A Report on Climate Science Claim Fact Checking System

Junhua Liu
The University of Melbourne

Abstract

This paper introduces a claim verification model. The task can be divided into information retrieval and claim classification. In the information retrieval section, the BM25 ranking algorithm, NER matching, and a pre-trained Bert model are adopted to retrieve relevant evidence that matches a given claim. The claim verification utilizes the Bert model for claim classification. Despite the low retrieval rate in the BM25 ranking model and the NER matching model, two pre-trained Bert models present satisfactory performance, suggesting their potential utility in future tasks.

1 Introduction

Claim verification is a complicated problem of significant interest in the Natural Language Processing (NLP) field, with many methods explored to approach this problem. With the introduction of Attention (Vaswani et al., 2017), in the deep Neural Network field, an improvement in the prediction accuracy in some Natural Language Models can be observed. The ability to incorporate context representation is an important feature of Attention. This feature is also important for Claim Verification models. Research on adopting Transformer models in claim verification has been performed and satisfactory results have been found.

This report aims to build and evaluate the performance of the claim verification task with the help of pre-trained Bert models on a small training dataset. Firstly, an explanation of the structure will be given to introduce the methodology. Secondly, the model will be discussed in detail. After that, evaluation and discussion will base on the model performance will be presented. At last, a conclusion will be given with some advice on future work.

2 Methodology and Reasoning

The evidence dataset contains 1,208,827 records, which is not a small dataset. Therefore, information retrieval methods are adopted in this project. The training dataset, in contrast, is a small dataset in terms of model training, containing only 1,227 records.

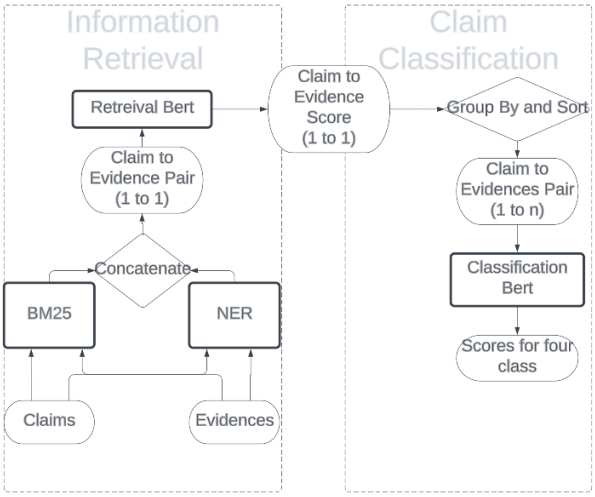

Figure 1: Claim verification pipeline.

The task aims to find relevant evidence to a claim in a database and then classify the claim into four labels, including SUPPORTS, REFUTES, NOT_ENOUGH_INFO, and DISPUTED, according to the retrieved information. Inspired by two Bert classification systems (Soleimani et al., 2020), a similar model design is used in this assignment. This task can be divided into two parts. The first part is information retrieval. In this process, evidence that correlated to the claim is extracted from the evidence database. In the second part, claim classification is performed using a pre-trained Bert Model.

2.1 Information Retrieval

A hybrid model is utilized for information retrieval. Considering the limited training dataset and the time complexity, a rule-based model using the BM25 ranking algorithm is used to reduce the selection range. This result is combined with evidence

retrieved by Name Entity Retrieval (NER) matching, as candidates. A Bert model is then adapted to classify this evidence into RELEVANT and IRRELEVANT groups.

The rule-based model with BM25 functions, explained by Tortman (Trotman et al., 2014), incorporates multiple factors, including term frequency, document length normalization, and inverse document frequency, to estimate the relevance between the given claim and each piece of evidence. 10 shreds of evidence with the highest score are chosen as candidates for each claim. This evidence is then transferred to the Bert model for fine selection. Another alternative to this method is TF-IDF. The latter method is not chosen by considering its high dependency on term frequency. It is worth mentioning that the given claims and evidence are short, which may result in underperformance using TF-IDF. BM25 incorporates term frequency normalization, which could be beneficial for short text ranking.

NER matching is used for information retrieval. Evidence may not directly prove the claim but may share some same-name entities. Therefore, finding less common name entities may offer help in retrieving information. Topic modelling methods are not considered because of their hardness in hyperparameter tuning and long processing times, which is essential when considering the number of evidence.

