Exploration of BERT-based models for Climate Science Claims Automated Fact Checking

UOM Student ID:1181506

Abstract

This report discusses the methods and results of BERT-based models for Machine Fact Checking for the topic of Climate Science using. Performance of the BERT (CrossEncoders) and Sentence-BERT (Bi-Encoder) models with different model hyperparameters (e.g. ratio for negative-sampling; number of dense layers used after the BERT layer) for the Retrieval and Classification task are evaluated, along with discussions of directions for further exploration.

1 Introduction

Climate change is a significant challenge threatening the existence of humanity that has been bought to the centre of public attention in recent times. Despite mounting scientific supporting evidence, climate-change speculators continue to undermine the scientific community's efforts to rally a global united front to comprehensively address this issue by making counterfactual climate claims to confuse public opinion. As the development of technology (i.e. social media platforms which allows opinion to be easily propagated) have become a key weapon for non-believers to sabotage this cause, it is imperative to also develop technology to counteract this damage. Automated Fact Checkers which can be applied on social media platforms such that misinformation can be censored or labelled are one such developments.

Common Fact Checking Models consists two primary components: Evidence Retrieval - which when given a claim searches a knowledge base for the most relevant scientific evidences; and Classification - which uses the retrieved relevant evidences to predict whether the claim is truthful[3]. In this problem, labels for classification include: "SUPPORTS", "REFUTES", "NOT ENOUGH INFO" and "DISPUTED".

This experiment explored using Deep Learning (DL) models based on BERT (such as CrossEncoders or SentenceBert)[2] to tackle both components, with different variations on model hyperparameters such as negative-sampling-ratios and number of additional dense layers after the BERT layer. Subject to computational resources and time constraints, improvements with next highest pursuit priority are detailed in section 5.

The model found to have the best result is the combination of a **BERT + 1 dense layer Retriever trained on negative-sampling-ratio=2 returning the top 4 predicted most relevant evidences**, and a **Bert + 2 dense layer Classifier that uses the top 2 most related retrievals from the retriever to make the final classification**.

2 Data

The dataset provided consisted of 1228 training claims, 154 development claims and 153 test claims for a Codalab competition. The labelled sets each contain 1-5 ground truth evidence and a classification label. 1,208,827 evidences were provided, inclusive of all ground truth evidences in train and dev; a qualitative analysis show not all are related to climate change.

To better facilitate evaluation based on labelled data, the train set was further split into a new-train set and new-test set such that the test set size matched that of the dev set. Hereon, the new-train dataset will be referred to as "train" data, the new-test as "test", and the original unlabelled test data as "future".

No pre-processing was performed as BERT layers are designed to handle raw text input using the WordPiece embedding algorithm[4].

A most-frequent label 0R model for the training set will predict "SUPPORT", yielding 44% accuracy on the dev set and 34% accuracy on the test set

3 Retrieval

The first DL model explored for Retrieval was a BERT CrossEncoder[6]: **BERT-transformer**

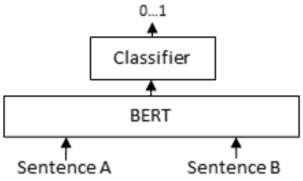

Figure 1: Architecture of *BERT Retriever/Classifier* (CrossEncoder) [5]

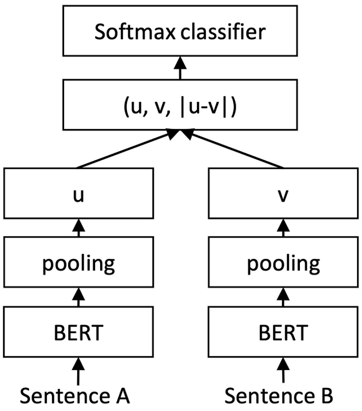

Figure 2: Architecture of *SBERT Retriever/Classifier* (Bi-Encoder) [6]

layer and a final dense layer reducing BERT's 768 dimension output down to a single score for each claim-evidence pair (*BERT Retriever*). The input to the model will be the concatenated claim and evidence that is then tokenised using BERT's WordPiece tokeniser - which then maps tokens/subtokens to an integer. Segment labels are applied to distinguish membership of sentences for token embeddings.

The second DL model explored for Retrieval was an SBERT (BERT Bi-Encoder) structure[1], using **two parallel BERT layer (one each for claim and evidence), before the 768 dimension output vector of the evidence is concatenated to the end of the claim's embedding, with an extra scalar value being the two vector's cosine similarity also concatenated to form a 1537 dimension vector, which is then reduced to a single score**. (*SBERT Retriever*).

