CAISA at SemEval-2025 Task 7: Multilingual and Cross-lingual Fact-Checked Claim Retrieval

Muqaddas Haroon[1] **Shaina Ashraf**[1,2] **Ipek Baris Schlicht**[3,4] **Lucie Flek**[1,2,5]
[1]University of Bonn, Germany
[2]Lamarr Institute for Machine Learning and Artificial Intelligence, Germany
[3]Universitat Politecnica de Valencia, Valencia, Spain
[4]DW Innovation, Bonn, Germany
[5]Bonn-Aachen International Center for Information Technology (b-it), Germany
s85mharo7@uni-bonn.de

Abstract

This paper describes our approach to the SemEval-2025 Task 7: Multilingual and Cross-lingual Fact-Checked Claim Retrieval on cross-lingual data. In this study, we developed a system to tackle the verified claim retrieval task, where the objective is to identify relevant, previously fact-checked claims from multilingual data based on a given input claim. We leveraged LLaMA, utilizing its ability to evaluate the relevance of retrieved claims within a retrieval-based fact-checking framework. This approach aimed to explore the impact of large language models (LLMs) on retrieval tasks and assess their effectiveness in enhancing fact-checking accuracy. Additionally, we integrated various embeddings, including e5-large, Jina embeddings, and the MPNet multilingual sentence transformer, to filter and rank a set of 500 candidate claims. These refined claims were then used as input for LLaMA, ensuring that only the most contextually relevant ones were assessed. Our team in the cross-lingual track scored s@10 0.57.

1 Introduction

The rapid proliferation of misinformation in recent years has raised significant concerns across governments, organizations, and individuals. False information spreads rapidly through online platforms, influencing public opinion and decision-making(Vosoughi et al., 2018). To combat this, numerous fact-checking organizations, such as FactCheck.org and Snopes, have emerged to manually verify claims. However, manual fact-checking is labor-intensive and struggles to keep pace with the high volume of misinformation circulating online (Thorne and Vlachos, 2018).

Misinformation is a global issue, requiring fact-checking solutions that extend beyond monolingual settings. To address this challenge, researchers have developed automated fact-checking systems, including verified claim retrieval, where the goal is to find previously fact-checked claims relevant to an input claim (Mansour et al., 2023). This task is particularly crucial as many false claims resurface in different forms across time and languages. While prior work has largely focused on monolingual claim retrieval, real-world misinformation often transcends language barriers, making multilingual claim retrieval a pressing issue.

As illustrated in Figure 1, we tackle multilingual verified claim retrieval by leveraging LLaMA (Touvron et al., 2023), a large language model (LLM) capable of understanding and retrieving fact-checked claims across different languages. To enhance retrieval precision, we incorporate Jina embeddings v2 (Günther et al., 2023)and the MPNet multilingual sentence transformer (Song et al., 2020), filtering and ranking candidate claims before passing them to LLaMA. By extending verified claim retrieval to a multilingual setting, we contribute to developing cross-lingual fact-checking systems that help mitigate misinformation more effectively.

2 Related Work

Multilingual Claim Retrieval. Pikuliak et al. (2023) introduced MultiClaim, the largest dataset for multilingual claim retrieval, featuring 28k posts (27 languages) and 206k fact-checks (39 languages). Their study shows that supervised fine-tuning improves retrieval performance over unsupervised methods.

Monolingual Claim Matching. Previous work has explored retrieval-based ranking for claim matching, using BERT and BM25 to rank check-worthy claims (Shaar et al., 2020). However, their dataset focused only on political claims in monolingual settings.

Cross-Lingual Claim Matching. Efforts to match social media claims with fact-checks across languages have utilized XLM-RoBERTa, BM25, and

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1377–1382
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

LaBSE (Kazemi et al., 2022). While effective monolingually, cross-lingual performance remains a challenge.

Multimodal Fact-Checking. *RAGAR: Your Falsehood Radar* explores Retrieval-Augmented Generation (RAG) for political claims, integrating textual and image inputs. While CoRAG and ToRAG reasoning techniques improve verification, reliance on GPT-4V (closed-source) limits adaptability (Khaliq et al., 2024).