Pretrained Bert models are adopted to perform the final retrieval task by considering the high performance of transformer models in NLP tasks. The result is selected from candidates from both BM25 and NER models for further classification.

2.2 Claim Verification

The Bert model is used to make the decision of accepting relevant evidence. Transformer models generally produce accurate predictions on NLP tasks because of the addition of Attention. As mentioned before, this mechanism is beneficial to capturing context relevance and resolving ambiguity. Both of which may resulting high performance in both information retrieval and claim classification tasks. Considering the relatively small dataset, pretrained Bert models are selected instead of training a Transformer model from scratch. Compared to rule-based methods, transformer models are chosen by considering their flexibility, better contextual understanding, and ease of implementation. Other ML methods, including RNN, are not chosen by considering their lack of long-range dependency capturing.

3 Modelling and Evaluation

In this section, descriptions of modelling will be given on all four models. Evaluations on these models will also be performed.

The first information retrieval model utilises the BM25 ranking algorithm. Before feeding this model, claims and evidence are pre-processed, during which stop words are removed and text is tokenized, lemmatized and changed to lower case. This step is essential for better matching accuracy. After that, every piece of evidence is matched with every claim to produce a score and the top 3 shreds of evidence are taken per claim as candidates.

In the second model, a matching dictionary is first created based on spaCy's NER functions. This dictionary is then used to match the name entity (NE) in the claim, extracted using the same method, to evidence that contains the same NE. To avoid overfitting on general entities, a hurdle of 500, which is introduced to reject common NE, on the number of corresponding evidence is set. The result claim-evidence pairs are then combined with the output of the BM25 model to serve as candidates for the final Bert model.

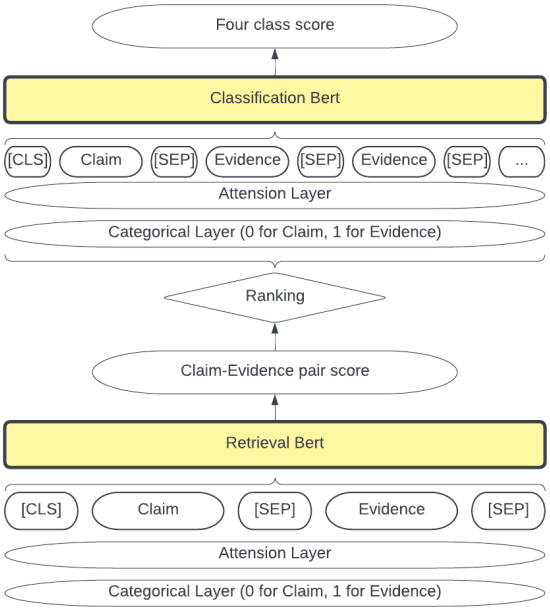

Figure 2: Structure of two Bert models

Considering the small number of the training set, the retrieval Bert model have a simple struc-

Method	True Positive	False Positive	False Negative	Retrieval Rate
BM25	743	11537	3379	0.180
NER	33	7532	4089	0.008
Retrieval Bert (BM25)	433	3180	3699	0.105
Retrieval Bert (BM25 + NER)	443	3175	3689	0.105

Table 1: Results of different retrieval methods.

True | Predict	SUPPORTS	REFUTES	NOT_ENOUGH_INFO	DISPUTED
SUPPORTS	101	13	12	5
REFUTES	8	36	2	2
NOT_ENOUGH_INFO	10	3	84	11
DISPUTED	7	4	5	22

Table 2: Confusion matrix for true and predicted labels in Classification Bert model.

ture. The training data for this Bert model is the labelled and filtered data from the BM25 model and NER model. This model contains a Bert layer, using the "bert-base-uncased" model as required, a dropout layer and a linear layer. The input claim-evidence pairs are firstly tokenized using the corresponding Bert tokenizer. These one-to-one pairs are tagged with 0s and 1s to indicate "claim" and "evidence". Attention tags are also included in the input. These three chunks of data are then fed into the Bert model. The weight of the final hidden layer for "[CLS]" is then taken and passed to a linear layer, which matches the data into the output score, through a drop-out layer. The output will be classified into RELATED or UNRELATED based on the score. The padding for claim and evidence is 80 for each. BCEWithLogisticLoss and Adam are chosen to be criterion and optimization functions.

