A BERT-based Method for Fact Checking of Climate Science Claims

Student ID: 1118749

Abstract

With the increase importance of fact checking nowadays, this paper investigate the pipeline of evidence retrieval and claim classification for climate science claims. TF-IDF is implemented initially to filter out the majority of unrelated evidences and narrow down the search range. Two BERT models are trained, with one of them uses hard negative mining to identify unrelated evidence with high similarity. Selection of related evidences is based on a score obtained through model stacking, including TF-IDF, Jaccard, DistilBERT and fine-tuned BERT. For claim classification, a BERT model is fine-tuned to accept a pair of claim and evidence to predict a label. The label is finally determined by choosing the one with the most votes, based on retrieved evidences. The method achieved a F-score of 0.179 in evidence retrieval and an accuracy of 0.5 in claim classification on the development set. For the final test set, it has a F-score of 0.145 and an accuracy of 0.532.

1 Introduction

In the era of the internet, people have unprecedented access to tons of information every day. However, there is a huge volume of fake and misleading content, which potentially could result in the distortion of public opinions. Therefore, fact checking has become an important area, which is composed of two tasks: evidence retrieval and claim classification. This paper focuses on a provided dataset regarding climate science and puts forward a model to automatically verify the claims based on a provided evidence set.

Claim: South Australia has the most expensive electricity in the world.

Evidences:

South Australia has the highest retail price for electricity in the country.

South Australia has the highest power prices in the world.

Label: SUPPORTS

Figure 1: Example of fact checking

The dataset is divided into a training set and a development set. The training set consists of 1,228 claims, while the development set contains 154 claims. The collection of evidences contains 1,208,827 pieces of text. For each claim, a minimum of 1 and a maximum of 5 evidences are retrieved. Claims are categorised into one of four labels based on the retrieved evidences, including SUPPORTS, REFUTES, NOT_ENOUGH_INFO and DISPUTED. Figure 1 presents an example of fact checking. Figure 2 shows the distribution of labels in the training set, and the development set follows the similar distribution.

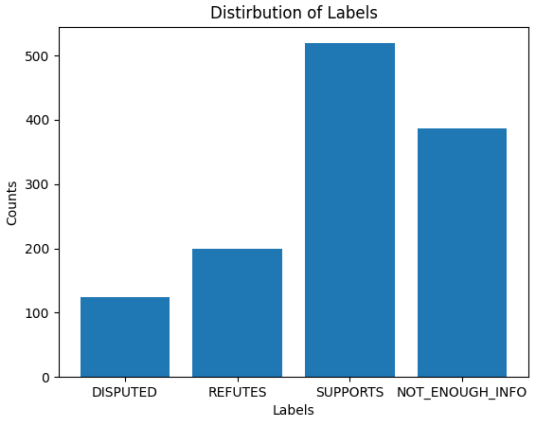

Figure 2: Distribution of labels

The proposed model combines muliple techniques, including Jaccard similariy, TF-IDF, and BERT, to retrieve related evidence for a given claim, which follows a two-step pipeline. In the first step, TF-IDF is applied to narrow down the search range and filter out less relevant evidences. The second step is to use model stacking to extract evidences with highest scores. Subsequently, a BERT-based model is employed to perform claim classification.

2 Related Works

Thorne et al. (2018) proposed a dataset specifically

for the task of Fact Extraction and VERification (FEVER), along with a method which achieved an accuracy of 31.87% in claim labelling and 50.91% in evidence retrieval. Within the dataset, a total of 185,445 claims are generated from Wikipedia, each of which can be classified into one of three labels: SUPPORTED, REFUTED or NOTENOUGHINFO (Thorne et al., 2018). Several scholars have contributed to the development of this task, and Bekoulis et al. (2021) conducted a thorough review on most of the recent methods proposed for this task.

Pre-trained models like BERT have been proved to perform well in a variety of NLP tasks. In the context of FEVER, a number of methods have been developed based on BERT to enhance performance. Nie et al. (2019) proposed a model combining TF-IDF, keyword matching and BERT, which obtained a F1-score of 74.62% in sentence retrieval and an accuracy of 72.56% in claim classification. Soleimani et al. (2020) conducted an experiment comparing the performance of various methods based on BERT, including pointwise, pairwise and hard negative mining. The results revealed that the highest F1-score achieved for evidence retrieval reached 69.66%, while the highest claim classification accuracy achieved was 71.86% (Soleimani et al., 2020). In addition to pre-trained models, Chen et al. (2022) attempted to decompose the claim to phrase-level to improve claim verification, and they achieved an accuracy of 76.91%.

3 Methods

3.1 Evidence Retrieval

3.1.1 Text Preprocessing

The provided textual data are not cleaned. Although BERT has its own tokenizer and does not require preprocessing as the original semantic structure is preferred to be maintained, alternative techniques such as Jaccard similarity and TF-IDF perform better on preprocessed data since noise and irrelevant information are eliminated. Therefore, a separate copy of the evidence dataset is created, containing cleaned texts. The process is completed by using NLTK and regular expressions, involving the removal of punctuations and stopwords, lowercase conversion and lemmatization.

