# A BERT-based system for Climate Evidence Retrieval and Claim Verification

**Yinghua Zhou**

## Abstract

This paper explores a BERT-based system for fact-checking unverified climate-related claims. We study the significance of TF-IDF cosine similarity in evidence retrieval and examine the impact of Hard Negative Mining during training. We also train a separate BERT model to classify claims, testing three different approaches for final label determination. Our system uncovers several insights and displays a somehow relatively competitive performance in the CodaLab-hosted competition.

## 1 Introduction

With the rapid dissemination of online climate information, automatic fact checking systems are vital to counter unverified or misleading claims.

In this paper, we focus on retrieving the relevant evidences from a given evidence pool for a particular claim, and then use the evidences retrieved in the previous step along with the claim itself to verify the truthfulness of the claim, by classifying the claim as "SUPPORTS", "REFUTES", "NOT_ENOUGH_INFO" or "DISPUTED".

Evidence retrieval, can be seen as a type of Natural Language Inference task, which involves determining the authenticity of a hypothesis (evidence) given a premise (claim). We investigate BERT on this task due to its promising achievements in the relevant research work. We then experiment on a specific Hard Negative Mining strategy to observe its effectiveness.

For Claim label classification, we examine on a separate BERT model, and compare three different strategies, namely rule-based aggregation, majority voting and evidence concatenation for BERT.

Given the evidence pool is significantly large, training BERT on every evidence for each claim is infeasible. We address this by investigating the Term Frequency Inverse Document Frequency (TF-IDF) cosine similarity between each claim and all evidences, as a way to filter out the ones that are less likely to be relevant given a claim, in order to reduce computational demands and focus on relatively promising evidences.

## 2 Related Work

A wealth of research has been conducted based on the FEVER dataset introduced in the work of Thorne et al. (2018), which is designed for the Fact Extraction and Verification challenge. These studies typically adopt a pipeline approach that encompasses document retrieval, sentence retrieval and claim verification. Our evidence retrieval task shares similarities with the first two components. There have been various models proposed for each pipeline component, while BERT has been demonstrated of its competitiveness and is widely adopted (Bekoulis et al., 2021), therefore we are interested in using BERT-based models for this study. As for claim label classification, some previous studies such as those by Soleimani et al. (2019) and Malon (2019) have focused on predicting individual claim-evidence pairs, followed by aggregation using rule-based approaches. Alternatively, some studies have directly employed neural methods in some ways on multiple potential evidences (Bekoulis et al., 2021). In our research, we will examine both of these strategies, alongside majority voting based on predictions for individual pairs.

## 3 Datasets

The datasets provided for this study consist of 1228 train claims, 154 development claims, and a set of 1208827 evidences. Drawing from the document retriever of the DrQA system (Chen et al., 2017), we explore TF-IDF cosine similarity in narrowing the potential relevant evidences. We observed that majority (over 90%) of ground truth evidence occurrences, each representing a true evidence for a specific claim, exhibit TF-IDF cosine similarity with their associated claim. Notably, 34.6% of these are among the top 100 in terms of TF-IDF

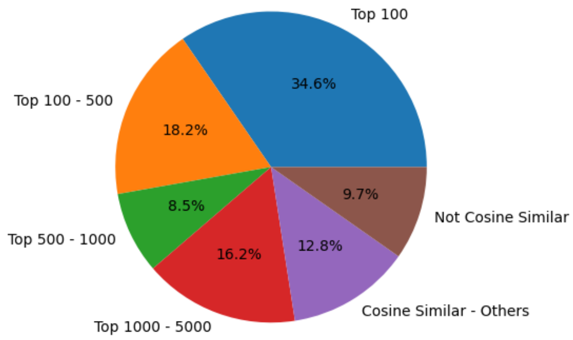

Figure 1: Proportion of Relevant Evidences in Each TF-IDF Cosine Similarity Range.

cosine similarity relative to their associated claims, meaning we can retrieve up to 34.6% of ground truth evidences if we consider only the top 100 TF-IDF cosine similar evidences for each claim. This proportion increases to over 60% when considering the top 1,000 TF-IDF cosine similar evidences, as shown in Figure 1. This finding underscores the potential effectiveness of pre-filtering the substantial volume of evidences for each claim, given that it is super computationally expensive and potentially ineffective to process all evidences for every claim using BERT.