End-to-End Multimodal Verification. An SBERT-BART framework combines evidence retrieval, claim verification, and explanation generation (Yao et al., 2023). However, SBERT struggles with cross-modal reasoning, affecting its ability to integrate text and visual data effectively.

3 Task Definition and Dataset

3.1 Task

The task at hand is for us to develop an effective system that is capable of retrieving previously fact-checked claims across multiple languages using a large language model (LLM), specifically leveraging LLaMA-3 8 B-Instruct (Large Language Model Meta AI). We selected LLaMA-3 8B-Instruct (Touvron et al., 2023) for our post-filtering step due to its strong performance in instruction-following tasks, especially in natural language inference (NLI) and zero-shot classification, which aligns closely with determining whether a fact-check is relevant to a given social media post. Our purpose is to increase the effectiveness in the identification and validation of claims associated with posts in different languages, using a model trained to differentiate between correct and incorrect claims based on historical data.

Our trained system will process a set of claims and determine whether each claim is relevant to the post. Our objective of this research is to develop an automated fact-checking system that can process multilingual claims and retrieve the most relevant facts for verification. The metric we used, "success@k," is a family of metrics that focuses on the performance of the top-k retrieved results We use this as it is crucial because, in many real-world scenarios, users primarily interact with the first few results. To be concise, Success@k metrics provide a way to evaluate the effectiveness of retrieval systems by focusing on the quality of the top-k results.

3.2 Dataset

This study utilizes the MultiClaim dataset, a large-scale multilingual dataset specifically designed for previously fact-checked claim retrieval (Peng et al., 2025). The dataset was built on top of the dataset by Pikuliak et al. (2023), which consists of 28,092 social media posts in 27 languages, 205,751 professionally fact-checked claims across 39 languages, and 31,305 verified post-claim connections, making it the most extensive and linguistically diverse dataset of its kind. The task organizers further introduced new data, including Turkish and Polish as new languages.

The dataset includes machine-translated versions of each post and claim, along with relevant metadata, enabling cross-lingual retrieval. The social media posts, primarily sourced from platforms such as Twitter and Facebook, frequently reference key political figures (e.g., Donald Trump) and organizations like the World Health Organization (WHO). The maximum post length in the dataset is 1,250 words, while fact-checked claims, particularly those related to political discourse, have a maximum length of approximately 600 words.

For evaluation purposes, we utilize the predefined post-claim pairings curated by the Task organizers, ensuring high-quality ground-truth data for retrieval-based fact-checking (Pikuliak et al., 2023).

4 System Overview

4.1 Pre-processing

Our preprocessing pipeline Figure 1 begins by cleaning and organizing raw posts and fact-checking claims, separating the original text, translations, and scores, while replacing missing "text" with OCR-extracted content. Next, we enrich the dataset by incorporating claim titles and summaries, merging all information based on predefined post-claim pairs, and splitting it into training, validation, and test sets.

Simultaneously, we extract names from URLs as extra metadata, clean the text, and generate summaries using LSA (latent semantic analysis) of the combined fields in our dataset using basic text analysis techniques. The final dataset undergoes language-specific cleaning (via spaCy, Stanza, and regex), alongside Unicode normalization, emoji

