Automated Fact-Checking System for Climate Change Claims TF-IDF Evidence Retrieval & Transformer Claim Classification

Sunchuangyu Huang, Wei Zhao, Xuan Wang
The University of Melbourne
{sunchuangyuh, weizhao1, xuanwang8}@student.unimelb.edu.au

Abstract

Climate change poses a significant threat, exacerbated by misinformation. Our project develops an automated fact-checking system for climate change claims, focusing on evidence retrieval and claim classification. The system uses TF-IDF for keyword extraction and Transformer-based models for classification. By processing a large dataset of evidence passages, we aim to reduce misinformation and enhance the reliability of public discourse on climate change. Preliminary results underscore the importance of effective evidence retrieval and highlight challenges with imbalanced datasets, guiding future improvements for a more robust and reliable system.

1 Introduction & Background

Climate change poses a significant threat to humanity, with rampant misinformation distorting public opinion and hindering effective action. Automated fact-checking systems can help address this issue by verifying claims using reliable evidence. Our system focuses on two key tasks: evidence retrieval and claim classification. Given a claim, it searches for relevant evidence from a knowledge source and classifies the claim as SUPPORTS, REFUTES, NOT ENOUGH INFO, or DISPUTED. In this project, we design and implement an automated fact-checking system specifically for climate change claims. Our approach leverages techniques learned from lectures and workshops, focusing on training models from scratch using a provided dataset, which includes claims and a large corpus of evidence passages, divided into training, validation, and test sets. By developing this system, we aim to reduce misinformation about climate change and enhance the reliability of public discourse. Our findings and methodologies could serve as a foundation for further research and development in automated fact-checking systems.

2 Dataset Exploratory Data Analysis

2.1 Large Evidence Knowledge System

The full evidence corpus comprises approximately $1,208,827$ records, forming a vast knowledge base for evidence retrieval.

Passage Length	Count
Very Short ($\leq$ 5 words)	$13,578$
Short (6-10 words)	$180,771$
Medium Short (11-20 words)	$547,481$
Medium (21-50 words)	$451,498$
Medium Long (51-100 words)	$15,217$
Long (101-200 words)	270
Very Long (> 200 words)	12

Table 1: Distribution of evidence passage lengths.

The length of evidence passages also plays a crucial role in evidence retrieval and claim classification tasks. To understand the distribution of evidence lengths, we categorized the passage lengths into different groups, from very short to very long. This diverse range of passage lengths highlights the need for effective preprocessing and model design to handle the varying evidence characteristics. In particular, the large number of short passages ($13,578$ passages with ≤ 5 words) may pose challenges when calculating cosine similarity with TF-IDF, as these short passages are more likely to be close to the claim context, potentially leading to poor performance in the evidence retrieval task.

2.2 Imbalanced Claim Training Data

The training dataset contains $1,228$ records and exhibits a significant imbalance in the distribution of claim labels. The data shows that the majority of claims are labeled as "SUPPORTS" (519 instances), while "DISPUTED" claims are the least frequent (124 instances).

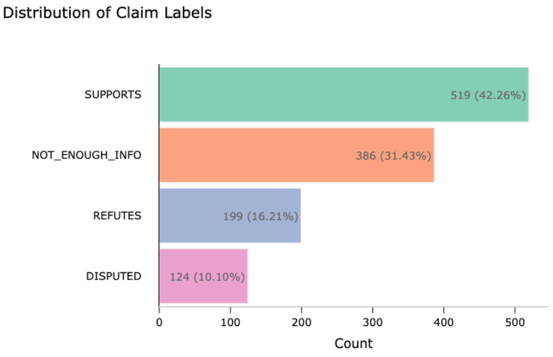


Figure 1: Claim labels distribution in the training set.

