Shouth NLP at SemEval-2025 Task 7: Multilingual Fact-Checking Retrieval Using Contrastive Learning

Santiago Lares Harbin and Juan Manuel Pérez
Universidad de San Andrés
slaresharbin@udesa.edu.ar

Abstract

We present a multilingual fact-checking retrieval system for the SemEval-2025 task of matching social media posts with relevant fact checks. Our approach utilizes a contrastive learning framework built on the multilingual E5 model architecture, fine-tuned on the provided dataset. The system achieves a Success@10 score of 0.867 on the official test set, with performance variations between languages. We demonstrate that input prefixes and language-specific corpus filtering significantly improve retrieval performance. Our analysis reveals interesting patterns in cross-lingual transfer, with specifically strong results on Malaysian and Thai languages. We make our code public for further research and development.

1 Introduction

The proliferation of misinformation on social media is referred by social scientists as a threat to the integrity of democracy and the public sphere (Chambers, 2021; Lewandowsky et al., 2023) This scenario has increasingly created an urgent need for automated fact-checking systems that can retrieve relevant fact-checks for dubious claims. The SemEval-2025 Task (Shahi et al., 2025) addresses this challenge by requiring participants to develop systems that can retrieve the most relevant fact checks for social media posts from a large corpus of 272,447 fact checks among multiple languages.

In this paper, we present our approach to this task, which utilizes a fine-tuned multilingual embedding model based on the E5 architecture (Wang et al., 2024). Our system employs a contrastive learning framework to optimize the semantic matching between posts and fact checks on fine-tuning. We focus in particular on multilingual representation learning and cross-lingual retrieval functions. The code for our system is publicly available on GitHub[1].

Our main contributions include: (1) an analysis of various model architectures and their performance among different languages, (2) an examination of how input prefixes affect retrieval accuracy, and (3) a demonstration of the importance of language-specific corpus filtering for improving retrieval performance. Our system achieved an average Success@10 score of 0.867 on the official test set, ranking 20th out of 28 published participants.

2 Background

Automated fact-checking has emerged as a critical tool in combating the spread of misinformation online. (Zhou and Zafarani, 2020) speak about the immense volume of in- formation shared online, and suggest and evaluate different methods to detect fake news. The SemEval-2025 Task (Peng et al., 2025) focuses especially on multilingual fact-check retrieval, where the goal is to match social media posts with relevant fact-checks from a large multilingual corpus. Most automated fact-checking system consist of three steps (Guo et al., 2022): first, the claim de- tection (is this check-worthy?); then, the evidence retrieval (given a claim, retrieve relevant informa- tion to verify it); and lastly and claim verification (given a claim and evidence, define if the claim is true or false).

The task (Peng et al., 2025) has two subtasks: monolingual retrieval and cross-lingual retrieval. We focus only on the monolingual subtask.

Recent advances in multilingual embedding models have shown promising results for multilingual retrieval tasks (Feng et al., 2022; Litschko et al., 2022). These models learn a shared semantic space between languages, allowing for effective comparison of text regardless of the source language. In particular, contrastive learning approaches have proven effective for training such models (Chen et al., 2020), as they explicitly op-

[1] https://github.com/sanlares/semeval_2025

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 2265–2269
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

timize for similarity between related texts while pushing unrelated texts apart in the embedding space. In this paper, we present our retrieval approach to this task, which uses a fine-tuned multilingual embedding model based on the E5 architecture (Wang et al., 2024). This model has demonstrated strong performance on multilingual retrieval benchmarks (Zhang et al., 2021) by combining transformer-based encoders with contrastive pre-training. We additionally fine-tune this model on the task-specific data to improve its performance for fact-check retrieval.

3 System Overview

Our system addresses the challenge of retrieving relevant fact checks from a large corpus of 272,447 documents by implementing a fine-tuned contrastive learning approach based on the multilingual E5 model architecture. The system consists of three main components: (1) a text embedding model, (2) a contrastive learning framework, and (3) a retrieval pipeline.

The core of our system is built upon the `Multilingual E5 Model` (Wang et al., 2024), which we improve with a custom architecture. The model architecture is illustrated in Figure 1, consisting of a transformer encoder, mean pooling layer, dense projection layer with GELU activation, and L2 normalization for cosine similarity calculation.

