CS7.501: Advanced Natural Language Processing
Interim Submission

Team - Ctrl+Alt+ElMo

Course Instructor - Dr. Manish Shrivastava

IIIT Hyderabad - October 7, 2024

1 Multilingual and Crosslingual Fact-Checked Claim Retrieval

1.1 Description

Multilingual and Crosslingual Fact-Checked Claim Retrieval focuses on developing a systems to retrieve relevant fact-checked claims for given social media posts across multiple languages, addressing the challenge of efficiently identifying previously fact-checked claims in a multilingual and cross-lingual context – supporting fact-checkers and researchers in their efforts to curb the global spread of misinformation.

1.2 Approach to the problem

We will be using something similar to retrieval part of RAG. We will implement retrieval method for the model to filter out the relevant documents (in this case posts). After, filtering we will use retrieval methods like BM25 etc. to get rank the most relevant fact checks. The final stage will be tuning hyperparameters of the retrieval and ranking model, choosing the appropriate Text embeddings (both multi-lingual and cross-lingual) and comparing their performances.

Model to Implement:

- BM25
- FastTTrack

And will be using different language embeddings like :

- GTR-T5-Large
- MiniLM-L12
- Sentence-T5-Large
- MiniLM-L2-Multilingual 118 En 0.74 0.38 0.16
- XLM-R

2 Progress So far

2.1 Dataset Cleaning

We utilize a comprehensive dataset consisting of several files that encompass fact-checks and posts in multiple languages, facilitating the training and evaluation of our multilingual and crosslingual claim retrieval system.

2.1.1 Fact-Checks

The dataset includes a file named `fact_checks.csv`, which contains a subset of 153,743 fact-checks across eight languages: Arabic (`ara`), German (`deu`), English (`eng`), French (`fra`), Malay (`msa`), Portuguese (`por`), Spanish (`spa`), and Thai (`tha`). This file serves as a key resource for understanding the various claims evaluated in different languages.

2.1.2 Posts

The `posts.csv` file encompasses all monolingual and crosslingual training and development posts. There is no overlap between the two subsets. This file supports the retrieval system by providing posts in 14 languages, enhancing the diversity of the training data.

2.1.3 Pairs

The `pairs.csv` file contains all training pairs, both monolingual and crosslingual. This resource is crucial for establishing connections between posts and their corresponding fact-checks, facilitating effective training for the retrieval model.

2.1.4 Steps of dataset cleaning

- We have done some pre-processing in posts like, removing emojis, extra punctuation or symbols.
- We have also cleaned the OCR transcripts for better readability.
- We have segregated the cross-lingual and mono-lingual(i.e., Translations of posts and fact-checkers in English).
- We have created json dump for faster loading as easy processing.

Conclusion: Mostly pre-processing is done, will make changes as we carry out the experiments.

2.2 Model Implementation

We have decided to focus on the implementation of the FastTrack Model, which is designed to handle multilingual and crosslingual retrieval tasks efficiently. FastTrack optimizes the retrieval process by utilizing a combination of dense and sparse retrieval techniques, allowing it to filter and rank relevant documents (in this case, social media posts) with high accuracy.

Our approach involves leveraging multilingual embeddings such as GTR-T5-Large, MiniLM-L12, and XLM-R, as they provide robust representations of text across different languages. For the retrieval phase, FastTrack uses these embeddings to map posts and fact-checked claims into a shared semantic space, ensuring that both monolingual and crosslingual matches are captured effectively.

The implementation is being carried out in a modular fashion, with simultaneous work on both monolingual and crosslingual aspects to ensure smooth transitions between the two. As of now, the model implementation is progressing steadily, and testing will commence soon to validate the retrieval accuracy.

3 Future Work

While **FastTrack** is the primary focus of our current implementation, we will also implement and evaluate the performance of **BM25**. BM25 is a widely used, effective, and interpretable sparse retrieval method, which ranks documents based on their term frequency and inverse document frequency.

In future experiments, we will:

- Compare the performance of **FastTrack** and **BM25** on the given dataset.
- Evaluate their respective retrieval capabilities using metrics such as **Mean Reciprocal Rank (MRR)** and **Success@K**.
- Analyze the impact of different language embeddings on both models' retrieval performance in monolingual and crosslingual contexts.

This comparative study will offer valuable insights into the trade-offs between dense (**FastTrack**) and sparse (**BM25**) retrieval methods, helping us determine which approach better serves the goal of accurate fact-check retrieval across multiple languages.

Note: The Dataset is attached to the link in the readme.md