective cross-lingual retrieval. To improve the model's discriminative ability, we incorporate a hard negative mining strategy

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 757–762

that dynamically selects challenging pairs that are semantically similar but do not correspond to the same fact-checked claim. This forces the model to learn more refined feature representations, improving its ability to distinguish between highly related but distinct claims. By leveraging this multilingual embedding space and fine-tuning it with task-specific constraints, our approach efficiently retrieves fact-checked claims, addressing the challenge of misinformation detection in a multilingual and cross-lingual setting.

2 Background

An automated fact-checking chain typically consists of several modules, each one in charge of performing one of the subtasks that make up the automation of the fact-checking process (Guo et al., 2022). Generally speaking, this process begins with the detection of whether a fact should be verified or not and ends, after identifying which facts can be verified with the same information, with the extraction, from some data source, of the original source or set of evidences that supports or refutes the fact.

Specifically, this task focuses on previously fact-checked claim retrieval (PFCR) (Shaar et al., 2022): that is, given a text making a claim and a set of claims that have already been verified, the objective is to rank the fact-checked claims so that those that are the most relevant with reference to the given claim. Following the process described above, this task would correspond to the last step of the automated fact-checking process.

Fact-checked claim retrieval has become an important research topic, mainly due to its potential applications in many challenging tasks such as the detection and correction of fake news and misinformation on digital platforms. This would allow to quickly identify this incorrect information, contrast it in real time, and improvements in search and recommendation systems by integrating mechanisms that prioritize verified and high quality information. Since automated systems can continuously crawl and monitor news websites, social networks and other online sources, they are capable of identify new claims and information as they emerge. Additionally, fact-checking requires prioritizing which claims should be checked. Advanced algorithms make these decisions based on factors such as virality, potential impact and source credibility (Nakov et al., 2021).

Once the information to be verified has been identified, machine learning models use historical data to pinpoint statements that may verify or contradict the fact being verified. This process consists of evaluating the similarity between each fact and previously verified facts, and assigning similarity scores to them, which significantly speeds up the extraction of verifying or contradicting statements. Moreover, rapid advancements in NLP in recent years have led to the emergence of many pre-trained Transformer-based models. These models are trained on vast corpora of unlabeled text and, thanks to their transfer learning capabilities, they can be adapted to various tasks, which makes them an interesting option for this task. They have been shown to be effective in extracting verifications and identifying contradictions for a given fact (Ünver, 2023).

Furthermore, pre-trained multilingual models based on Transformers, such as mBERT and XLM-RoBERTa, have proven to be effective in the task of crosslingual fact-checking (Kazemi et al., 2022). Trained on a large multilingual corpora, they are capable of handling multiple languages, making them capable of performing the extraction of information that verifies or contradict a fact even when the texts are in different languages.

Therefore, in this study, we tested a self-alignment pretraining strategy based on XLM-RoBERTa and metric learning. Our approach involves training a multilingual language model to match social media posts with relevant fact-checks by leveraging semantic representations. Using hard pair mining and optimization with a multi-similarity loss function, the model learns to identify difficult to distinguish examples, improving the matching accuracy between posts and claims. We fine-tuned the model on a shared embedding space and used MultiSimilarityLoss (Wang et al., 2019) to refine similarity and dissimilarity relationships.

3 System overview

Figure 1 illustrates the architecture of our system for claim identification in fact-checking publications. Our approach is based on a self-alignment pretraining method similar to SapBERT (Liu et al., 2021), which involves training a multilingual language model, such as XLM-RoBERTa (Conneau et al., 2019), through metric learning to match social media posts with relevant fact-checks. In this way, the trained model is able to identify seman-

tic representations through a process of hard pair mining and optimization using a multiple similarity loss function.

Specifically, the pipeline of our system is organized into several stages, as shown in Figure 1. First, data loading and processing are performed. For this, we use the three datasets provided by the organizers: `posts.csv`, which contains texts of social media posts, including OCR of each post; `fact_check.csv`, which gathers verified facts (claims); and `pairs.csv`, which links posts to relevant claims. Text encoding is then carried out using XLM-RoBERTa, which generates embeddings for both posts and claims. These embeddings capture the semantic meaning of the texts in a shared vector space, facilitating similarity matching.