The model processes both posts and fact checks using language-specific prefixes (`query:` and `passage:` respectively) to optimize the semantic matching between them.

We trained our model using the Multiple Negatives Ranking Loss (MNRL) (Jadwin and Huang, 2023), which can be formulated as:

$$L = -\log \frac{\exp(sim(x_i, x_i^+)/\tau)}{\sum_{j=1}^{N} \exp(sim(x_i, x_j)/\tau)} \quad (1)$$

where:

- x_i represents the anchor text embedding (post)
- x_i^+ represents the positive pair embedding (matching fact-check)
- x_j represents all other samples in the batch (negative pairs)
- τ is a temperature parameter that controls the sharpness of the distribution
- $sim(\cdot,\cdot)$ is the cosine similarity function

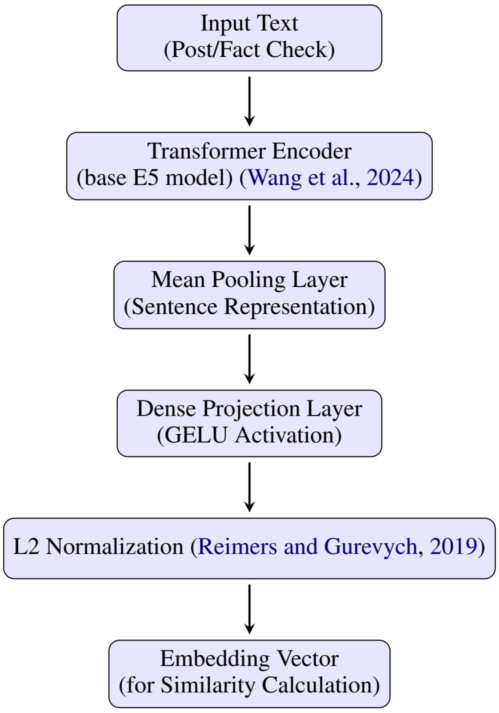
Input Text
(Post/Fact Check)
Transformer Encoder
(base E5 model) (Wang et al., 2024)
Mean Pooling Layer
(Sentence Representation)
Dense Projection Layer
(GELU Activation)
L2 Normalization (Reimers and Gurevych, 2019)
Embedding Vector
(for Similarity Calculation)

Figure 1: Architecture of the embedding model used in our system. The model processes both posts and fact checks using language-specific prefixes to optimize semantic matching.

We used in-batch negatives for computational efficiency while maintaining effective contrastive learning. The temperature parameter τ is dynamically adjusted during training to optimize the learning process.

The retrieval process follows these steps:

1. **Language Detection**: The system first identifies the language of the input post.
2. **Corpus Filtering**: The fact-check corpus is filtered to prioritize documents in the same language as the input post.
3. **Embedding Generation**: The model generates embeddings for both the input post and the filtered fact checks.
4. **Similarity Ranking**: The system computes Pearson cosine similarity between the post and fact check embeddings.
5. **Top-K Selection**: The 10 most similar fact checks are retrieved and ranked.

This pipeline is set for both monolingual and cross-lingual retrieval, with special consideration for language-specific distinctions in the embedding space.

4 Experimental Setup

4.1 Dataset

The dataset used for this study was derived from the official SemEval-2025 competition data. We divided the training set into training (80%) and validation (20%) subsets while maintaining the language distribution for both sets.

For preprocessing, we utilized the `load.py` file recommended by the competition organizers. Our approach involved creating semantic representations for both posts and fact checks. For posts, we concatenated the OCR-extracted text (`ocr` column) with the manual transcription (`text` column) from the `posts.csv` file. Similarly, for fact checks, we combined the `claim` and `title` columns from the `fact_checks.csv` file. For model input, relevant prefixes (e.g., `query:`, `passage:`) were prepended to these concatenated texts prior to tokenization to guide the model's contextual understanding.

In addition to fine-tuning our primary model, we conducted experiments with various base models available in the SentenceTransformers framework, testing different prefix combinations that yielded varying performance results.

