Fact-Checked Claim Retrieval

Suyash Sethia 2021114010
Jainit Bafna 2021114003

Retrieval Wetrieval

Abstract

The task involves retrieving relevant fact-checked claims for given social media posts in a multilingual and cross-lingual setup using the MultiClaim dataset. Various approaches were tested and evaluated, with significant potential to mitigate misinformation effectively. This document provides an overview of the methodology, dataset, experiments, results, and the dual-encoder pipeline in detail.

Introduction

The retrieval of fact-checked claims is crucial in combating misinformation across multiple languages. The MultiClaim dataset comprises over 200,000 fact-checked claims and 28,000 social media posts in 27 languages. This project focuses on implementing and evaluating models for retrieving fact-checked claims, starting with English monolingual setups and extending to multilingual setups.

Literature Survey and related works

- **Exploring Dual Encoder Architectures for Question Answering**(Link): This paper explores dual encoders models ability to retrieve correct answers . the dataset is used is MS MARCO, open domain NQ and the MultiReQA benchmarks
- **Retrieval-Enhanced Dual Encoder Training for Product Matching**(Link): The Product matching is the task of finding the same prod-

uct in a catalog for a specific query product. Dual encoders had been proposed as a state-of-the-art solution

- **ColBERT: efficient and effective Passage Search via Contextualized Late Interaction over BERT**(Link): we present ColBERT, a novel ranking model that adapts deep LMs (in particular, BERT) for efficient retrieval. ColBERT introduces a late interaction architecture that independently encodes the query and the document using BERT and then employs a cheap yet powerful interaction step that models their one-grained similarity

Dataset Description

The MultiClaim dataset consists of three main parts:

- **Fact-Checks:** Contains fact-check claims with fields like ID, claim, instances, and titles.
- **Posts:** Contains social media posts with fields like ID, instances, OCR data, verdicts, and text.
- **Fact-Check-Post Mapping:** Maps fact-checks to social media posts, along with language pair information.

Experimental Details

- **Data Splits:** We Divide the data into 80:20 split for train and validation set .we shuffle the data to avoid any biases.
- **Hardware Setup** : We had access to ADA with 4 GPU'S

Preprocessing

The preprocessing involved:

- Using English translations for all multilingual data.
- Applying regex functions to clean text by removing links and handling elongations.
- Lemmatization with Porter Stemmer for BM25-based word-matching algorithms.

Approaches and Results

BM25

BM25 was used for initial baselining: Hyperparameters used were :

- K: 1.25
- B: 0.75

Metric	Value
Success@10	0.5734
Success@50	0.6898
Success@100	0.7312

Table 1: BM25 Results

Embedding Models

- Sentence encoders like *all-MiniLM-L6-v2*, *paraphrase-MiniLM-L12-v2* *sentence-t5-base* and *all-mpnet-base-v2* were evaluated.
- A FAISS vector database was used to store claim embeddings and retrieve top candidates.
- Both posts and claims were encoded using the same model and the similarity scores were calculated using cosine similarity.

Model	Success@10	Success@50	Success@100
all-MiniLM-L6-v2	0.6523	0.7423	0.7712
all-mpnet-base-v2	0.6534	0.7411	0.7878
sentence-t5-base	0.6358	0.7311	0.7690
paraphrase-MiniLM-L12-v2	0.6454	0.7232	0.7621

Table 2: Embedding Model Results

ColBERT Pre-Trained

Pre-trained Colbert model was used from hugging-faceLink.

- The colbert model uses seperate encoders for query and documents.
- Pre-trained Colbert model trained on task of ranking documents based on query
- The ColBERT model achieved marginally better results than BM25 but not better than embeddings model because, although colbert was trained on the similar task but the dataset was completely different.

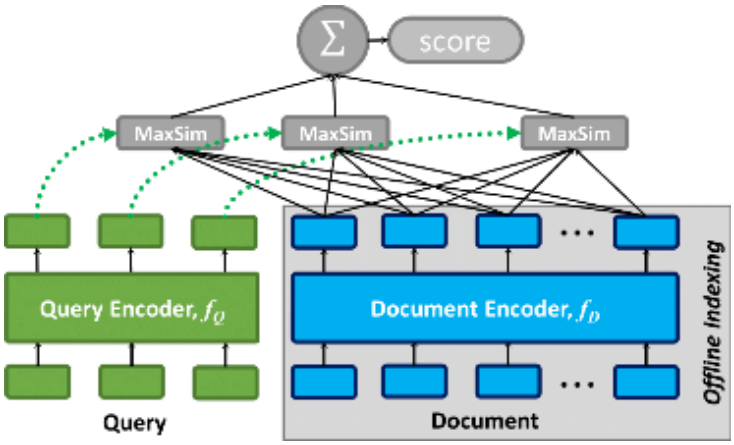

Figure 1: ColBert Architecture

Model	Success@10	Success@50	Success@100
ColBERT Pre-Trained	0.6124	0.7019	0.7399

Table 3: ColBERT Results

Dual-Encoder Architecture

The dual-encoder architecture outperformed other approaches:

Mathematical Description of Dual-Encoder Pipeline

Hyperparameters used were :

Model Name	Loss function	Success@10	Success@50	Success@100
all-mpnet-base-v2	(Cosine Loss)	0.6952	0.7719	0.8108
all-mpnet-base-v2	(Hyperbolic Loss)	0.7152	0.7819	0.8308
all-MiniLM-L6-v2	(Hyperbolic Loss)	0.7012	0.7741	0.8211
all-MiniLM-L6-v2	(Cosine Loss)	0.6812	0.7361	0.7544

Table 4: Dual-Encoder Results with Model Names

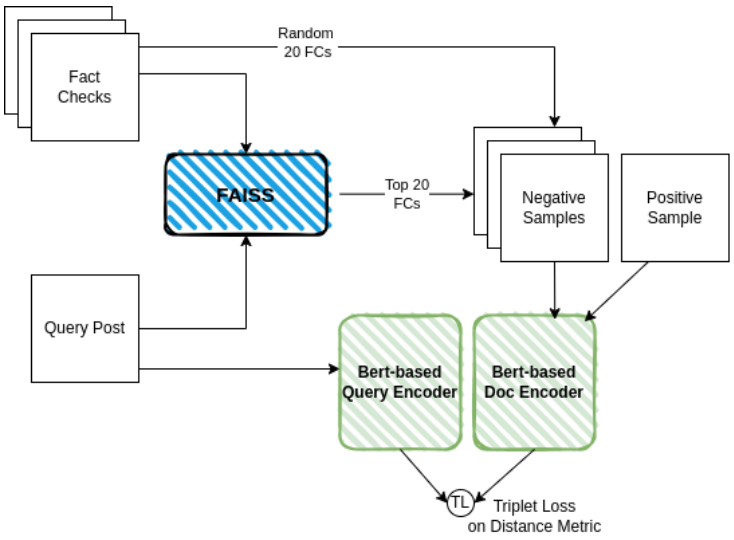

Figure 2: Dual-Encoder Architecture for Fact-Checked Claim Retrieval

- K: 1.25
- B: 0.75

The dual-encoder methodology consists of two components: a **Query Encoder** f_q and a **Document Encoder** f_d, both BERT-based models.

1. Triplet Definition

A triplet $T(Q, P, N)$ is defined as:

- Q: Query post embedding (from social media text).
- P: Positive sample embedding (fact-checked claim that matches the query).
- N: Negative sample embeddings (non-matching fact-checked claims).

2. Embedding Representations

- Query embedding: $\mathbf{q} = f_q(Q)$, where f_q maps the query text to an embedding $\mathbf{q} \in \mathbb{R}^d$.
- Positive sample embedding: $\mathbf{p} = f_d(P)$, where f_d maps the fact-check text to an embedding $\mathbf{p} \in \mathbb{R}^d$.
- Negative sample embedding: $\mathbf{n}_i = f_d(N_i)$, where N_i is a negative fact-check sample.

3. Similarity Metric

The cosine similarity between embeddings is defined as:

$$S(x, y) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}, \quad \text{for } \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$$

The hyperbolic distance between embeddings, is given by:

$$d_{\mathcal{H}}(x, y) = \operatorname{arcosh}\left(1 + 2\frac{\|\mathbf{x} - \mathbf{y}\|^2}{(1 - \|\mathbf{x}\|^2)(1 - \|\mathbf{y}\|^2)}\right), \quad \text{for } \|\mathbf{x}\|, \|\mathbf{y}\| < 1$$

where $\|\mathbf{x}\|$ denotes the Euclidean norm of $\mathbf{x}$, and arcosh is the inverse hyperbolic cosine function.