The classification Bert model has the same structure as the retrieval Bert model. To produce the input, all RELATED claim-evidence pairs are grouped by the claim and sorted by the final score. Up to 3 top pieces of evidence will be selected. Instead of one-to-one pairs in the retrieval model, input tokens are one-to-many claim-to-evidences pairs. These texts are concatenated into a single list with category tagging and attention tagging using the same method. The training process is the same as the previous model, except the final layer maps to the scores of four output classes. Finally, classification is made based on these scores of four classes. There are some classification rules for better accuracy. Firstly, if both SUPPORT and REFUTES scores are less than 0.6 the output will be assigned to NOT_ENOUGH_INFO. Secondly, if the difference between SUPPORT and REFUTES is less than 0.05, the output will be assigned to DISPUTED. Otherwise, the result will be assigned based on the top score. The criterion and optimization functions are Adam and CrossEntropyLoss with class weight. BertForSequenceClassification and BertTokenizer are used in this model.

3.1 Claim Verification

From Table 1, it could be observed that the retrieval rate is low using this combination of models. The overall performance of BM25 heavily outweighs the performance of NER. Even though introducing NER result in a minor improvement in the model, compared to its trade-off with extra processing time, this operation is not recommended. Satisfactory performance on evidence elimination can be observed using the Bert model, with a trade-off of a significant reduction in the correct evidence retrieval number.

An explanation for the low retrieval rate in BM25 is the inadequate term weights. Similar to TF-IDF, BM25 ranking heavily relies on the document length. In this task, claims and evidences are short and therefore lead to underperformance in this ranking algorithm. NER methods were added aiming to compensate for the inability of capturing detailed phrases. The unexpected low performance may be explained by the ambiguity in the short text. It is also worth mentioning that strict hurdles on acceptance rate and noisy data may also contribute to underperformance.

The retrieval Bert model shows satisfactory performance on false evidence elimination, which is expected. The model is fine-tuned on the official pre-trained Bert model, which is not specialized in evidence retrieval. Moreover, the training dataset

True | Predict	SUPPORTS	REFUTES	NOT_ENOUGH_INFO	DISPUTED
SUPPORTS	348	60	100	11
REFUTES	25	129	41	4
NOT_ENOUGH_INFO	101	65	211	9
DISPUTED	45	25	36	18

Table 3: Confusion matrix for true and predicted labels for overall model.

is small, which might lead to under-training and overfitting in the Bert Model. These reasons may explain the high true evidence rejection rate.

The classification Bert model shows a satisfactory result of 0.748 in the training set, given the correct evidence. It is worth mentioning that the final prediction is modified by rules, which does not directly explain the performance of the Bert model alone. Because of the uncertainty in the one-to-one claim-to-evidence relationship, the model is trained using a one-to-many format as input. To reduce the information of positional information in the evidence part, an additional dataset is added by resampling with changed order of evidence.

The final accuracy in the local training set is 0.575. It is worth mentioning that this result is contributed by both retrieval and classification models. Even though the wrong evidences are retrieved, the Bert model can make relatively correct classification on the close relevant evidence. It is worth mentioning that this model tends to classify uncertain records into NOT_ENOUGH_INFO. This feature is introduced using matching rules based on the output score of the final Bert model.

4 Conclusion and Future Work

In conclusion, the claim verification task is approached using both information retrieval and classification models in this report. In general, MB25 and NER retrieval models reveal unreliable performance and both two Bert models show relatively acceptable results. It is concluded that simple independent relevance ranking and IR technics are unsuitable for information retrieval, and, even given a limited training dataset, Bert models are usable for claim verification tasks.

In future work, the information retrieval system could be improved with different methods. These include using combinations of methods, e.g., MB25 ranking on NER result, and using better algorithms. The classification model may be improved by adopting more proper pre-trained models.

References

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.

Andrew Trotman, Antti Puurula, and Blake Burgess. 2014. Improvements to bm25 and language models examined. In *Proceedings of the 19th Australasian Document Computing Symposium*, ADCS '14, page 58–65, New York, NY, USA. Association for Computing Machinery.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.