Fact-Checking Climate Statements: A Combined Approach Using DPR with BERT models

Student ID: 1292534

Abstract

In the digital age, the internet is rife with misleading narratives and false claims, particularly when it comes to climate change. This paper proposes a Fact-checking system that leverages BERT models and the BM25 information retrieval model to combat this problem through a claim validation process. This claims Fact-checking system utilizes BERT for evidence classification and utilizes multiple evidence retrieval tools or techniques. And has demonstrated a compelling capacity to verify the veracity of a claim and fetch the corresponding supporting evidence.

1 Introduction

Due to the large amount of fake information about climate change on the Internet (Veltri and Atanasova, 2017), there is an urgent need for reliable mechanisms to discern fact from fiction. This paper introduces an innovative system designed to address this issue through the power of the BERT models.

The Fact-checking system described in this paper uses ALBERT (Lan et al., 2019), which is a lite version of BERT, for classifying and flagging claims on individual pieces of evidence. At the same time, this paper also attempts to use the base uncased BERT to make uniform predictions for all evidence together as one input. Both processes incorporate three techniques for retrieving evidence - DPR, Word2Vec, and BM25 (Karpukhin et al., 2020; Robertson and Walker, 1994; Mikolov et al., 2013). This work is to improve the credibility and efficiency of automated Fact-checking systems and to assist in the accurate sharing of information about climate change on the Internet. The results of this study yield valuable insights and ideas for a potential solution to mitigate the proliferation of disinformation in this critical area of climate change. As the figure will present a brief review of the whole architecture.

2 Methodology

This section outlines the architecture of the Fact-checking system, and its key components, training data and training process.

2.1 System Architecture

The system primarily operates on a BERT-based model, along with an optional choice among DPR, Word2Vec, and BM25 for evidence embedding or dictionary building, which assists in retrieving the most relevant evidence for the BERT classifier model from a large corpus.

The system pipeline consists of two primary stages: evidence embedding and claim classification. The initial stage involves creating embeddings or dictionary for all available evidence within a large corpus, using three optional embedding methods. This process calculates the similarity between the claim and each piece of evidence to identify the top 'k' pieces of evidence most relevant to the claim. The second stage involves inputting these 'k' pieces of evidence into the BERT-based model for classification. This model is designed to categorize the claim into one of four possible classes. The final output includes the assigned claim label and the IDs of the supporting evidence. As figure 1 will give a brief review of the whole architecture.

2.2 Model Components

2.2.1 Embedding Models

This subsection will describe the details of the embedding models used in the system, including Dense Passage Retrieval (DPR), Word2Vec, and BM25. The BM25 information retrieval model, as a non-embedding retrieval model, will also be discussed in this subsection.

Dense Passage Retrieval (DPR) In this project, DPR is employed to create high-dimensional vector representations of both claims and evidences. To generate these embeddings, This system utilized

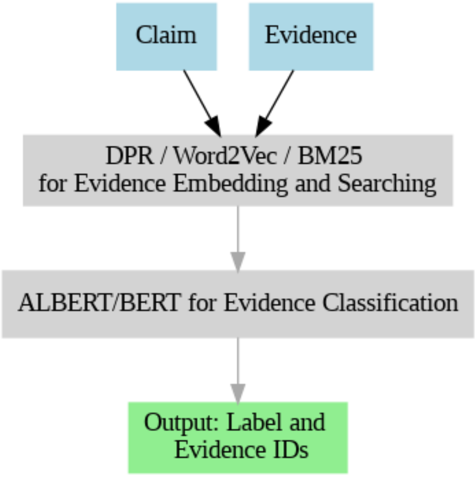
Claim
Evidence
DPR / Word2Vec / BM25
for Evidence Embedding and Searching
ALBERT/BERT for Evidence Classification
Output: Label and
Evidence IDs

Figure 1: Fact Checking System Architecture

the pre-trained models provided by Hugging Face (Wolf et al., 2020). The claim is fed into the DPR question encoder, which outputs a 768-dimensional vector encapsulating the semantic essence of the claim. Similarly, each piece of evidence is processed by the DPR context encoder to produce a corresponding vector representation. These embeddings enable the system to measure the semantic closeness between the claim and each piece of evidence, facilitating the subsequent evidence retrieval process.

Word2Vec In the pursuit of identifying the most effective and well-tested embedding method for claim validation, the Word2Vec model has also been employed in this project. Word2Vec creates embeddings by learning from the words within the Google News corpus (Mikolov et al., 2013), thereby capturing inherent semantic and syntactic patterns. The choice of Word2Vec is motivated by its ability to encapsulate semantic similarities, which is anticipated to aid in pinpointing relevant evidence in relation to a claim.