4.2 Implementation Details

The system was implemented using the following configuration:

- **Base Model**: multilingual-e5-large-instruct
- **Framework**: SentenceTransformers (v3.4.1)
- **Deep Learning Backend**: PyTorch (v2.5.1)
- **Python Version**: 3.12.7
- **Computing Environment**: macOS 15.1 (24B83)

The model was trained with the following hyperparameters:

The primary evaluation metric for our system was Success@10 (S@10), which measures the proportion of posts for which the correct fact check appears in the top 10 retrieved results. We evaluated the model's performance in both monolingual and cross-lingual settings:

Parameter	**Value**
Maximum Sequence Length	512 tokens
Pooling Mode	Mean
Embedding Dimension	384
Batch Size	4
Evaluation Batch Size	16
Gradient Accumulation Steps	2
Learning Rate	2e-5
Temperature	0.05
Mixed Precision Training	Enabled

Table 1: Model training hyperparameters

5 Results

We evaluate our system's performance on both the validation and test sets, analyzing the impact of different model configurations and language-specific performance. Our system achieved an average Success@10 score of **0.867** on the official test set, ranking 20th out of 28 published participants.

The results demonstrate considerable variability in model performance across languages, with Malaysian (msa) showing the strongest performance at 0.978 and English (eng) and Portuguese (por) showing the lowest at 0.796. This disparity suggests that linguistic factors may play a significant role in retrieval performance.

Table 3 shows the performance comparison of different base models and input prefix configurations on the validation set. The results demonstrate that the multilingual-e5-large-instruct model with appropriate prefixes consistently outperforms other configurations. The base model $intfloat/multilingual - e5 - large - instruct$ (Wang et al., 2024) improved from an average Success@10 of 0.834 to 0.867 when fine- tuned with the training dataset, along with the rec- ommended query and passage prefixes. This im- provement was consistent for all languages tested in the competition.

5.1 Impact of Input Prefixes

Our experiments revealed that the choice of input prefixes significantly impacts model performance. Using the recommended prefixes on multilingual-e5-large-instruct base model ('passage:' and 'query:') improved the average Success@10 score by 0.013 points (from 0.821 to 0.834) compared to using no prefixes. This improvement was consistent across all languages, with the largest gains observed in:

- English: +0.022 (from 0.766 to 0.788)
- Spanish: +0.012 (from 0.857 to 0.869)

Language	Success@10	Relative Performance
Malaysian (msa)	**0.978**	+12.8%
Thai (tha)	0.940	+8.4%
Arabic (ara)	0.912	+5.2%
French (fra)	0.894	+3.1%
Spanish (spa)	0.866	-0.1%
German (deu)	0.858	-1.0%
Turkish (tur)	0.828	-4.5%
Polish (pol)	0.806	-7.0%
English (eng)	0.796	-8.2%
Portuguese (por)	0.796	-8.2%
Average	0.867	–

Table 2: Official test set results of the fine-tuned multilingual-e5-large-instruct model by language. Languages are ordered by Success@10 score. Best performance is shown in **bold**. The model shows strong performance on Malaysian and Thai languages.

Base Model	Posts Prefix	Facts Prefix	Success@10
`multilingual-e5-large-instruct` (Wang et al., 2024)	passage:	query:	**0.834**
`multilingual-e5-large-instruct` (Wang et al., 2024)	–	–	0.821
`multilingual-e5-small` (Wang et al., 2024)	–	–	0.795
`e5-large-v2` (Wang et al., 2022)	–	–	0.758
`modernbert-embed-base` (Nussbaum et al., 2024)	`search_query:`	`search_document:`	0.779
`paraphrase-multilingual-mpnet-base-v2` (Reimers and Gurevych, 2019)	–	–	0.736

Table 3: Performance comparison of base models on the validation set. Models are ordered by Success@10 score. Best result is shown in **bold**.

- Portuguese: +0.018 (from 0.781 to 0.799)

These results suggest that appropriate input formatting plays a vital role in improving the model's cross-lingual understanding and retrieval functions.

