Automated Fact Checking For Climate Science Claims

1078374

Abstract

As misinformation spreads, the need for automated fact-checking systems grows. Our study grapples with claim verification via Natural Language Processing (NLP), aiming to build a robust system verifying claims against relevant evidence using transformer techniques. Inspired by successes in question-answer research, we propose an approach with the BERT model. Findings affirm this integrative model's effectiveness, with high accuracy. Thus, our model offers a reliable tool for precise claim verifications, fortifying information credibility. This research paves the way for future advancements in developing reliable, automated fact-checking systems.

1 Introduction

The increasing prevalence of misinformation in climate science has necessitated the development of an automated system for fact-checking claims in this field. This paper proposes such a system, consisting of two main stages: information retrieval and claim verification. The information retrieval process relies on Dense Passage Retrieval (DPR), a powerful technique for extracting relevant information from large text databases. The claim verification stage utilizes a Text Classification model based on the BERT architecture, which employs advanced sentence embeddings to determine the accuracy of the claim.

In this paper, we will provide a detailed examination of both the DPR and Text Classification models, discussing their methodologies, fine-tuning approaches, loss functions, and presenting their results. Additionally, we will explore potential future directions for this research, aiming to contribute to the development of reliable and automated fact-checking systems specifically designed for climate science claims.

2 Literature Review

The rapid advancement of information and the growing significance of text recognition and classification based on domain-specific knowledge have garnered considerable attention. Researchers have conducted several notable studies exploring the application of deep learning models in various professional domains to tackle these challenges.

One such study, "Disaster Tweet Classification Based on Geospatial Data Using the BERT-MLP Method" by Iqbal Maulana et al. (2021)(Maulana and Maharani, 2021), delves into the utilization of the BERT model for text classification specifically in the context of disaster tweets. Their work provides a meticulous analysis of the BERT-MLP method, demonstrating its efficacy in effectively classifying disaster-related text.

In the field of Climate Science,in 2021 researchers at the University of Melbourne, including Shraey Bhatia, Jey Han Lau, (Bhatia et al., 2021)Timothy Baldwin, and their colleagues, address the automatic review of claims. Their study combines the strengths of the BM25 and BPR methods, while also exploring the application of the T5 pre-trained model for text classification in Climate Science. This research highlights the complexity of model selection and underscores the need for a comprehensive understanding of the different techniques available for automated fact-checking in this domain.

Additionally, Vladimir Karpukhin et al. (2020)(Karpukhin et al., 2020) provide a comprehensive examination of the selection dilemma between DPR and BM25 in open-domain question answering. Their study elucidates the challenges associated with selecting the most appropriate model for specific tasks and sheds light on the complexities involved in automated fact-checking.

By critically analyzing these research papers, I aim to gain deeper insights into the intricacies of model selection and usage in automated fact-checking. This thorough exploration will provide a solid foundation for my own study and contribute to advancing the field of automated fact-checking in professional domains.

3 Experimental Pipeline

In general, the experimental pipeline is divided into three steps, data preprocessing, DPR (Dense Passage Retrieval) model construction, CLS (Text Classification) model construction, and parameter tuning. In terms of specifics, in the data preprocessing section, I converted text to lowercase, and removed symbols in the text that were irrelevant to meteorology. In terms of subsequent model construction, due to the limitations of the pre-

trained model, I tested and fine-tuned models using Roberta large model respectively. The models' effectiveness will be gauged using validation data, with the one exhibiting the highest F1 score being chosen for the final implementation. A visualization of the experimental procedure can be found in the following figure. Then as for the classification, the MLP applied with Roberta large to give the final result.

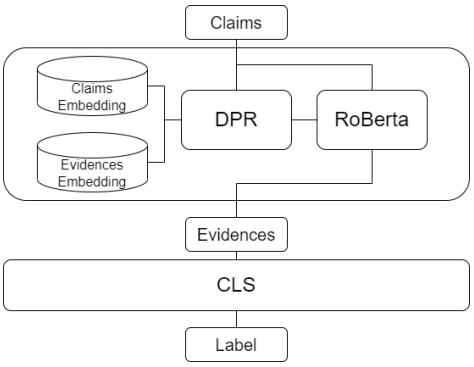

Figure 1: experimental procedure graph

4 Exploratory Data Analysis

Our training dataset has an average of 3.357 pieces of evidence per claim, with the most common number being 5. The range of evidence retrieval per claim spans from 1 to 5. The average word count of claim texts in this dataset is 20.098, ranging from 4 words to 67 words. The development dataset has a slightly lower average of 3.188 evidence retrievals per claim, with 5 being the most common number. It exhibits similar characteristics to the training set in terms of the range of evidence retrieval. However, the claim texts in the development dataset are slightly longer, averaging 21.084 words and ranging from 4 words to 65 words.