4. Triplet Loss

The training objective is to minimize the triplet loss:

$$\mathcal{L}_{\text{triplet}} = \max\left(0, S(\mathbf{q}, \mathbf{n}_i) - S(\mathbf{q}, \mathbf{p}) + \alpha\right)$$

where α is a margin parameter ensuring a sufficient gap between positive and negative similarities.

5. FAISS-Based Negative Sampling

- Top k candidates are retrieved from the FAISS database.
- Random N negatives are added for diversity.

6. Training Objective

The overall loss is minimized over all triplets:

$$\mathcal{L} = \frac{1}{|T|} \sum_{T(Q,P,N)} \mathcal{L}_{\text{triplet}}$$

7. Inference

During inference:

- Compute query embedding $\mathbf{q} = f_q(Q)$.
- Compare $\mathbf{q}$ with all fact-check embeddings $\mathbf{p}_i$ using cosine similarity.
- Retrieve the top k fact-checks with highest similarity.

Conclusion

Among all the models, the dual-encoder architecture consistently outperformed baseline methods and other advanced pre-trained models across all evaluation metrics. These results emphasize the dual encoder's superior capability in identifying relevant fact-check claims, making it an invaluable framework for addressing challenges in misinformation detection.

The dual-encoder architecture, optimized with triplet loss and FAISS-based sampling, demonstrated state-of-the-art performance in the task of fact-checked claim retrieval. This approach highlights its potential as a robust tool for combating misinformation through efficient and accurate retrieval of fact-checked claims.

The effectiveness of the tested models is summarized in the table below, which compares their performance on the Success@10, Success@50, and Success@100 metrics:

Model	**Success@10**	**Success@50**	**Success@100**
BM25	0.5735	0.6899	0.7312
all-MiniLM-L6-v2	0.6523	0.7423	0.7712
all-mpnet-base-v2	0.6534	0.7411	0.7878
sentence-t5-base	0.6358	0.7311	0.7690
paraphrase-MiniLM-L12-v2	0.6434	0.7232	0.7621
ColBERT Pre-trained	0.6124	0.7019	0.7399
Dual-Encoder	**0.7012**	**0.7791**	**0.8202**

Table 5: Performance comparison of retrieval models based on Success@k metrics.

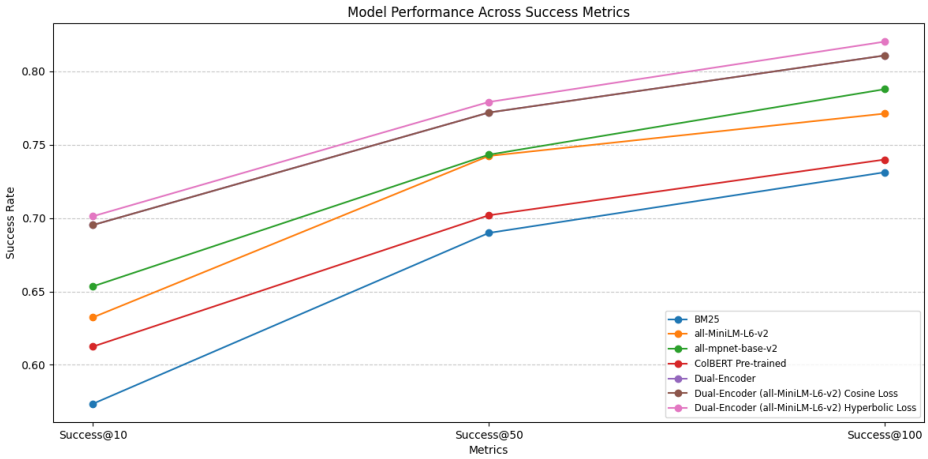

Figure 3: Results

Error Analysis

- The dataset contains claims which are very lexically and semantically very similar to each other . these examples create ambiguity and some claims have 100's of similar claims .
- – images of red sky in china are real
  – the sky of china is red in a strange phenomenon
  – picture of the sky in china turning red.
  – the sky turned blood red in china

  Only success@k was used as metric due the following ambiguity in dataset .

Ablation Study

- We tried different pre-processing steps to see the results including lemmatisation , stemming . but the results were quite low for embeddings based models .
- We tried many different sentence encoders for retrieval of negative samples for dual encoder model .
- We tried ensemble models to get to get better latency . but since the models were saved in vector database ensemble models reduced success@k scores and did not have large impact on latency

Scalability

- We have saved the models on one-drive and built a gradio-app which can be used to retrieve the top 10 documents given a query .

Future Work

- Extension of this project to multilingual and cross lingual setup .