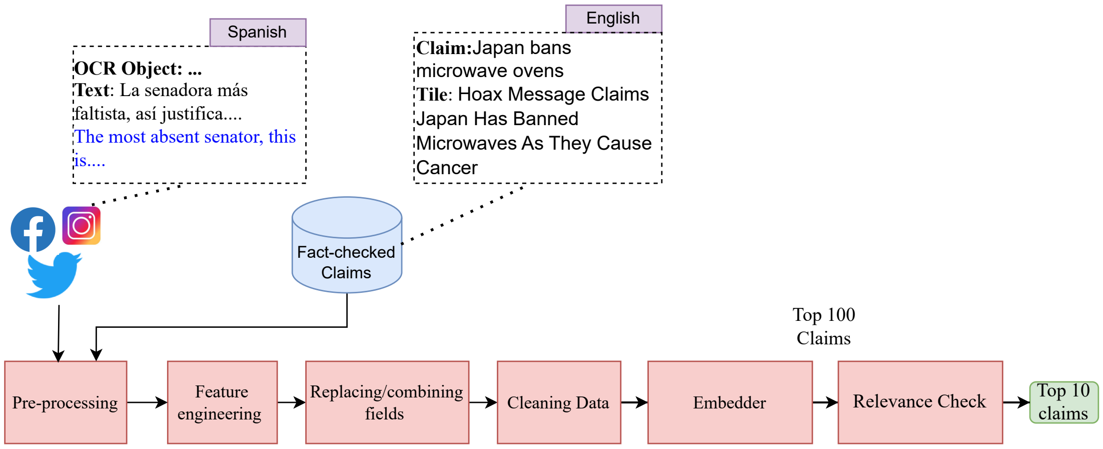
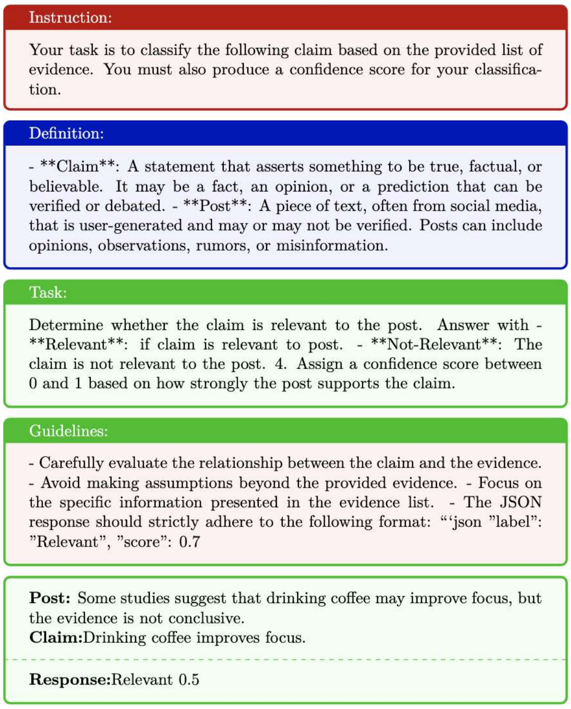

Spanish
OCR Object: ...
Text: La senadora más faltista, así justifica....
The most absent senator, this is....
English
Claim:Japan bans microwave ovens
Tile: Hoax Message Claims Japan Has Banned Microwaves As They Cause Cancer
Fact-checked Claims
Top 100 Claims
Pre-processing
Feature engineering
Replacing/combining fields
Cleaning Data
Embedder
Relevance Check
Top 10 claims

Figure 1: A complete overview of the system. We apply pre-processing steps for cleaning pairs of posts and fact-checked claims, vectorize them with Jina embeddings and finally rerank them with Llama.

conversion, and translation (if needed) to ensure standardized, low-noise text. These steps enhance the quality of embedding generation and summarization for downstream tasks.



Instruction:

Your task is to classify the following claim based on the provided list of evidence. You must also produce a confidence score for your classification.

Definition:

- **Claim**: A statement that asserts something to be true, factual, or believable. It may be a fact, an opinion, or a prediction that can be verified or debated. - **Post**: A piece of text, often from social media, that is user-generated and may or may not be verified. Posts can include opinions, observations, rumors, or misinformation.

Task:

Determine whether the claim is relevant to the post. Answer with - **Relevant**: if claim is relevant to post. - **Not-Relevant**: The claim is not relevant to the post. 4. Assign a confidence score between 0 and 1 based on how strongly the post supports the claim.