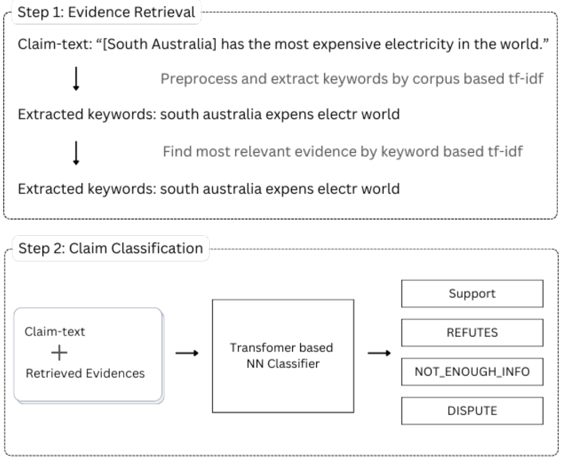
This imbalanced dataset can lead to issues in model training and evaluation, such as biased model predictions, unstable model convergence, and inaccurate evaluation metrics.

2.3 Preprocessing

The preprocessing stage is crucial for improving the performance of our automated fact-checking system. The key preprocessing steps include expanding contractions, converting to lowercase, handling concatenated words, removing punctuation, tokenizing the text, removing stopwords, and applying stemming. These steps transform the raw data into a more structured and informative format, which is then used as input for the keyword extraction, evidence retrieval, and claim classification components of our system.

3 Workflow

The workflow of our automated fact-checking system consists of two main components: evidence retrieval and claim classification. The evidence retrieval component uses the extracted keywords to search the corpus and retrieve the most relevant evidence passages. The claim classification component then takes the claim text and the retrieved evidence as input, and predicts the appropriate label using a Transformer-based neural network model. This two-stage workflow, as shown in Figure-2, allows our automated fact-checking system to leverage the large evidence corpus and make informed decisions on the veracity of the given claims. The following sections provide a detailed explanation of each component and the role of keyword extraction in the overall process.


Step 1: Evidence Retrieval
Claim-text: "[South Australia] has the most expensive electricity in the world."
Preprocess and extract keywords by corpus based tf-idf
Extracted keywords: south australia expens electr world
Find most relevant evidence by keyword based tf-idf
Extracted keywords: south australia expens electr world
Step 2: Claim Classification
Claim-text
+
Retrieved Evidences
Transfomer based NN Classifier
Support
REFUTES
NOT_ENOUGH_INFO
DISPUTE

xw

Figure 2: Two Stage Workflow Diagram

3.1 Preliminary Study

3.1.1 TF-IDF (Term Frequency-Inverse Document Frequency)

In the initial phase, we used Term Frequency-Inverse Document Frequency (TF-IDF) as the primary technique for keyword extraction and evidence retrieval. TF-IDF is a commonly used statistical method in text mining and information retrieval that reflects how important a word is to a document in a corpus. According to Aggarwal's article 2023, It composed by three parts:

1. **Term Frequency (TF):** This measures the frequency of a word in a specific document. It is calculated by dividing the number of occurrences of the word by the total number of words in the document. This metric gives higher weight to terms that appear more frequently in a single document.

$$d = \text{Total number of terms in document}$$
$$t = \text{NO. of times term appears in } d$$
$$\text{TF}(t, d) = \frac{t}{d}$$

2. **Inverse Document Frequency (IDF)**: This measures the importance of the term across a set of documents. IDF decreases as the number of documents containing the word increases, which helps to adjust for the fact that some words appear more frequently in general. IDF is calculated as the logarithm of the ratio of the total number of documents to the number of documents containing the term.

$n = \text{no. of documents containing word } W$

N = N is the total no. of documents

$$\text{IDF}(t, D) = \log(\frac{N}{n})$$

3. **TF-IDF Calculation**: The TF-IDF value is obtained by multiplying TF and IDF. This value is higher for a term that is more characteristic of a particular document, thereby helping to differentiate documents within the same corpus based on their key terms.

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D)$$

The TF-IDF approach allowed us to identify the most relevant terms within the claim text and match them against the evidence corpus to find the most relevant passages. Through our experiments, we observed that the TF-IDF-based evidence retrieval exhibited challenges with the diverse range of evidence passage lengths. The large number of short passages in the corpus were more likely to be close to the claim context, leading to poor performance in the evidence retrieval task.