The next step involves hard pair mining, using the MultiSimilarityMiner library to identify examples that present high similarity in the embedding space and are therefore more difficult to distinguish, as they are not present in the `pairs.csv` file. Once the pairs and hard pairs are identified, the model is trained using MultiSimilarityLoss as the loss function, which adjusts the similarity and dissimilarity relationships between pairs of embeddings. Additionally, AdamW is employed as an optimizer to efficiently update the model parameters.

Figure 2 shows the pipeline used for the evaluation of the model. A previously trained model was used to identify the most relevant claims related to social media posts. In order to speed up the search time, a dictionary of embeddings of the facts was created. These claims are tokenized and passed through the trained model to obtain their representations in the form of embeddings. These representations are stored in a list and concatenated with their corresponding `fact_id`. Once this fact dictionary is available, to extract the top 10 most related claims to an input post, the embedding of the post is obtained using the trained model and the distance between the embedding of the post and the pre-generated embeddings of the claims is calculated. From the distance matrix, the indices of the closest (i.e., most similar) facts are selected and the first 10 results are returned.

4 Experimental setup

For experimentation, the datasets provided by the organizers were used for both training and testing. In the training process, a set of 153,742 facts, 24,431 posts, and 25,742 pairings were available, as shown in Table 1. For the testing phase, the facts set consists of 272,446 examples, and the goal is to find the 10 most related facts for each post (with a total of 8,275 posts in the test set).

It is important to note that the facts set is considerably large, with 272,446 examples. However, with our model and approach, the search time is significantly reduced, as it is only necessary to extract the 10 embeddings of facts most related to the embedding of the input post. In contrast, many other current approaches would require a sequential search through the entire set of facts for each post, which is much less efficient.

For training, XLM-RoBERTa was used as the base model, together with MultiSimilarityLoss as the loss function and MultiSimilarityMiner for hard pair generation. In addition, AdamW was used as optimizer, with a learning rate of 2e-5 and a weight decay of 0.01. The model was trained for a total of 20 epochs.

Table 1: Dataset distribution

Type	Number
Train	
Pairs	25.742
Facts	153.742
Posts	24.431
Test	
Facts	272.446
Posts	8.275

5 Results

In this task of identifying the most relevant facts related to the posts, S@10 was used as the reference metric to evaluate the performance of the models. S@10 measures the proportion of posts for which at least one of the ten retrieved facts is relevant.

The detailed analysis of the results, presented in Table 2, reveals crucial information about the performance of the top three teams and our team in this task, including the official ranking in several languages. Overall, PINGAN AI leads the ranking with an S@10 (avg) of 0.96, standing out as the most effective model. It is followed by PALI and TIFIN India, with average scores of 0.9472 and 0.938 respectively, demonstrating their high accuracy in fact retrieval. In contrast, UMUTeam

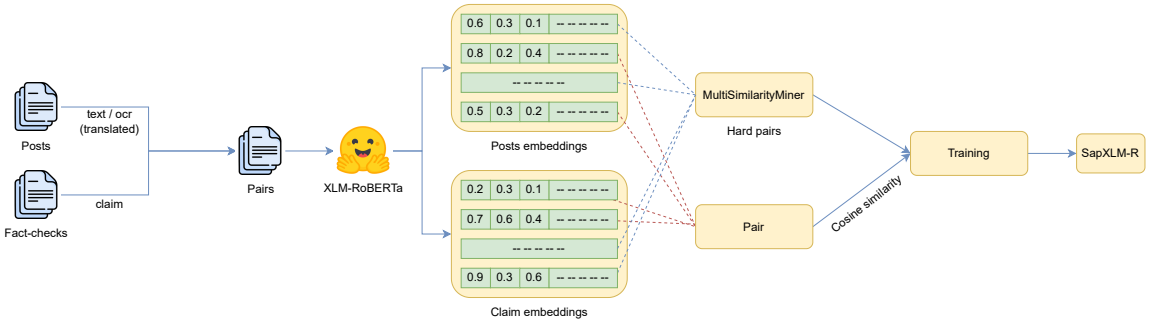

Figure 1: System architecture pipeline.