Guidelines:

- Carefully evaluate the relationship between the claim and the evidence. - Avoid making assumptions beyond the provided evidence. - Focus on the specific information presented in the evidence list. - The JSON response should strictly adhere to the following format: “‘json ”label”: ”Relevant”, ”score”: 0.7

Post: Some studies suggest that drinking coffee may improve focus, but the evidence is not conclusive.
Claim: Drinking coffee improves focus.

Response: Relevant 0.5

Figure 2: Prompt that we used for relevance check

4.2 Prompting Methodology

The related works showed that LLMs performed better when the prompt had details (Peskine et al., 2023). Our approach Figure 2 uses a structured, instruction-based prompt (zero-shot) with a task description to classify text into "relevant" or "not relevant" on similarity. We use zero-shot without description to assess the abilities of our prompt if no definition is provided as well. With these prompts, we can understand how adding details to the prompt increases the accuracy of our result. We further ask the model to give us a score according to the relevancy, and according to that we sort our claims for our top 10 result.

4.3 Evaluation

The system will take a given social media post, generate its semantic representation using state-of-the-art embedding models, and compare it against a database of verified claims. The system is designed to work in multiple stages, beginning with data preprocessing, where both posts and claims are normalized and tokenized to ensure consistency.

4.3.1 Models

Next, multilingual embeddings are generated separately for posts and claims to capture semantic similarities across different languages. We use embeddings for research such as jina embeddings, multilingual-e5-large-instruct and mpnet-multilingual: Jina embeddings v2 by Günther et al. (2023), which relies on English translations and filters out 500 claims for our LLaMA model; and second, paraphrase-multilingual-mpnet-base-v2, which processes posts and claims in their original languages and filters out 500 claims. MPNet introduces a novel pre-training approach that combines the strengths of BERT and XLNet while addressing their respective limitations, which is better for our embedding generation (Song et al., 2020). Lastly multilingual-e5-large-instruct (Wang et al., 2024), which is a state of the art multilingual text embedding model based on xlm-roberta-large. It is instruction-tuned, enhancing the quality of the embeddings.

4.3.2 Claim Ranking and Retrieval:

As mentioned before, we have utilized Llama as our model for selecting the top 10 claims for each post. For this purpose, we use our Llama model in a zero-shot setting (with and without task description). To refine the retrieved claims, an additional verification step using our LLaMA model determines whether the claim is relevant or not, along with a score ranging from 1 to 0, where 1 indicates it is highly relevant and 0 indicates it is not relevant at all. The final output is a ranked list of fact-checks, providing users with the most relevant and reliable information to verify the claim. The ranking is done by utilising the score provided by Llama. The system is designed to be efficient, scalable, and applicable across multiple languages, addressing the growing challenge of misinformation in the digital space.

4.3.3 Experimental Setup

We use Llama as we want to explore the possibility of how well can LLM perform as a fact retriever on different prompt settings. As mentioned above, we use MPNet multilingual and multilingual e5-large [1] for a purely cross-lingual setting that captures nuanced language variations, making it highly effective in cross-lingual claim retrieval. It helps in providing a balance between semantic understanding and efficiency, improving relevance ranking. Its context-aware embeddings ensure better alignment across different languages. We use Jina embeddings [2] for retrieving the top 500 claims using English translations and then map them to the original language, which is then given to Llama. For retrieval, cosine similarity is computed between a given post and the fact-check claims, and the top 500 claims with the highest similarity scores are selected for further evaluation.

5 Results

We are using success@10 as our metric for explaining how well our model performed. This measures the probability or percentage of times that at least one relevant item is found within the top 'k' results of a ranked list. If our retrieved result contains the correct claim from the top 10 claims, we add to our score 1 else, it is 0. LLaMA excels in multilingual claim retrieval due to its strong contextual and semantic understanding. Its attention-based architecture helps detect reworded claims across languages. We find that adding details to the prompt further improves accuracy by aligning it with fact-checking datasets.