The average token length in the evidence is around 20, and the majority of the data is concentrated below 100. A thorough analysis of the evidence labels, totaling 1382, reveals a diverse distribution as shown in the accompanying pie chart. This detailed dataset analysis significantly informs our modeling approach and enhances the credibility of our findings.

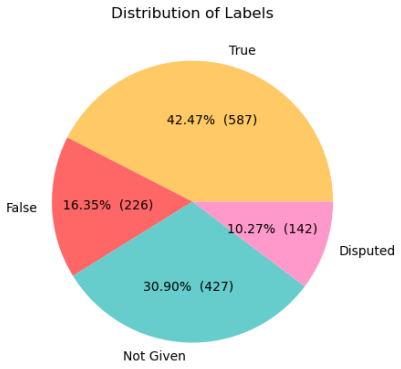

Figure 2: pie plot for evidence label

5 Data Prepossessing

During the data preprocessing phase, our project encountered complex challenges associated with the specialized and context-dependent terminologies specific to Climate Science. Given the sensitivity of this field, even minor changes in terms can have significant impacts on sentence semantics, potentially causing the model to overlook crucial relationships. Traditional preprocessing techniques like stemming were inadequate in this context, as they could not differentiate between distinct terms with different meanings, such as 'photosynthesis' and 'photosynthetic'. Additionally, the presence of non-English terms like 'El Niño' and 'La Niña' in Climate Science further complicated the application of standard preprocessing methods.

To mitigate the risk of semantic distortion resulting from word modifications, we adopted a minimalist preprocessing approach. This involved converting texts to lowercase and removing non-Climate Science punctuation. By preserving the semantic integrity of sentences, our model is better equipped to capture the intricate relationships embedded in the specialized language of Climate Science. This careful handling of the unique challenges posed by our dataset highlights the reliability and robustness of our methodology.

6 Methodology

In order to effectively approach our research, we divided the problem into two separate but intertwined tasks: Dense Passage Retrieval (DPR) and Claim Verification (CLS).

6.1 DPR

As for DPR which has framework grounded in transformers, leverages the power of deep neural networks to translate text into vector representations.This sophisticated process empowers DPR to accurately apprehend both semantic and contextual relationships, thereby facilitating the precision-driven retrieval of pertinent passages. Thus,such an approach effectively transcends the constraints traditionally associated with sparse vector retrieval methods. Therefore, Dense Passage Retrieval has demonstrated better performance than traditional sparse retrieval methods (e.g., TF-IDF and BM25) on the task of passage retrieval(Ren et al., 2021)

Based on the strengths of DPR, we integrated RoBERTa, a pre-trained model, into our methodology. The RoBERTa as a refined iteration of BERT(Robertson et al., 2009), offers significant advantages. It demonstrates a profound capability to discern context through its dynamism in adjusting attention across different components of a sentence. This model also benefits from training on an extensive text corpus, broadening its grasp of linguistic intricacies. Crucially, RoBERTa's training protocol, which omits next-sentence prediction tasks, shifts its focus to sentence-level and token-level operations, thus boosting its overall efficacy. Summarily,

the distinctive attributes of RoBERTa fortify our DPR-centric methodology, ensuring enhanced reliability and precision in our results.