3.1.2 Jaccard Similarity

Jaccard similarity calculates the size of the intersection divided by the size of the union of two sets, which returns a high score if two texts share many words. Assuming that a claim and its related evidences should describe the same topic, it is reasonable to expect they should have shared entities, for example, locations or organizations. Nonetheless, it cannot work with synonym sets, for example, high price and expensive. Named Entity Recognition (NER) tagger is also put into consideration as it is specifically designed to capture the entities. However, it is less efficient and its improvement in performance is limited.

3.1.3 TF-IDF

TF-IDF is a powerful technique for information retrieval as it evaluates the importance of words based on their occurrence and frequency, rather than the syntax and grammatical structure of the sentence. It is able to identify words that are highly representative of a given sentence, with high efficiency. To facilitate the comparison between a claim and the evidence dataset, a TF-IDF vectorizer is applied to the evidence dataset. The vectorizer transforms the textual data into TF-IDF vectors, representing the importance of each word in the evidence texts. When assessing the similarity between a claim and an evidence, the claim is first transformed into TF-IDF vectors and cosine similarity is calculated between two vectors.

3.1.4 DistilBERT

The above techniques does not take semantic features into account. To implement an efficient way of capturing such features, DistilBERT is selected, which is a compact version of BERT. All evidences are encoded and passed into the *distilbert-base-uncased model*. The CLS token from the model's output is extracted as the embeddings for evidences. When computing the similarity, an embedding is generated for the claim and cosine similarity is computed between the claim embedding and the evidence embedding. This approach manages to obtain semantic similarities efficiently.

3.1.5 Fine-tuned BERT

As shown in figure 3, a model is constructed based on BERT to classify if an evidence is related to a claim, with 0 representing non-related and 1 as related. Binary cross entropy is set as the loss function, using the Adam optimizer with learning rate equals 2e-5. A sigmoid function is applied to the model's output (CLS token) to constrain it within the range of 0 to 1. For the selection of pretrained model, *bert-base-uncased* is used since

there is no huge improvement if using large bert models (Soleimani et al., 2020).

Related evidences for each claim are extracted and paired together with the claim to form positive cases in the training data for this model. Two strategies are implemented to create negative cases. The first one is randomly sampling from the evidence set. While this approach is simple to apply, it may provide cases which are too easy for the model to discriminate from positive cases. Inspired by the hard negative mining (HNM) proposed by Soleimani et al. (2020), negative cases are created by extracting evidences which are the most similar to the claim using TF-IDF similarity. It is expected that training with HNM could make the model better at identifying similar but unrelated content, improving the overall robustness. To ensure the dataset is not biased, the sampling size is set to be equal to the positive cases.

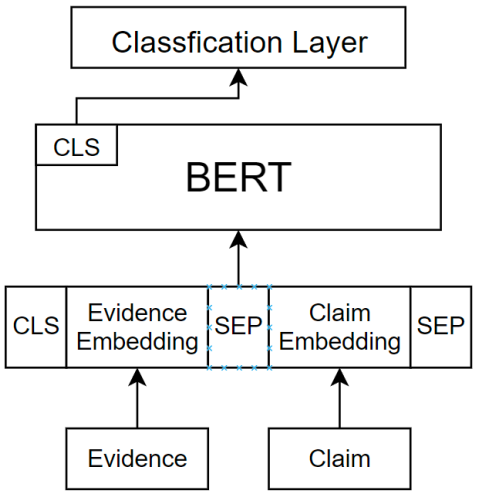

Figure 3: Structure of evidence classifier

3.2 Claim Classification

A claim classifier is modelled similarly as figure 3 depicts. Instead, a softmax function is applied to the model's output to return the probability distribution of labels. Based on the retrieved evidence, the claim is classified by selecting the label with the highest probability. For the training data, related evidences of each claim are extracted and paired together with the claim and the given label.

4 Modelling

4.1 Evidence Retrieval

The process of evidence retrieval follows a two-step pipeline, initialised with a filter to extract top 150 candidates that potentially related to the claim so as to narrow down the search range. Afterwards a score is calculated for each candidate, combining several similarity metrics. Finally, 5 evidences are retrieved as the related ones to the claim. Figure 4 presents the compete pipeline of this step.

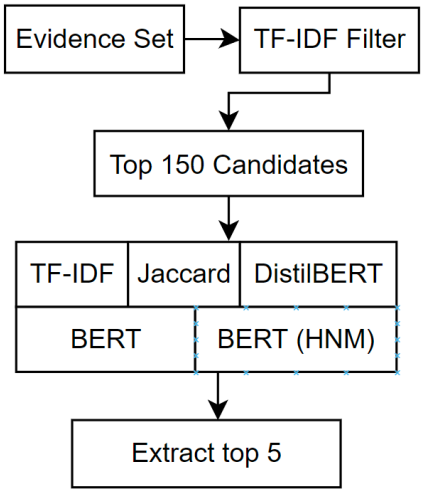

Figure 4: Pipeline of evidence retrieval

4.1.1 Initial Filter

As a initial filter, it is necessary to balance between efficiency and performance. Table 1 shows the number of correct evidences retrieved and the execution time. For each filter, it extracts the top 100 most similar evidences for all claims in the development dataset. Benefit by the pre-calculated vectors and keyword matching, TF-IDF dominates other methods in both efficiency and performance, therefore is selected as the initial filter.