3.2 Evidence Retrieval - Keyword Extraction

The keyword extraction process is a crucial step in the evidence retrieval component of our automated fact-checking system. By identifying the most relevant keywords from the claim text, we can effectively search the evidence corpus and retrieve the most relevant passages. The process involves transforming the preprocessed claim text into a TF-IDF vector, which assigns a numerical value to each word based on its importance within the evidence corpus. The top N values in the TF-IDF vector are then identified, and the corresponding keywords are extracted. The keyword extraction process can be envisaged as a clustering exercise. When the claim text and the evidence passages share common keywords, they are "closer" in the TF-IDF vector space, resulting in a smaller angle when calculating the cosine similarity. This clustering of similar keywords allows the evidence retrieval component to efficiently identify the most relevant evidence passages for a given claim (Mustafa et al., 2021).

$$\text{cosine\_similarity}(\mathbf{A}, \mathbf{B}) = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

3.3 Classification Model

3.3.1 Model Comparison

To improve the claim classification task, we explored the use of custom Transformer and LSTM-based models. The Transformer model, introduced in the "Attention Is All You Need" paper (Vaswani et al., 2017), uniquely relies entirely on attention mechanisms, discarding traditional recurrent and convolutional neural network architectures. Key features of the Transformer include:

1. **Architecture**: Composed of an encoder and a decoder, each containing multiple layers with attention mechanisms and fully connected networks.
2. **Attention Mechanisms**: Utilizes scaled dot-product attention and multi-head attention to process various parts of the input data simultaneously, enhancing parallelization and efficiency.
3. **Positional Encoding**: Adds positional encoding to input embeddings to maintain sequence order, compensating for the absence of recurrence in the model.

To evaluate and compare the performance of these models, we trained both models on the train-claims dataset containing 1228 rows of training data.

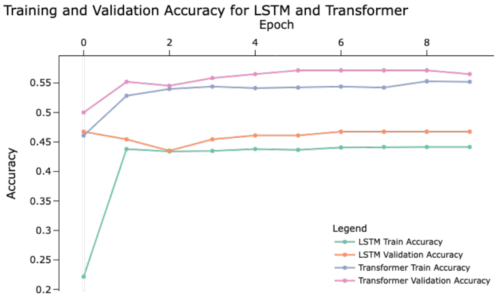


Figure 3: Training and Validation Accuracy for Transformer and LSTM Models

As shown in Figures 3, the Transformer model outperformed the LSTM model in both training and validation accuracy. The Transformer model achieved a validation accuracy of 57.14%, whereas the LSTM model achieved a lower validation accuracy at 46.75%. These results may come from that the Transformer's ability to capture long-range dependencies and contextual relationships through its self-attention mechanism, which is particularly beneficial for understanding complex claim-evidence relationships even with a very small training dataset.

3.3.2 Transformer

Given the superior performance of the Transformer model, it became our primary choice for the claim classification task (Vaswani et al., 2017). Transformer architectures have demonstrated state-of-the-art performance in various NLP tasks, including text classification. We designed a Transformer-based model that takes the claim text and the retrieved evidence passages as input and predicts the appropriate claim label. The model's ability to capture the contextual relationships between the claim and evidence proved crucial in improving the classification accuracy. However, the initial results from the Transformer-based classification model were not as promising as expected, with an accuracy of only 57.14%. There is a need for further refinements in the evidence retrieval process, as the quality and relevance of the retrieved evidence directly impact the classification performance.

3.4 Claim Classification System

The claim classification component takes the claim text and the retrieved evidence passages as input and predicts the label for the claim, which can be one of the following: SUPPORTS, REFUTES, NOT_ENOUGH_INFO, or DISPUTED. The input preparation step combines the claim text and the retrieved evidence passages into a single input sequence for the classification model. A Transformer-based neural network model is then used to learn the relationship between the claim and the evidence, and predict the appropriate label. The classification model is trained on a labeled dataset and evaluated on a held-out development set to ensure robust and accurate predictions. Once the model is trained and evaluated, it is used to make predictions on the test claims, and the predicted labels along with the associated evidence passages are compiled into the final output format.