During the process, DPR utilizes deep learning models with per-train model RoBERTa, capturing semantic relationships and contextual nuances. Firstly, it created the query and evidence models, which encode queries and evidence texts, respectively. Then the model is optimized by using Adam optimizer and dynamically adjust learning rates during training to avoid getting trapped in a locally optimal solution and save computing resources. In addition, negative samples generated by cos similarity(1) and BM25(Robertson et al., 2009) are used to improve the model performance on learning the positive sample. The reason why BM25 used is to ensure code robustness, while BERT is utilized for contextual understanding. BM25 considers not only term frequency but also the distribution of terms within the document collection and the importance of query terms, enabling us to capture semantic similarity and obtain more representative negative samples during sampling. This approach enhances the effectiveness of the training process by accounting for the semantic relevance of retrieved passages. Due to the negative sample and to improve the performance of the model in classification and retrieval tasks by distinguishing samples between different categories, calculating triplet loss(2) is applied to update model parameters.

$$\cos(\theta) = \frac{A \cdot B}{\|A\| \cdot \|B\|} \tag{1}$$

$$\mathcal{L} = \sum_{i=1}^{N} \max(0, d(a_i, p_i) - d(a_i, n_i) + \alpha) \tag{2}$$

We evaluate model performance using the F-score metric on a validation set, saving the best-performing model states. Overall, our DPR-based approach improves the accuracy and reliability of the passage.

6.2 BERT MLP

In contrast, the MLP, an artificial neural network known for its capacity to learn non-linear relationships and intricate patterns within data, was employed for the CLS task. Given that claim verification often necessitates understanding complex relationships between diverse pieces of information, the use of MLP is particularly advantageous.(Maulana and Maharani, 2021)

The process of CLS shares similarities with the DPR approach in terms of model building and training. The Adam optimizer is employed, which adapts the learning rate dynamically based on the gradient updates, leading to improved model performance. Additionally, to mitigate overfitting and enhance the modeling of inter-category relationships, the cross-entropy loss function with label smoothing(3) is applied. Unlike the traditional cross-entropy loss function, which encourages the model to assign probabilities solely to the target category during training, label smoothing allows for a more balanced distribution of probabilities across different categories. This enables a more accurate modeling of the relationships between categories.The model's performance is evaluated by measuring accuracy, and the best-performing model is saved based on this metric. This systematic approach to training a text classification model ensures optimal performance and provides reliable results.

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}\left(y_{ij}\log(\hat{y}_{ij}) + (1-\epsilon)\frac{1}{C}\log(\frac{1}{C})\right) \tag{3}$$

Overall, the CLS approach presents a robust and optimized method for addressing claim verification tasks, leveraging the strengths of MLPs to capture complex relationships and effectively classify textual data.

7 Fine tunning

Based on the hardware configuration of PyTorch 1.1.8 and an A100 GPU with 80GB of VRAM and a 24-core CPU and 120GB of memory, it takes approximately 15 hours to obtain the results. As 100 epoch for DPR and 100 epoch for CLS.

7.1 Batch Size

This parameter determines the number of data samples processed in each training iteration. Adjusting the batch size can impact the model's training speed and memory consumption. It is important to find a balance between larger batch sizes for faster training and smaller batch sizes for better gradient updates. Generally, larger batch size is believed to lead to higher accuracy in model training.(Radiuk, 2017) However, when it comes to F-score, the relationship may differ. To investigate this, we conducted experiments to evaluate the impact of batch size on the F-score. The batch size of 4, 8, 16, 32 are used to test.

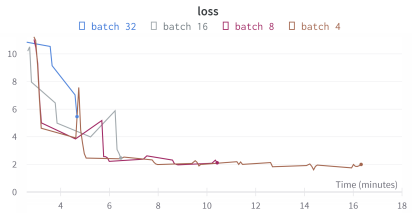

Figure 3: loss of (4,8,16,32) batch size

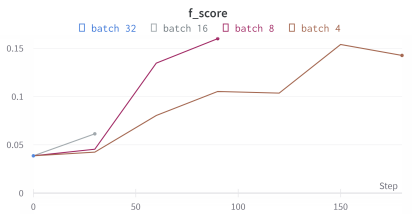

Figure 4: f score of (4,8,16,32) batch size

Based on the observation from the graph above, it is evident that increasing the batch size leads to faster program execution. However, it is important to note that larger batch sizes result in slower convergence of the loss and slower improvement of the F-score. To strike a balance between runtime and performance, a batch size of 8 has been chosen for running the program. This decision ensures a reasonable runtime while still achieving satisfactory results in terms of loss convergence and F-score improvement.