The loss functions used to train both models are binary log-loss applied on the model outputs, after transforming by this link function: $\frac{e^{logit}}{1+e^{logit}}$. The data used to train this model consists of every ground truth pair of claim-evidence labelled 1, and sampled pairs of claims and non-ground truth evidences labelled 0. Negative sampling is used to prevent training data from being oversized, as using every claim-evidence pair leads to $O(n * m)$ memory complexity for n claims and m evidences. Additionally, without negative sampling, mini-batch sampling in the stochastic gradient descent step used to train DL models will rarely or never samples positives, causing under-fitting. A negative sampling ratio (*ns-r*) was tuned as an integer hyperparameter, where the value denotes how many negative instances to sample for each positive sample (i.e. ns-r=2 means 2 negative samples for each positive sample).

To perform prediction, every claim-evidence pair was put through the model, with relevancy of each evidence to the claim determined by predicted probabilities (i.e. the output score transformed by the link function).

3.0.1 Disadvantages and pragmatic experimental choices

A disadvantage of the BERT retriever is its inefficiency during prediction: whereas the 1.2 million evidences for *SBERT Retriever* can be precomputed using its independent evidence-BERT layer and stored, thus speeding up predictions by only needing to run one claim sentence through the claim-BERT layer and then put the concatenated vector through the dense layer to get the claim-evidence score, the *BERT Retriever* which jointly feeds in concatenated claim and evidence through a single BERT layer has complexity $O(n * m)$ for n claims and m evidences, whereas *SBERT Retriever* has complexity $O(n + m)$. $O(n * m)$ complexity for 1.2 million evidences was infeasible for predicting a set of 154 claims, so this experiment was conducted only on the 3443 used ground truth evidences in the labelled datasets (both negative sampling and prediction only looked in this reduced pool). Evidence reduction via the unsupervised learning method of clustering, and also the supervised learning methods of performing a pre-step rough ranking of evidences for each claim based on cosine similarity and embedding (Word2Vec average, Doc2Vec average, BERT embeddings, tf-idf with Singular Value Decomposition) were attempted, but with results either not reducing the evidence enough or leaving out ground truth evidences in the rough top 10000 evidence ranking, this final evidence reduction choice was adopted out of pragmatism.

To keep the experiment comparable, the *SBERT Retriever* also underwent this evidence reduction. The ns-r=1 model for SBERT Retreiver trained and predicted on the full evidence corpus was also tried in pursuit of optimal performance on BERT-based models, but its ns-r=2 and ns-r=4 versions were omitted due to computational resource constraints. The Retrievers trained on reduced samples will be denoted "-r" in results.

3.0.2 Result

The evaluation metric used for the retrieval step is F1-score, which penalises false positives as well as false negatives. An additional hyperparameter K being "the number of top evidences to retrieve" was introduced, as each claim has a different numbers of ground truth evidences so even a retriever that can rank all evidences perfectly according the hypothetical "true" ranking will likely not have the highest F1-score.

Observing results, the *BERT Retriever* using negative sampling ratio of 2 and top K=4 evidences had the highest development F1-score with 0.3374, with a test F1-score of 0.3029.

The most important result from Table 1 is that SBERT models fail to find almost any True Positives, meaning none of the ground truth evidences made the top K in the predicted rankings. A possible explanation is that the independent parallel BERT layers are not sufficient in learning the relationships in this corpus of claims and evidences that the single BERT has learnt, which allows the latter to correctly rank at least a few ground truth evidences within the top K for this task. The performance of different values of K between 2 and 5 are quite stable, with all BERT-r models topping at 4 or 5. Increasing the negative-sampling-ratio was also shown to increase the F1-Score, with the ns-r=2 closely outperforming ns-r=4. This is reasonable as the evidence corpus is all climate-based, so more negative instances were required for the model to learn to distinguish within this more semantically coherent corpus of evidence, whilst having too many negative instances will cause models to be overfit to the negative class and underfit the positive class, thus failing to produce an overall predicted ranking that is similar to the ground truth. Though this result is produced based on crudely reduced evidence, it should translate to fact-checkers with better evidence reduction methods, which all scalable fact checkers should have.

3.1 Predicting Number of Evidences for Each Claim

An attempt to use supervised learning classification to predict the number of evidence a given claim would use, using the BERT + 1 dense layer model with input being the claim sentence and output being the number of ground truth evidences, was undertaken in hope of increasing the precision of the retriever. However, the classifier's performance did not beat the baseline, so setting a blanket constant K for number of evidence to use as hyperparameter is the best solution for this problem.

4 Classification

Only the Cross-Encoder structure (*"BERT"*) discussed in Retrieval was explored for classification (*BERT Classifier*), as the Bi-Encoder structure (*"SBERT"*) classifier was abandoned during training due to low accuracy on training mini-batches. Classifiers were trained with 4 outputs and with CrossEntropy Loss (on output 4 dimensional vector transformed by soft-max).