In practice, both the claim and every piece of evidence is represented as Word2Vec embeddings. These vectors are subsequently subjected to a distance metric analysis. The calculated distances, inversely proportional to the semantic similarity, used as a ranking mechanism to measure the relevance of each piece of evidence to the claim.

BM25 BM25, a prominent retrieval function in Information Retrieval, is introduced in this project as an alternative to the vector space model for evidence retrieval. The choice of BM25 stems from its distinctive approach to text retrieval, which potentially offers a complementary perspective to the vector embeddings used by DPR and Word2Vec.

In the practical implementation, a comprehensive index is created for the entire corpus, where each piece of evidence is indexed by its unique identifier. This index forms the backbone for fast and efficient text retrieval. The system incorporates both BM25 and BM25F scoring algorithms, harnessing their respective strengths to achieve a more robust and effective retrieval process (Pérez-Iglesias et al., 2009). BM25 is adept at capturing term frequency-inverse document frequency, while BM25F extends BM25 by considering multiple fields with different weights, providing an additional layer of refinement to the retrieval process. By utilizing the power of Whoosh, a Python-based search engine known for its speed, the system is able to swiftly navigate through the large climate change corpus and retrieve the most pertinent pieces of evidence for the fact-checking process.

2.2.2 Classification Model

In this subsection, the details of BERT and ALBERT models used for classification purposes will be presented. With the process of making uniform predictions for all evidence using the based, uncased BERT be explained.

2.3 Training

ALBERT With the retrieval of potential evidence using the approaches detailed previously, the next crucial step is to classify the claim's label based on found evidence. For this task, an ALBERT model is utilized, which is a lighter and faster transformer model compared to BERT. This model (2) is designed to handle the classification of a single claim in one of the three categories: 'SUPPORTS', 'REFUTES', or 'NOT ENOUGH INFO'.

Each claim in the dataset, along with its corresponding evidence and label, is processed and transformed into a 300-length vector with an ALBERT tokenizer. Notably, claims labelled as 'DISPUTED' are excluded from this stage, as the aim is to classify only the claims with definite labels. Each claim-evidence pair is then tokenized and truncated or padded to a fixed length, ensuring consistent input size for the model.

The ALBERT-based classifier used in this project is a custom-built module that integrates the ALBERT model with two fully connected layers, interspersed with dropout layers for regularization. The model extracts the pooled output from AL-

BERT, which is then fed through the subsequent layers to finally produce a 3-dimensional vector to represent claim labels.

The model is trained using a Cross-Entropy Loss function and optimized with Adam. The parameters are updated to minimize the discrepancy between the predicted labels and actual labels. At the same time, an early stopping mechanism was applied during training to prevent overfitting. If the validation loss does not decrease for a specified number of consecutive epochs, the training is halted, and the best model parameters are saved. This could prevent the chance of overfitting and save training time.

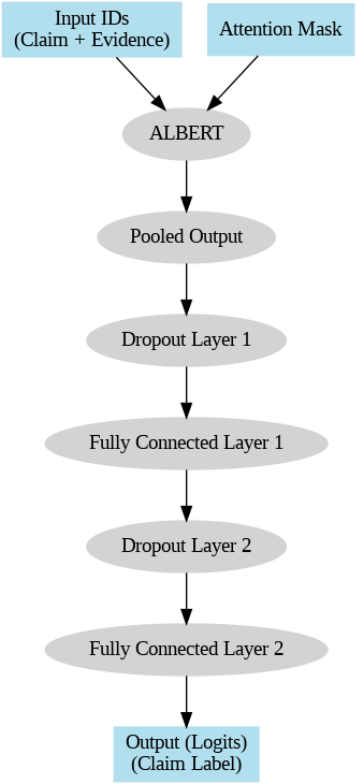

Figure 2: Fact Checking System Architecture

BERT In this system, the underlying BERT model was also deployed. In contrast to the ALBERT model, the BERT model in this task considers all evidence for each claim uniformly during the training process. This means that the model is trained to classify the evidence for each claim as a whole, rather than individually.

A noteworthy aspect of this model is its ability to handle data marked as "disagree", a feature that is not incorporated into the ALBERT model. This allows the BERT model to gain insight from a wider range of data, which may enhance the comprehensiveness of the Fact-checking process.