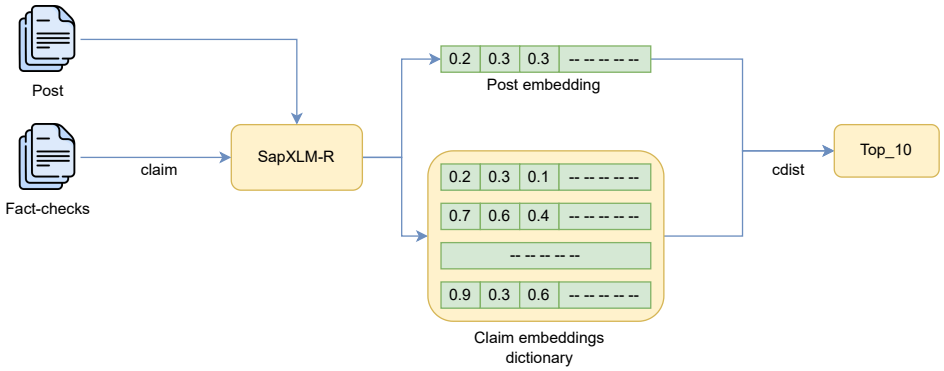

Figure 2: Evaluation approach pipeline.

ranks 25th with an S@10 (avg) of 0.544, reflecting a much lower performance compared to the leaders.

Analyzing the performance by language, it can be seen that in ENG, the top three teams show solid results, especially PINGAN AI (0.916) and PALI (0.904), while UMUTeam achieves only 0.414, indicating significant difficulties in this language. In SPA, PINGAN AI and PALI perform excellently (0.974 and 0.97, respectively), while UMUTeam scores only 0.42, suggesting considerable room for improvement. In THA and MSA, PINGAN AI, PALI and TIFIN India achieve perfect or near perfect scores, demonstrating exceptional performance. However, although UMUTeam shows relatively better performance in these languages (0.786 and 0.688 respectively), it still lags behind the leading teams. The lowest overall performance is observed in POL. While PINGAN AI and PALI score 0.926 and 0.888, respectively, UMUTeam scores only 0.464, which is its lowest score in this language assessment.

Analysis of these results shows that language diversity significantly affects model performance, with generally higher scores observed for THA and MSA, in contrast to ENG and POL. This variability suggests that the models used may not be equally optimized for all languages. Although our team (UMUTeam) did not achieve a high score in the ranking, our approach has notable advantages. Unlike other approaches that may require separate models or language-specific settings, our approach uses a single model for all languages and achieves a hit rate above 50% for most languages. It also has a relatively short search time, making it a more efficient and scalable solution.

6 Conclusion

Our participation in the Multilingual and Crosslingual Fact-Checked Claim Retrieval task (SemEval 2025) addressed the challenge of identifying relevant fact-checked claims for social media posts across multiple languages. We proposed a solution based on XLM-RoBERTa, leveraging its multilingual capabilities to generate contextual embeddings that enable semantic comparisons between posts and fact-checks. By incorporating a metric learning pipeline with hard negative mining, our approach effectively distinguished between highly related but distinct claims, improving cross-lingual retrieval accuracy.

Table 2: Official ranking reported by language and final average

Language	#1 PINGAN AI	#2 PALI	#3 TIFIN India	...	#25 UMUTeam
English	0.916	0.904	0.88	...	0.414
French	0.972	0.954	0.954	...	0.564
German	0.958	0.936	0.936	...	0.384
Portuguese	0.926	0.908	0.902	...	0.47
Spanish	0.974	0.97	0.96	...	0.42
Thai	0.9945	1.000	0.9945	...	0.7869
Malay	1.000	1.000	1.000	...	0.6882
Arabic	0.986	0.982	0.966	...	0.716
Turkish	0.948	0.93	0.904	...	0.538
Polish	0.926	0.888	0.886	...	0.464
Average	0.96	0.9472	0.938	...	0.544

The experimental results demonstrated that our system efficiently retrieved relevant claims across various languages, achieving a hit rate above 50% for most of them. Notably, the model showed competitive performance in Thai and Malay, highlighting its effectiveness in languages with simpler grammatical structures. However, the system's performance was lower in English and Polish, indicating challenges in processing complex linguistic nuances. These results underscore the impact of language diversity on model performance, suggesting the need for language-specific optimizations.