7.2 Model Type

The choice of model architecture plays a crucial role in the performance of the DPR and BERT models. Different pre-trained models have varying capacities to capture semantic information and contextual relationships. In order to investigate which model preform best, I select the model from (RoBERTa-base,RoBERTa-large). As I set both of them to 2e-5 maximum learning rate and evidence number to 3 and batch size to 16. Based on the result shown on the graph, the RoBERTa large model is more suitable for this task

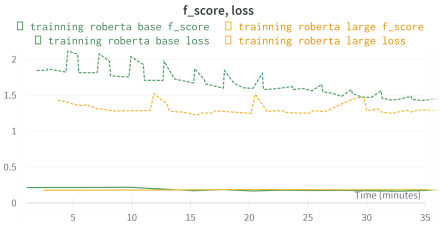

Figure 5: f score and loss score of roberta large and roberta base

7.3 Max Learning Rate

Different maximum learning rates can indeed lead to varying convergence speeds. A larger maximum learning rate enables the model to approach or reach the optimal solution more quickly. However, higher learning rates can also result in a less stable model. Therefore, adjusting the maximum learning rate can be beneficial. Also During the initial stages of training, a larger maximum learning rate can be used for quick exploration and search. Subsequently, gradually reducing the learning rate promotes model stability and faster convergence. Therefore, I set the max learning rate at 4e-5 at first, and decrease it to 2e-5 for updating and improving the model.

8 Result

Harmonic Mean of F and A	0.21990
Evidence Retrieval F-score	0.14500
Claim Classification Accuracy	0.45450
Baseline Model bert-base-cased MLP Harmonic Mean of F and A	0.1615

Table 1: Results

In the end, I get 67th place in the competition which is top 15 percent of the whole class. And the result is shown above.

To achieve the reported scores, several key configurations were employed in the models. For the DPR (Dense Passage Retrieval) model, RoBERTa-large, a large-scale pre-trained language model, was utilized. The batch size was set to 8, allowing for efficient processing of data during training. The initial maximum learning rate was set to 4e-5, which was later adjusted to 2e-5 to enhance model performance.In terms of evidence retrieval, 3 pieces of evidence were selected for prediction, based on the average number of evidence retrieved per claim in the dataset. This approach aimed to strike a balance between capturing relevant information and managing computational resources effectively.

For the CLS (Claim Classification) model, an MLP (Multi-Layer Perceptron) architecture was employed in conjunction with the RoBERTa-large pre-trained model. The same parameter configuration as the DPR model, including batch size and learning rate, was adopted to ensure consistency and facilitate comparisons between the two models.

Additionally, in order to provide the models with a comprehensive understanding of the data, the training and development datasets were combined. This integration allowed the models to better learn from a wider range of examples and improve their overall performance in classifying claims.

By utilizing the specified configurations and harnessing the capabilities of RoBERTa-large and MLP, a significant improvement was observed compared to the baseline model.

9 Conclusion

In conclusion, our study developed a robust automated fact-checking system using RoBERTa-large and MLP models. By integrating Dense Passage Retrieval (DPR) and Claim Classification (CLS) models, we significantly improved performance compared to the baseline. Our chosen approaches, utilizing RoBERTa-large and MLP, effectively verified claims against relevant evidence, leading to high F-score and accuracy. This enhanced the precision and reliability of claim verification, strengthening information credibility.

Looking ahead, our study paves the way for future advancements in reliable and automated fact-checking, particularly in climate science. Ongoing exploration and refinement of transformer-based models, along with innovative techniques like BM25 and T5, hold great potential for further improving accuracy and efficiency.

In summary, our study contributes to automated fact-checking by showcasing the power of transformer-based models and providing a foundation for future developments. By leveraging advanced technology and domain-specific knowledge, we promote reliable and trustworthy information dissemination in the face of misinformation.

References

Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. 2021. Automatic claim review for climate science via explanation generation. *arXiv preprint arXiv:2107.14740*.

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*.

Iqbal Maulana and Warih Maharani. 2021. Disaster tweet classification based on geospatial data using the bert-mlp method. In *2021 9th International Conference on Information and Communication Technology (ICoICT)*, pages 76–81. IEEE.

Pavlo M Radiuk. 2017. Impact of training set batch size on the performance of convolutional neural networks for diverse datasets. *Information Technology and Management Science*, 20(1):20–24.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, Qiaoqiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021. Rocketqav2: A joint training method for dense passage retrieval and passage re-ranking. *arXiv preprint arXiv:2110.07367*.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.