## 4 Methods

The model we employ in this study, BERT, which stands for Bidirectional Encoder Representations from Transformers (Devlin et al., 2019), is a multi-layer bidirectional model that leverages the multi-head self-attention mechanism of the Transformer's encoder architecture (Vaswani et al., 2017). It is pre-trained on the objectives of Masked Language Modeling and Next Sentence Predictions, allowing effective and efficient adaptation to downstream tasks through a fine tuning process.

The input to the BERT model is a sequence of tokens, each will be mapped to an embedding vector. Sentence pairs are packed into a single sequence, padded with a [CLS] token at the beginning, and a [SEP] token between sentence pairs and at the end:

```
[CLS] Claim [SEP] Evidence [SEP]
```

Alongside we also incorporate segment and position embeddings to denote the sentence belonging and position information of a token respectively. The [CLS] token is used as a representation of the entire sequence, encapsulating the relationship between a claim and evidence when packed together. It is normalized via an additional linear layer for the final output. We utilise the BERT-base version throughout this study.

### 4.1 Evidence Retrieval

To begin with, in training the BERT model for evidence retrieval, we randomly sample negative evidences from the entire evidence pool for each claim, while including all ground truth evidences. Each claim-evidence pair is treated as a single training sample, and the task is treated as a binary classification problem of determining evidence relevance to the claim. Given the small proportion of selected evidences, we also discover the impact of dynamically refreshing negative evidences each epoch. We then examine the approach that samples negative evidences randomly from a specified top TF-IDF cosine similarity range. Given the inherent randomness in this process, we compare each approach's best performance across 5 experiments over 10 epochs, to roughly picture their effectiveness.

In validation or testing, we pre-select the first 1000 TF-IDF cosine similar evidences for each claim, the number is chosen based on the development set performance. Each claim-evidence pair is passed through the fine-tuned BERT model and ranked by their sigmoid layer probability outcomes. The top five evidences at maximum are selected as the predicted relevant ones for a claim, mirroring the fact of there are max of five evidences per claim in the datasets. In addition, we attempt to apply a threshold to balance recall and precision.

We employ Binary-Cross Entropy as the loss function for this task, and train the model for one epoch on selected samples with a learning rate of 2e-5 and a batch size of 64. We set a confidence threshold of 0.9, whereby the model selects only evidences with a confidence of over 90%. For the approach of randomly sampling negative evidences from the top TF-IDF cosine similar ones for training, we set the range to top 5000 for all claims, while maintaining a 1:5 positive-to-negative evidence ratio for a claim and balance classes by assigning class weights based on this ratio. These hyper-parameters are tuned based on development set performance.

Inspired by Soleimani et al. (2019), we will explore the effectiveness of a specific offline Hard Negative Mining strategy. Post initial training, we test the model on all training claims to gather as many false positives as true positives per claim. We then retrain the model over multiple epochs using this refined dataset, with some random samples for generalization, to observe the development set performance changes. We reduce the learning rate to 2e-7 to manage overfitting.

### 4.2 Claim Classification

We train a distinct BERT model for this task, initially considering only ground truth evidences and their associated claims as training samples. We explore three approaches under two broad categories for this task. We will also compare the best performance across 5 experiments each over 10 epochs.

The first category focuses on training each individual claim-evidence pair, labeled the same as the associated claim, while excluding claims labeled as 'DISPUTED' for training due to the inherent inability to distinguish each individual evidence. we experiment with both majority voting and rule-based aggregation to determine a claim's final label. Majority voting picks the most common label among a claim's evidences with random tie-breaking. Rule-based aggregation follows these guidelines: a claim is labeled as 'SUPPORTS' or 'REFUTES' if one exists without the other; 'NOT_ENOUGH_INFO' if there are neither 'SUPPORTS' nor 'REFUTES'; and 'DISPUTED' if there's at least one 'SUPPORTS' and one 'REFUTES'.

