Enhancing Fact-Verification: Leveraging Transformer Models for Evidence Retrieval and Classification

Alberto García Martín

University of Salamanca

Introduction

The rise of misinformation online necessitates robust fact-checking systems. The **FEVER[1] shared task** addresses this by providing a vast dataset of 185,445 claims from a 2017 Wikipedia dump, aiming to **automate the verification process**. This dataset includes evidence sets categorizing claims as Supported, Refuted, or NotEnoughInfo.

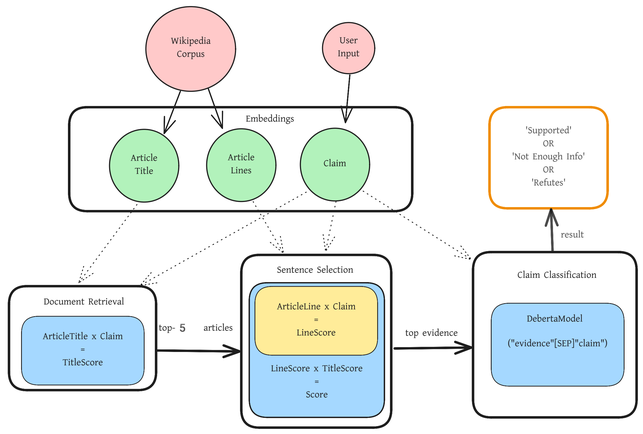

Figure 1. Diagram showcasing the proposed pipeline.

Techniques like entity linking and the Enhanced Sequential Inference Model (ESIM) have been developed for document retrieval and sentence selection, respectively, to rank evidence sentences. The final classification of claims leverages a textual entailment model to determine their veracity. Our paper builds upon these methodologies, **refining the pipeline for enhanced efficiency** and establishing a strong baseline for future research.

Related Works

The FEVER shared task has sparked diverse strategies for automating fact-checking.

- Initial approaches used standard retrieval methods like **TF-IDF**, enhanced by Hanselowski et al. with **NER** for better Wikipedia page retrieval.
- Soleimani et al. and Jiang et al. advanced the field with **transformer-based models**, larger model sizes, and sentence concatenation techniques.
- DeHaven et al.'s **BEVERS** system achieved **SOTA** performance by fine-tuning standard pipeline components, highlighting the power of optimization over novelty.
- BEVERS' document retrieval used **fuzzy string search** prioritizing document titles, and "point-wise" ranking for sentence selection, augmented by "evidence-based re-retrieval".

Current research aims to create robust, efficient, and accurate fact-checking systems to combat the spread of misinformation online.

Document Retrieval

In our FEVER pipeline, we've optimized document retrieval using a **leading embedding model** from the Massive Text Embedding Benchmark (MTEB). This model excels across various linguistic tasks, underpinning our method to identify and rank relevant Wikipedia articles effectively.

- Utilized **"bge-small-en"** model from the C-Pack suite for state-of-the-art document embedding.
- Performed **top-k similarity search** to retrieve relevant Wikipedia articles.
- Aimed to **surpass the baseline** FEVER TF-IDF retrieval model, with a focus on computational efficiency using vector databases that implement methods like Faiss.

Sentence Selection

This process involves **embedding each sentence** from the retrieved articles using the same sophisticated model applied in the document retrieval phase.

We enhance accuracy and context-awareness by **weighing the similarity scores** of both the article titles and the content of each sentence. The title, encapsulating the article's essence, and the detailed content analysis together enable us to identify the sentences most pertinent to the claim.

For evaluation purposes, we select the top five sentences from the weighted scores. In practice, we **focus on the top sentence** to streamline efficiency.

Claim Classification

Here we discern the veracity of claims by categorizing them as 'SUPPORTS', 'REFUTES', or 'NOT ENOUGH INFO'. We employ the **DeBERTaV3 model** as a base model, fine-tuned on the FEVER dataset.

Claim: Roman Atwood is a content creator. (**Supported**) **Evidence:** `[wiki/Roman_Atwood]` He is best known for his vlogs, where he posts updates about his life on a daily basis.
Claim: Furia is adapted from a short story by Anna Politkovskaya. (**Refuted**) **Evidence:** `[wiki/Furia_(film)]` Furia is a 1999 French romantic drama film directed by Alexandre Aja, ..., adapted from the science fiction short story Graffiti by Julio Cortázar.
Claim: Afghanistan is the source of the Kushan dynasty. (**NotEnoughInfo**)

Figure 2. Three examples from the FEVER dataset.

Before training we **refine the training data**, ensuring that each claim is paired with the most relevant evidence.

Training entails feeding the model with **claim-evidence pairs**, differentiated by a unique separator token, guiding the model to predict one of the three veracity labels. This process enables the model to evaluate the evidence in context to the claim.

In validation, we measure the model's **accuracy**, confirming its real-world applicability and generalization. Through DeBERTa-v3, we have not only improved upon the baseline but also have set a **new bar for accuracy** in claim classification within the FEVER challenge.

Experiments

Our pipeline for the FEVER task integrates a **Qdrant vector database** for document and sentence retrieval, improving speed and performance despite the computational intensity of embedding the entire Wikipedia dataset. We addressed the challenges of large-scale data processing and memory constraints by **offloading embeddings to disk** storage.

We employed the "deberta-v3-large" model, chosen for its **balance between parameter size and expected performance**. Training the claim classification model, we faced limitations in computational resources, allowing only for two epochs of training.

The unbalanced training dataset posed a risk of bias, which we countered by ensuring **balanced validation and test sets**. Our groundwork establishes a robust fact-checking system with room for further refinement.

Results

Our trained claim classification model achieved notable accuracy on the development dataset, surpassing the baseline FEVER score and demonstrating **competitive performance alongside state-of-the-art models**. Accuracy improved from 87.79% to 88.64% across two epochs, underscoring significant potential beyond the initial baseline of 50.91%.

Training Loss	Epoch	Step	Validation Loss	Accuracy
0.2748	1.0	7134	0.3707	87.79%
0.1482	2.0	14268	0.3771	88.64%

Table 1. Results of the trained claim classification model on the development dataset.

Limited computational resources restricted our evaluation of the full pipeline to a small subset (∼5%) of the test dataset. Thus, our results are **preliminary**.

FEVER Score	Evidence F1	Label Accuracy
0.73	0.41	76%

Table 2. Preliminary results of the overall pipeline on a subset of the test dataset.

The evidence retrieval process emerged as a **bottleneck**, highlighting an area ripe for enhancement. Despite this, our FEVER score aligns with average submissions, indicating our pipeline's robustness.

Conclusion

Further development could refine the evidence retrieval phase and capitalize on the full potential of our model. Our current results, despite limitations, suggest that our model's quality could significantly **influence the advancement of the field of automated fact verification**.

References

[1] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: A Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819. Association for Computational Linguistics.

albertogm@usal.es