Cross-Lingual Fact-Check Retrieval Using Contrastive Learning and Knowledge Distillation

Web Retrieval and Mining Final Project Proposal

Yi-Xuan JIANG ID: b10902010@ntu.edu.tw

James ID: r13921031@ntu.edu.tw

Ting-Kuan HSIEH ID: b10701166@ntu.edu.tw

1 Introduction to the Problem

Misinformation continues to spread rapidly across global social media platforms, transcending language barriers and posing significant challenges to fact-checkers worldwide. When a potentially misleading claim emerges, fact-checkers must determine if the claim has been previously fact-checked—a task that becomes exponentially more difficult when considering multiple languages. The SemEval-2025 Task 7 addresses this critical challenge by focusing on multilingual and crosslingual fact-checked claim retrieval.

The objective of our project is to develop an effective system that can retrieve relevant fact-checked claims for given social media posts across languages. This will help fact-checkers quickly identify if a claim has been previously verified or debunked, even when the original claim and fact-check exist in different languages. Our system aims to bridge the language gap and improve the efficiency of fact-checking workflows.

This problem is worth solving for several compelling reasons:

1. **Resource Optimization**: Fact-checking organizations have limited resources. Automated retrieval systems can help prioritize truly novel claims over recycled misinformation.

2. **Cross-Border Information Flow**: Misinformation often spreads across language boundaries, being translated and modified as it moves between communities. A crosslingual system can identify these patterns.

3. **Timely Response**: The speed of response is crucial in combating viral misinformation. Automated multilingual retrieval can significantly reduce the time needed to identify previously fact-checked claims.

4. **Broader Coverage**: Many fact-checking organizations can only monitor content in specific languages. A crosslingual system expands their effective coverage.

5. **Societal Impact**: By accelerating and improving fact-checking processes, our work contributes to reducing the harmful effects of misinformation on public discourse and decision-making.

By participating in this SemEval shared task, we aim to advance the state-of-the-art in multilingual and crosslingual fact-checking technologies while developing practical tools that can be integrated into real-world fact-checking workflows.

2 Methodology

Our approach to solving this multilingual and crosslingual fact-checked claim retrieval challenge will combine several advanced NLP techniques. We plan to explore the following methodologies:

2.1 Multilingual Dense Retrieval Architecture

We will implement a dual-encoder architecture based on pre-trained multilingual language models such as XLM-RoBERTa or mBERT. This approach will encode both the social media posts and fact-checked claims into a shared embedding space where similarity can be efficiently computed.

To enhance the multilingual capabilities, we'll fine-tune the model using:

- **Contrastive Learning**: We'll employ in-batch negatives and hard negative mining to train the model to distinguish between relevant and irrelevant fact-checks. For each post, we'll treat matched fact-checks as positive examples and other fact-checks as negative examples.
- **Knowledge Distillation**: We plan to leverage larger teacher models (potentially multilingual-T5 or BLOOM) to distill knowledge into our more efficient student models.
- **Translation-Augmented Training**: To address the crosslingual challenges specifically, we'll augment our training data with translations, teaching the model to identify semantic equivalence across languages.

2.2 Retrieval Enhancement Techniques

To improve retrieval performance, we'll implement:

- **Query Expansion**: Using multilingual WordNet or BabelNet to add synonyms and related terms to the query.
- **Reranking**: A two-stage retrieval approach where initial candidates are retrieved using lightweight BM25 or semantic search, then reranked using a more sophisticated cross-encoder model.
- **OCR and Multimodal Integration**: We'll develop techniques to leverage the OCR content from images, integrating this textual information with the post text for a more comprehensive representation.

2.3 Language-Specific Components

To address language-specific nuances:

- **Language Identification**: We'll implement robust language detection to route content to specialized models when appropriate.
- **Language-Specific Preprocessing**: Customize tokenization and normalization for different language families.

2.4 Anticipated Challenges

We anticipate several challenges:

- **Semantic Drift Across Languages**: Concepts may not map perfectly between languages, requiring techniques that can bridge these gaps.
- **Computational Efficiency**: Balancing model performance with practical inference speed requirements.
- **Data Imbalance**: Some languages may have significantly more training examples than others.

- **Cultural Context**: Misinformation often relies on cultural references that may not translate well across languages.
- **Translation Quality**: The quality of translated text in the dataset may vary and impact model performance.

3 Experiments

3.1 Evaluation Metrics

In line with the SemEval task requirements, we will evaluate our system primarily using success@10 (whether the correct fact-check appears in the top 10 retrieved results). Additionally, for our internal development, we'll track:

- Mean Reciprocal Rank (MRR)
- Normalized Discounted Cumulative Gain (NDCG)
- Precision and Recall at various cutoffs (P@k, R@k)

3.2 Ablation Studies

We plan to conduct ablation studies to understand the contribution of each component:

- Comparing performance with and without OCR text
- Evaluating different embedding models (mBERT vs. XLM-R vs. LaBSE)
- Measuring the impact of knowledge distillation
- Assessing the value of translation augmentation

3.3 Datasets and Resources

Our experiments will leverage:

- **Primary Dataset**: The MultiClaim dataset provided by the SemEval-2025 Task 7 organizers
- **Additional Resources**:
  - CLEF CheckThat! Lab datasets for transfer learning
  - MultiLingual Misinformation Dataset (MLMD)
  - Fact-checking websites like Snopes, PolitiFact, and international equivalents
  - Hugging Face's Transformers library for model implementation
  - PyTorch for deep learning framework
  - FAISS for efficient similarity search

3.4 Experimental Setup

We'll implement a rigorous cross-validation approach:

- Split the training data into 5 folds while preserving language distribution
- Use 4 folds for training and 1 for validation
- Perform hyperparameter tuning using Bayesian optimization
- Evaluate on the development set provided by the task organizers

We plan to conduct language-specific analyses to identify potential strengths and weaknesses in our approach across different language pairs.

4 Schedule

Our proposed timeline for the project is as follows, aligned with the course deadlines:

Date	**Course Milestone**	**Project Activities**
May 2, 2025	Proposal submission deadline	Submit this proposal document
May 3-8		Data exploration and preprocessing Initialize project repository and documentation
May 9, 2025	Proposal feedback released	Review feedback and adjust plans accordingly
May 10-16		Implement baseline models (BM25, TF-IDF) Begin data preprocessing pipeline
May 17-23		Implement multilingual embedding models Train initial dual-encoder architecture
May 24-30		Develop contrastive learning approach Add knowledge distillation components
May 31-June 4		System integration and testing Error analysis and model refinement Prepare presentation slides and finalize report
June 6, 2025	Final presentation & submission	Deliver oral presentation Submit final report and code

This schedule allows for iterative development with regular evaluation points to ensure we're making progress toward our goals. We've allocated additional time for error analysis and refinement to address any unexpected challenges that may arise during development.

5 References

1. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of EMNLP-IJCNLP 2019.

2. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of ACL 2020.

3. Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2022). Language-agnostic BERT Sentence Embedding. In Proceedings of ACL 2022.

4. Nakov, P., Barrón-Cedeño, A., Elsayed, T., Suwaileh, R., Haouari, F., Babulkov, N., Markov, I., & Da San Martino, G. (2021). The CLEF-2021 CheckThat! Lab on Detecting Check-Worthy Claims, Previously Fact-Checked Claims, and Fake News. In Proceedings of ECIR 2021.

5. Shaar, S., Babulkov, N., Da San Martino, G., & Nakov, P. (2020). That is a Known Lie: Detecting Previously Fact-Checked Claims. In Proceedings of ACL 2020.