The second category of approach aggregates the evidences of a claim before inputting them into the model. We explore concatenation of evidences into a single text sequence, after random shuffling the evidence set for a claim to mitigate the problem of the evidences at the end being truncated every time due to BERT's input sequence length limit.

It is important to recognise the performance of this claim label classification task heavily depends on the preceding evidence retrieval task. Training and validating on ground truth evidences may fail to reflect practical performance due to discrepancies between the trained data and actual test data, since the test evidences are derived from the previous task and may contain noise, potentially confusing the model. Thus, we aim to explore the impact of incorporating predicted evidences, obtained from the evidence retrieval model of the same pipeline, into the training data for this task.

| **Model** | **F1-Score** | **Recall** | **Precision** |
|---|---|---|---|
| {A} | 0.144 | 0.191 | 0.116 |
| {B} | 0.193 | 0.215 | 0.175 |
| {C} | 0.229 | 0.301 | 0.185 |
| {D} | 0.218 | 0.252 | 0.192 |

Table 1: Best performance comparison of the proposed approaches on the dev set over 10 epochs in 5 experiments: 'A' denotes fixed random sampling from the entire evidence pool; 'B' is its dynamically refreshed variant; 'C' represents sampling from the top TF-IDF range; 'D' refers to 'B' further trained with Hard Negative Mining.

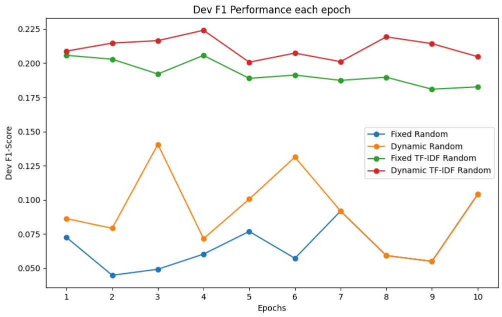

Figure 2: F1 Performance of each Evidence Retrieval approach on the development set.

## 5 Results

Figure 2 presents the performance of our evidence retrieval methods in a single experiment on the development set. Due to the inherent randomness in our sampling strategies, the results vary, especially when negative evidences are dynamically resampled each epoch. However patterns are similar in general as observed. The approach with a fixed set of negative evidences randomly sampled from the entire evidence pool under-performs, which we attribute to the limited variety of evidences seen during training, and the relative triviality of the majority of these samples.

This statement can be further supported when we compare the approach to the dynamically refreshed variant. The best F1-score performance largely improves under the dynamic method, as shown in Table 1, suggesting that exposure to a wider array of negative evidences aids in generalization and potentially exposes the model to more informative samples.

The best performing approach considers sampling negative evidences from the top TF-IDF cosine similarity range. This approach not only highlights a certain degree of correlation between TF-IDF cosine similarity and the relevance of a claim-evidence pair, but also emphasizes the importance of training on meaningful samples to enable the

| Model | Accuracy (T) | Accuracy (P) |
|---|---|---|
| {A} | 0.526 | 0.552 |
| {B} | 0.604 | 0.552 |
| {C} | 0.713 | 0.416 |
| {D} | 0.519 | 0.54 |
| {E} | 0.545 | 0.57 |
| {F} | 0.526 | 0.5 |

Table 2: Best performance comparison of the proposed approaches on the dev set across five experiments each over 10 epochs. Results are based on accuracy with both **T**rue and **P**redicted evidences from the preceding ER model. 'A' represents Rule-based Aggregation, 'B' Majority Voting, 'C' evidence concatenation for BERT, while 'D', 'E', and 'F' denote 'A', 'B', and 'C' respectively, trained with predicted evidences.

model to capture more sophisticated patterns. However, it is still not sufficiently capable of capturing deep semantics relationships between a claim and its evidences, while the performance gain from using the top TF-IDF range is modest, and that there may be room to further leverage this relationship for enhanced performance.