By considering all the evidence together, the BERT model provides a more comprehensive understanding and evaluation of each claim, which may improve its performance in Fact-checking tasks. However, the model does not have this capability when isolating which specific EVIDENCE needs to be finally returned

3 Experiments and Results

In this section, the experimental setup and results of the retrieval and classification models are presented. The experiments involve two major steps: firstly, the training data is processed and fed into the evidence retrieval models, which are BM25, DPR, and Word2Vec. Secondly, the retrieved evidence is passed to either one of the claim classification models: BERT or ALBERT. The performance of each model combination is evaluated using the F-score for evidence retrieval and Accuracy for claim classification. The Harmonic Mean of F-score and Accuracy is used as the overall performance metric. And please note that the first table 1 is not based on either model, but a direct retrieval with the 5 most relevant pieces of evidence.

Model	**F-score**
BM25	0.13
DPR	0.08
Word2Vec	0.06

Table 1: Performance of Evidence Retrieval Models(on dev)

Table 1 showcases the performance of the different evidence retrieval models, which demonstrates that BM25 performs the best among the three models.

Retrieval	**Model**	**HMF(dev)**	**HMF(test)**
BM25	ALBERT	0.13	0.12
DPR	ALBERT	0.08	0.07
Word2Vec	ALBERT	0.07	-
BM25	BERT	0.16	0.18
DPR	BERT	0.08	0.06
Word2Vec	BERT	0.09	-

Table 2: Harmonic Mean of F-score and Accuracy (HMF) for Different Retrieval and Model Combinations

Table 2 compares the performance of different combinations of retrieval models and classification models. The best performance is obtained by the combination of BM25 and BERT, indicating that this is the most effective model pairing for our task.

From these results, it can be inferred that integrating the BM25 retrieval model with the BERT classification model results in the most efficient and accurate system for claim verification. Therefore, the final model selection for the claim Fact-checking is BM25 with the base BERT model.

4 Discussion

This section reflects on the challenges encountered and potential improvements for the Fact-checking system.

Evidence Dataset Quality: The evidence dataset is inherently noisy. In a random sample test(200 pieces of evidence), only approximately 10 per cent of the data pertained to weather-related topics. Thus, combined with the restrictions on using external data for the task, significantly increased the difficulty and volume of queries. Therefore, the application of Positive-Unlabeled (PU) learning as a pre-processing step for the evidence dataset could potentially enhance the system's performance by filtering out irrelevant evidence as a future improvement option.

The choose of K nearest: The choice of K, is the number of evidence documents returned by the BM25 search, which has greatly impacted prediction performance. Considering both the quantity of returned evidence and the quality of the BERT classifier. K=5 was selected,

BM25 usually retrieves relevant evidence within the top 50 documents. However, a larger K could strain the classifier's capacity. If future improvements can enhance the BERT classifier's performance, experimenting with larger K values may be viable, potentially leading to increased prediction performance.

Training Data for BERT Models: The quantity of training data for both BERT models was insufficient, which adversely affected their performance. In practice, the majority of claims were classified as "NOT ENOUGH INFO", indicating that the models struggled to confidently classify the claims due to the lack of diverse training samples. Future work could involve the collection of more diverse and balanced training data to improve the model's performance.

Performance of DPR and Word2Vec: In this study, it was observed that both DPR and Word2Vec models tend to favor sentences with similar structures during retrieval, even when the semantic content of these sentences might differ significantly. For instance, sentences that follow the structure "person X said statement Y" would be considered similar by these models, regardless of the actual content of the statements. This structural bias might have led to the lower performance of these models in retrieving the most relevant evidence.

Despite these results, it is important to note that these findings do not undermine the potential capabilities of DPR and Word2Vec. It suggests that these models may need further fine-tuning to better handle the specific nature of this task. The frequency-based retrieval model BM25, despite being less sophisticated than the embedding-based models, was able to perform well due to its robustness to structural similarities.

One potential direction for future work could be developing a custom embedding model trained specifically on a claim-type dataset. This could potentially mitigate the structural bias of the existing models and improve the overall retrieval performance.

5 Conclusion

This paper presents an experimental Fact-checking system designed to address the problem of misleading narratives and false claims about climate change that proliferate on the Internet. It utilizes the BERT model and various evidence retrieval techniques, including BM25, Dense Passage Retrieval (DPR), and Word2Vec. The system demonstrates the ability to verify the veracity of a claim and obtain appropriate supporting evidence.

References

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Joaquín Pérez-Iglesias, José R Pérez-Agüera, Víctor Fresno, and Yuval Z Feinstein. 2009. Integrating the probabilistic models bm25/bm25f into lucene. *arXiv preprint arXiv:0911.5046*.

Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*, pages 232–241. Springer.

Giuseppe A Veltri and Dimitrinka Atanasova. 2017. Climate change on twitter: Content, media ecology and information sharing behaviour. *Public understanding of science*, 26(6):721–737.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.