Table 1: Performance of filters on dev set

Method	#Correct	Time(s)
Jaccard	174	570.3
TF-IDF	186	42.4
DistilBERT	115	402.7

4.1.2 Model Stacking

To improve the accuracy of evidence retrieval, model stacking is utilised to evaluate each candidate from multiple dimensions. Table 2 presents the allocated weights for each metric. The values are determined by evaluating performance on the development set. As candidates all have relatively high TF-IDF similarity, more weights is put on comparing semantic similarity. The BERT classifier trained on randomly sampled dataset is

applied to filter out obviously unrelated evidences. Meanwhile, another classifier trained using hard negative mining is expected to distinguish unrelated evidences which have high similarities.

Table 2: Model stacking

Method	Weight
TF-IDF	0.05
Jaccard	0.05
DistilBERT	0.2
BERT	0.4
BERT (HNM)	0.3

4.2 Claim Classification

In the process of claim classification, each retrieved evidence is paired with the claim and subsequently passed into the BERT classifier. The final classification result is determined through a voting mechanism, which selects the label with the most votes. Note that a claim will not be classified as NOT_ENOUGH_INFO unless it is the only label returned, so that other valuable evidences can gain higher priorities.

5 Evaluation

5.1 Results

Performance is evaluated by F-score of evidence retrieval, accuracy of claim classification and harmonic mean of F-score and accuracy. Table 3 presents the method's performance on both the development and test sets. For the test set, both the ongoing and final evaluations are presented. Note that models are trained on data combining both the training and development sets to maximize the performance on the test set.

Table 3: Performance on dev and test set

Dataset	Mean	F-score	Accuracy
Dev	0.263	0.179	0.5
Test (Ongoing)	0.26	0.174	0.513
Test (Final)	0.228	0.145	0.532

5.2 Critical Analysis

The F-score in evidence retrieval is lower than the performance reported in other papers regarding the FEVER dataset. The difference may stem from several factors, such as variations in the dataset. The volume of given dataset is limited comparing to the FEVER dataset. Besides, there are many claims labelling as NOT_ENOUGH_INFO, and the corresponding retrieved evidences may not have direct relation with the claim, which could lead the classifier to learn useless features. Furthermore, a notable challenge is the existence of numerous similar contexts that can be regarded as potentially related to a claim. While hard negative mining can provide some alleviation to the problem, the improvement is still limited. Finally, the method may be overfitting on the development set, which leads to a drop in the final evaluation of the test set.

Performance in claim classification is relatively better. The training set is biased since most of claims are labelled as SUPPORTS or NOT_ENOUGH_INFO as shown in figure 2, which may let the classifier prone to predicting these two labels. It is likely that label distribution in the test set is similar to the training set so that the prediction accuracies are close.

6 Conclusion

Several techniques are investigated for evidence retrieval, including TF-IDF, Jaccard, DistilBERT and fine-tuned BERT. The evidence is first filtered using TF-IDF, which has strong performance and efficiency comparing to other techniques. In addition, hard negative mining is implemented to enhance the model's ability of detecting unrelated evidences with high similarity. Evidences are finally retrieved based on a score obtained through model stacking. For claim classification, a fine-tuned BERT model is utilised to predict a label given a pair of claim and evidence. The final label is determined by selecting the label which has the most votes.

For the development set, the method achieved a F-score of 0.179 in evidence retrieval and an accuracy of 0.5 in claim classification. Meanwhile, it has a F-score of 0.145 and an accuracy of 0.532 on the final test set. Future research could place more emphasis on enhancing evidence retrieval, particularly in dealing with unrelated evidences that have high similarity.

References

Giannis Bekoulis, Christina Papagiannopoulou, and Nikos Deligiannis. 2021. A review on fact extraction and verification. *ACM Computing Surveys (CSUR)*, 55(1):1–35.

Jiangjie Chen, Qiaoben Bao, Changzhi Sun, Xinbo Zhang, Jiaze Chen, Hao Zhou, Yanghua Xiao, and

Lei Li. 2022. Loren: Logic-regularized reasoning for interpretable fact verification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10482–10491.

Yixin Nie, Songhe Wang, and Mohit Bansal. 2019. Revealing the importance of semantic retrieval for machine reading at scale. *arXiv preprint arXiv:1909.08041*.

Amir Soleimani, Christof Monz, and Marcel Worring. 2020. Bert for evidence retrieval and claim verification. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*, pages 359–366. Springer.

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