We observed a modest improvement in F1-score of 0.02 with the proposed Hard Negative Mining strategy based on the basic dynamic refreshing variant compared to its best performance. However, we were not able to experiment this strategy on the top TF-IDF cosine similarity range sampling approach due to computational constraints. This constraint primarily stems from the extended duration required to find hard negatives for the model as observed. However, this finding implies that the TF-IDF approach is potentially more robust towards general negative evidences.

As for label prediction, the rule-based aggregation method is in fact highly sensitive to the prediction of each individual claim-evidence pair, as a single incorrect prediction can mislead the final claim label. This is evident from its under-performance as shown in Table 2. Majority voting is more resilient, especially when the performance of the preceding evidence retrieval task is unreliable. However trade-offs apply, especially we have to inherently give up the 'DISPUTED' choice, and the performance is to a greater extent instability dominated.

The best performing approach when isolating the evidence retrieval component and focus solely on the ground truth evidences for claim classification, is the one that concatenates evidences for each claim. This technique leverages synergistic information, allowing the model to capture a deeper relationship between the claim and its evidences. However, it highly relies on the quality of the predicted evidences from the preceding task in practice, as any noise can largely distort the attributes of the entire concatenation. This dependency is illustrated by the substantial drop in accuracy observed when testing on predicted, rather than ground-truth, evidences for the development set.

Our system, as evaluated in the CodaLab competition, employed the top TF-IDF cosine similarity for evidence retrieval and majority voting for claim label classification with incorporation of predicted evidences in the training samples. We noticed a significant F1 performance decrease from 0.2308 to 0.1637, in evidence retrieval compared to our ongoing evaluations, a trend that also happened to other systems. This may be contributed by the potential overfitting or the model under-performs in a group of samples with certain specific patterns.

The claim label classification performance of our system also experienced a significant accuracy drop from 0.5 to 0.44160 compared to ongoing evaluations, although similar accuracy was observed in previous ongoing submissions. While the inherent instability of majority voting can contribute to this discrepancy, it also reflects that simply incorporating predicted evidences into classifier training may not be strictly sufficient for capturing task patterns and ensuring practical reliability. Nevertheless, the system demonstrated relatively competitive results in the competition, despite the system needs various improvements that could be explored in the future.

## 6 Conclusion

In this paper, we presented a BERT-based pipeline for evidence retrieval and claim label classification, and evaluated several alternatives within each component. Our findings indicate that utilising TF-IDF cosine similarity can effectively filters relevant samples from a large evidence pool. Furthermore, we found that Hard Negative Mining can potentially be helpful. It is also noted that the performance of evidence retrieval can significantly impact the downstream claim label classification task. Results have revealed many potential improvements on this pipeline, while a crucial aspect for future research will be to capture deeper semantic relationships between claims and evidences to further improve model performance, and that exploring an effective way to resolve or compensate the inaccuracy of evidence retrieval is essential for claim classification.

## References

Giannis Bekoulis, Christina Papagiannopoulou, and Nikos Deligiannis. 2021. A review on fact extraction and verification. *Computing Research Repository*, arXiv:2010.03001. Version 5.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *Computing Research Repository*, arXiv:1704.00051. Version 2.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Computing Research Repository*, arXiv:1810.04805. Version 2.

Christopher Malon. 2019. Team papelo: Transformer networks at fever. *Computing Research Repository*, arXiv:1901.02534. Version 1.

Amir Soleimani, Christof Monz, and Marcel Worring. 2019. Bert for evidence retrieval and claim verification. *Computing Research Repository*, arXiv:1910.02655. Version 1.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *Computing Research Repository*, arXiv:1803.05355. Version 3.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Computing Research Repository*, arXiv:1706.03762. Version 5.