Despite not achieving top ranks, our unified model approach proved to be efficient and scalable, offering a significant advantage over language-specific solutions. The system's rapid search time and ability to handle multilingual input with a single model demonstrate its potential for real-world misinformation detection applications.

It is worth noting that we trained the model using the full training set provided by the organizers, without holding out a separate validation set. This decision was motivated by the retrieval nature of the task, where the goal was to optimize a embedding space to compare posts and fact-checked claims efficiently. While this choice allowed us to leverage all available data for training robust representations, it also limited our ability to perform detailed hyperparameter tuning and ablation studies. We acknowledge this limitation and consider incorporating a validation split and more granular experiments in future work. Besides, we will explore domain adaptation, cross-lingual transfer learning techniques, and addressing the challenges posed by languages with complex grammatical structures.

Finally, we aim to apply our fact-checked claim retrieval system to domains involving political discourse (Garcia-Díaz et al., 2023) and harmful or emotionally charged narratives, such as hate speech and hope speech (Pan et al., 2025a,b). In this sense, integrating multilingual claim retrieval into these pipelines could help validate or challenge politically biased or manipulative statements, especially in settings where real-time verification is critical. In future work, we plan to explore domain adaptation strategies and joint modeling approaches that combine claim verification with stance and intention detection, allowing for a more comprehensive analysis of ideological and affective discourse.

Acknowledgments

This work is part of the research project LT-SWM (TED2021-131167B-I00) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. This work is also part of the research project LaTe4PoliticES (PID2022-138099OB-I00) funded by MCIN/AEI/10.13039/501100011033 and the European Fund for Regional Development (ERDF)-a way to make Europe, and the research project "Services based on language technologies for political microtargeting" (22252/PDC/23) funded by the Autonomous Community of the Region of Murcia through the Regional Support Program for the Transfer and Valorization of Knowledge and Scientific Entrepreneurship of the Seneca Foundation, Science and Technology Agency of the Region of Murcia. Mr. Tomás Bernal-Beltrán is supported by University of Murcia through the predoctoral programme.

References

Mariano Bartolomé. 2021. Redes sociales, desinformación, cibersoberanía y vigilancia digital: una visión desde la ciberseguridad. *RESI: Revista de estudios en seguridad internacional*, 7(2):167–185.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

José Antonio Garcia-Díaz, Salud M Jiménez Zafra, María Teresa Martín Valdivia, Francisco García-Sánchez, Luis Alfonso Ureña López, and Rafael Valencia García. 2023. Overview of PoliticES at IberLEF 2023: Political Ideology Detection in Spanish Texts. *Procesamiento del Lenguaje Natural*, 71:409–416.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Ashkan Kazemi, Zehua Li, Verónica Pérez-Rosas, Scott A Hale, and Rada Mihalcea. 2022. Matching tweets with applicable fact-checks across languages. *arXiv preprint arXiv:2202.07094*.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. In *30th International Joint Conference on Artificial Intelligence, IJCAI 2021*, pages 4551–4558. International Joint Conferences on Artificial Intelligence.

Ronghao Pan, José Antonio García-Díaz, and Rafael Valencia-García. 2025a. Optimizing few-shot learning through a consistent retrieval extraction system for hate speech detection. *Procesamiento del Lenguaje Natural*, 74:241–252.

Ronghao Pan, José Antonio García-Díaz, and Rafael Valencia-García. 2025b. Spanish mtlhatecorpus 2023: Multi-task learning for hate speech detection to identify speech type, target, target group and intensity. *Computer Standards & Interfaces*, 94:103990.

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 Mária Bieliková. 2023. Multilingual previously fact-checked claim retrieval. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16477–16500.

Shaden Shaar, Nikola Georgiev, Firoj Alam, Giovanni Da San Martino, Aisha Mohamed, and Preslav Nakov. 2022. Assisting the human fact-checkers: Detecting all previously fact-checked claims in a document. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2069–2080.

Akın Ünver. 2023. Emerging technologies and automated fact-checking: tools, techniques and algorithms. *Techniques and Algorithms (August 29, 2023)*.

Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. 2019. Multi-similarity loss with general pair weighting for deep metric learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5017–5025. IEEE.