5.2 Corpus Reduction Analysis

Our experiments showed that performance improves significantly when the fact-check corpus is reduced in size. Specifically, when using only the subset of approximately 16,000 fact checks corresponding to the validation set (versus the full corpus of 272,447), Success@10 scores improved by an average of 10.2%. This improvement demonstrates the impact of corpus size on retrieval accuracy. For more detailed experiments on corpus reduction with the fine-tuned model, see Table 4 in Appendix A.

Our analysis revealed several key outcomes:

- The base model multilingual-e5-large-instruct consistently outperforms other options.
- The implemented model training architecture is well-suited for the provided dataset training, allowing to obtain a fine-tuned model that outperforms the best base model tested in experiments.
- Appropriate input prefixes can provide substantial performance gains, improving Success@10 by 0.013 points on average.
- There is considerable variability in model performance between languages, suggesting that language-specific tuning could yield extended improvements.
- Corpus reduction by language dramatically improves retrieval accuracy, highlighting the importance of effective pre-filtering strategies.

Subsequent work could focus on addressing the performance discrepancy among languages, perhaps through language-specific fine-tuning or more sophisticated cross-lingual transfer techniques. Moreover, exploring ensemble methods that combine multiple embedding models might improve retrieval accuracy for challenging languages.

6 Acknowledgments

We would like to thank the SemEval-2025 task organizers for creating this challenging and impactful

task, and for providing thorough evaluation metrics and datasets.

References

Simone Chambers. 2021. Truth, deliberative democracy, and the virtues of accuracy: Is fake news destroying the public sphere? *Political Studies*, 69(1):147–163.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. *A Simple Framework for Contrastive Learning of Visual Representations*, volume 119 of *PMLR*, pages 1597–1607. Virtual. Editors: Daumé, Hal and Singh, Aarti.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. *Preprint*, arXiv:2007.01852.

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

Addison Jadwin and Catherine Huang. 2023. Improving minbert performance on multiple tasks through in-domain pretraining, negatives ranking loss learning, and hyperparameter optimization.

Stephan Lewandowsky, Ullrich K.H. Ecker, John Cook, Sander van der Linden, Jon Roozenbeek, and Naomi Oreskes. 2023. Misinformation and the epistemic integrity of democracy. *Current Opinion in Psychology*, 54:101711.

Robert Litschko, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2022. On cross-lingual retrieval with multilingual text encoders. *Information Retrieval Journal*, 25(2):149–183.

Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed: Training a reproducible long context text embedder. *Preprint*, arXiv:2402.01613.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Gautam Kishore Shahi, Preslav Nakov, Giovanni Da San Martino, Wenjun Gao, Firoj Alam, Hamdy Mubarak, Arthur Brack, Ussama Yaqub, Bader Alshemali, Fahim Alam, Sara Mourad, Temitope Ndapa Nakashole, Tien Lahouasse, Oliver Kutz, Ralf Möller, and Kalina Bontcheva. 2025. Semeval-2025 task 1: Multilingual and cross-lingual fact-checking retrieval. In *Proceedings of the 19th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

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. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. tydi: A multi-lingual benchmark for dense retrieval. *Preprint*, arXiv:2108.08787.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.*, 53(5).

A Corpus Reduction Analysis

Table 4 shows the impact of corpus reduction on the fine-tuned model's performance. The results demonstrate that using language-specific subsets of the fact-check corpus significantly improves retrieval performance.

Model	Posts Prefix	Facts Prefix	S@10
multi-e5	query:	passage:	**0.934**
multi-e5	–	–	0.933
multi-e5	post:	fact check:	0.934
multi-e5	This is a post...	This is a fact...	0.936
multi-e5	<[language]>	–	0.933
multi-e5	The following...	The following...	0.934

Table 4: Fine-tuned model performance with different configurations when using a reduced corpus. All experiments were conducted on the validation set with a corpus containing only fact checks in the same language as the query posts, resulting in higher overall scores compared to the full corpus evaluation.

Our experiments showed that performance improves significantly when the fact-check corpus is reduced to include only documents in the same language as the query post. For example, when retrieving from a language-specific subset of approximately 16,000 fact checks (versus the full corpus of 272,447), Success@10 scores improved by an average of 10.2%.