Metric Success@10	Value
Llama zero-shot	11%
Jina	55.2%
e5-large	54.3%
MPnet	22%
Llama + Jina (task description)	57.7%
Llama + e5-large(task description)	54.3%
Llama + MPNet (task description)	38.6%

Table 1: Summary of Results. The fine-tuning

We evaluated zero-shot learning, MpNet, e5-large and Jina embedding for claim verification. According to our results table 1, simple **Zero-shot** learning performed poorly as the model failed to understand the task without details, but **zero-shot with task description** was better in results. We also see that the choice of embedders plays an important part in our accuracy, as E5-large performs better than MpNet, and Jina embeddings are much better than E5-large. Jina embedding outperformed both, excelling in semantic similarity and factual consistency, making it more effective for nuanced claim verification. However, it still faced challenges with implicit entailment, requiring external world knowledge.

Our results showed that LLaMA provided a success value of 57.7% using Jina embeddings based on the metric success@10. Although a few competitors achieved higher rankings, our system performed well overall.

6 Conclusion

In conclusion, our approach to retrieving previously fact-checked claims using the LLaMA model across multiple languages has shown promising

[1] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

[2] https://huggingface.co/jinaai/jina-embeddings-v2-base-en

results, achieving a success@10 value of 55.7 percent. This demonstrates that LLaMA can effectively assist in automating fact-checking tasks, making the process more efficient and scalable. LLaMA performs well in multilingual claim retrieval because it has a strong understanding of context and semantics across different languages. As a large language model (LLM), it is trained on diverse multilingual data, allowing it to recognize paraphrased or reworded claims. Its attention-based architecture helps it detect long-range dependencies, making it effective in finding related fact-checked claims even when the wording differs. Additionally, using LLaMA with task description prompt on claim verification tasks improves its accuracy by aligning its understanding with real-world fact-checking datasets. While the current success rate is encouraging, further refinements and optimizations are needed to improve the accuracy and reliability of claim retrieval across diverse languages.

Despite its strengths, LLaMA has some limitations. It sometimes overgeneralizes, meaning it might incorrectly match a claim with a similar but factually incorrect statement. Additionally, since LLaMA is not a dedicated fact-checking model, it may hallucinate relationships between claims that do not exist.

Acknowledgments

This work is supported by the DeFaktS project funded by the German Federal Ministry of Education and Research (BMBF), and by the Lamarr Institute for Machine Learning and Artificial Intelligence.

References

Michael Günther, Jackmin Ong, Isabelle Mohr, Alaeddine Abdessalem, Tanguy Abel, Mohammad Kalim Akram, Susana Guzman, Georgios Mastrapas, Saba Sturua, Bo Wang, and 1 others. 2023. Jina embeddings 2: 8192-token general-purpose text embeddings for long documents. *arXiv preprint arXiv:2310.19923*.

Ashkan Kazemi, Zehua Li, Verónica Pérez-Rosas, Scott A. Hale, and Rada Mihalcea. 2022. Matching tweets with applicable fact-checks across languages. In *Proceedings of the Workshop on Multi-Modal Fake News and Hate-Speech Detection (DE-FACTIFY 2022) co-located with the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI 2022), Virtual Event, Vancouver, Canada, February 27, 2022*, volume 3199 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Mohammed Abdul Khaliq, Paul Yu-Chun Chang, Mingyang Ma, Bernhard Pflugfelder, and Filip Miletić. 2024. RAGAR, your falsehood radar: RAG-augmented reasoning for political fact-checking using multimodal large language models. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 280–296, Miami, Florida, USA. Association for Computational Linguistics.

Watheq Mansour, Tamer Elsayed, and Abdulaziz Al-Ali. 2023. This is not new! spotting previously-verified claims over twitter. *Information Processing & Management*, 60(4):103414.

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. Association for Computational Linguistics.

Youri Peskine, Damir Korenčić, Ivan Grubisic, Paolo Papotti, Raphael Troncy, and Paolo Rosso. 2023. Definitions matter: Guiding GPT for multi-label classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4054–4063, Singapore. Association for Computational Linguistics.

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.

Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. 2020. That is a known lie: Detecting previously fact-checked claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3607–3618, Online. Association for Computational Linguistics.

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.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

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, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2733–2743.