The dataset used to train consists only of the claims and ground-truth evidence pairs, with no negative-sampling required, and instead the report explored adding a 50-neuron hidden layer before the final output. At prediction time, the 4-dimensional vector outputs of each claim-evidence pair were element-wise added, with the label with highest summed predicted value used as the final prediction.

4.0.1 Result

The classifier performances reported in table 2 are based on classifications made using evidences retrieved by the aforementioned best *BERT Retriever* in section 3 (with hyperparameter ns-r=2). The report chose to only discuss this set of accuracies rather than accuracies from classifications based on ground truth labels because ultimately the classifier will only classify based on retrieved evidences, so its best to consider the results that includes error propagation. Another benefit from reporting this score is that it represents the final performance of the whole system, rather than just the classification part.

When using retriever retrieved evidences to make predictions, the 2 dense layer *BERT Classifier* had highest dev accuracy for K=2 at 0.539, which had a test accuracy 0.487.

Table 2 shows that whilst the K that gives the

Model	**ns-r\K**	1	2	3	4	5
BERT-r	1	0.123	0.182	0.190	0.197	0.197
BERT-r	2	0.239	0.307	0.326	**0.337**	0.327
BERT-r	4	0.237	0.279	0.331	0.332	0.324
SBERT-r	1	0	0	0	0	0
SBERT-r	2	0.003	0.002	0.002	0.002	0.003
SBERT-r	4	0	0	0	0	0
SBERT	1	0	0	0	0	0

Table 1: Dev set F-1 scores of different BERT based Retrievers on different hyperparameters

Dense Layers	1	2
1	0.357	0.526
2	0.370	**0.539**
3	0.338	0.5065
4	0.364	0.513
5	0.370	0.533

Table 2: Dev set accuracy scores of BERT based classifiers for different number of dense layers based on BERT-r ns-r=2 retrievals for different values of top K retrieved evidences

best retrieval F1-score is 4, the *BERT Classifiers* prefer to use just the top 2 retrieved documents, likely because incorporating more documents could introduce noise from both variance within the classifier when classifying based on correctly retrieved evidences, but also error propagation from incorrectly retrieved evidences. Models with 2 dense layers outperforming that of 1 dense layer by a substantial margins also suggests that the 4-way classification's complexity demands the extra layer's flexibility to better capture the relationship between BERT outputs and final labels. The final chosen *BERT Classifier* beat the 0R accuracy on the test set by 0.15, a substantial result.

5 Conclusion and Future Recommendations

This report explored using BERT-based Deep Learning models to build both parts of a Retrieval-Classification structured model for Climate Science Claims Automated Fact Checking. The *BERT Retriever* and *BERT Classifier* both outperformed their SBERT counterparts, and it was discovered that the classifier performs better when predicting based on less number of top evidence than the K top evidences that provides the best retrieval performance based on the development F1-score. The best BERT retriever has a test F1-score of 0.303 and accuracy 0.487, giving a test harmonic mean (metric used for competition) of

$$\frac{2}{\frac{1}{0.303} + \frac{1}{0.487}} = 0.373$$

A priority future research area for improvement on model performance for this task is better evidence reduction methods, for example using Part of Speech tags or Name Extraction to extract nouns from the claims and evidences before repeating the aforementioned attempted supervised and unsupervised evidence reduction methods. This is likely to improve evidence reduction performance as representations on the nouns of each sentence will be more topic coherent than representations based on raw sentence/stop-word-removed-and-stemmed sentences.

From an architectural perspective, further dense layers could be considered as the experiments showed that 2 dense layer models outperformed those with 1. Batch size and learning rate are also important hyperparameters that should be tuned, provided that there is sufficient supporting hardware (i.e. GPU) to support large batch sizes in training. With the second part of the system being a classification problem, common tricks to imporve classification models such as oversampling minority classes until all classes are equally represented should not be overlooked, in spite of the models utilising advanced Deep Learning architectures.

Codalab competition returned 0.105 as retrieval F1-score and 0.403 as accuracy, based on using the best development hyperparameters for a *BERT Retriever* and *BERT Classifier*. This F1-score and accuracy combined for for 0.167 harmonic mean, a respectable outcome for an exploration of BERT-based models and different choices around training and prediction for tackling the problem of Climate Science Claims Automated Fact Checking.

References

[1] DAIR.AI. Sentencebert — semantically meaningful sentence embeddings the right way. 2020. Accessed: 2023-04-23.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. Accessed: 2023-04-25.

[3] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, 10:178–206, 02 2022. Accessed: 2023-04-24.

[4] Chetna Khanna. Wordpiece: Subword-based tokenization algorithm. 2021. Accessed: 2023-05-14.

[5] Nil Reimers. sentence-transformers (github repository). 2021. Accessed: 2023-05-14.

[6] Nils Reimers. Cross-encoders. 2022. Accessed: 2023-04-23.