4 Experiments and Evaluation

4.1 Evaluation Methodology

To evaluate the performance of our automated fact-checking system, we use a combination of standard evaluation metrics commonly used in the field of text classification. We report the following metrics for fact checking system:

1. **F-score**: The F-score considers both precision and recall, making it suitable for evaluating the evidence retrieval performance. It reflects the balance between retrieving relevant evidence and minimizing false positives in the retrieval process.

2. **Accuracy**: The accuracy measures the proportion of correctly classified claims, which is essential for evaluating the classification performance. It indicates the model's ability to classify claims accurately based on the provided evidence.

3. **Harmonic Mean**: The Harmonic Mean combines the F-score and Accuracy to provide a balanced evaluation metric. It offers a comprehensive assessment by considering both the evidence retrieval quality and the claim classification accuracy.

4.2 Experimental Results

	Validation	Test
F-score	0.04299	0.03310
Accuracy	0.55844	0.40790
Harmonic Mean	0.07984	0.06120

Table 2: Fact Checking System Performance

4.3 Results and Analysis

The results indicate that the poor evidence retrieval significantly impacts the final claim classification accuracy. While the accuracy may appear satisfactory when analyzing the validation data, the model demonstrates poor generalization when evaluated on the held-out test set. The significantly lower F-score and Harmonic Mean on the test set compared to the validation set suggest that the model has overfit to the evaluation data and struggles to maintain performance on new, unseen claims. This discrepancy highlights the importance of thorough testing and the need for a robust evidence retrieval component to ensure the system can generalize well to real-world scenarios.

5 Discussion and Conclusion

The results of our automated fact-checking system for climate change claims have revealed several key insights. One of the primary findings is the significant impact of the evidence retrieval component on the overall performance of the claim classification task. The significantly lower F-score and Harmonic Mean observed on the held-out test set, compared to the validation data, highlights the crucial role of the evidence retrieval process in the system's ability to accurately verify the claims. In addition, the poor generalization observed on the test set suggests that the model has likely over-fit to the evaluation data and struggles to maintain its performance on new, unseen claims. This discrepancy underscores the importance of thorough testing and the need for a robust evidence retrieval component that can effectively handle the diverse characteristics of the evidence corpus. Moreover, the diverse range of evidence passage lengths, particularly the large number of short passages, poses a significant challenge for the TF-IDF-based approach used in the preliminary evidence retrieval process. The inability of the TF-IDF method to effectively capture the contextual relationships between the claim and the evidence passages has likely contributed to the sub-optimal performance in the evidence retrieval task. Furthermore, the imbalanced nature of the training data, with a disproportionate number of "SUPPORTS" claims, presents an additional challenge that requires careful consideration. This class imbalance can lead to biased model predictions, unstable model convergence, and inaccurate evaluation metrics, all of which can hinder the of a reliable fact-checking system.

6 Limitation

The key limitations of our automated fact-checking system for climate change claims primarily revolve around the challenges faced in the evidence retrieval component and the imbalanced nature of the training data. One of the significant limitations is the reliance on the TF-IDF-based approach for evidence retrieval. While the TF-IDF method is widely used, it has inherent difficulties in effectively capturing the contextual relationships between the claim and the evidence passages. The diverse range of evidence passage lengths, particularly the large number of short passages, poses a significant challenge for the TF-IDF-based cosine similarity calculations, leading to sub-optimal performance in the evidence retrieval task. Another limitation is the imbalanced nature of the training data, with a disproportionate number of "SUPPORTS" claims compared to other label categories. This class imbalance can lead to biased model predictions, unstable model convergence, and inaccurate evaluation metrics, which can hinder the development of a fair and reliable automated fact-checking system. The discrepancy observed between the validation and test set performance also highlights a limitation in the model's ability to generalize. The significant drop in performance on the held-out test set suggests that the model has likely over-fit to the evaluation data and struggles to maintain its effectiveness on new, unseen claims. This limitation underscores the importance of thorough testing and the need for a more robust model design that can consistently perform well on diverse and unseen data.

7 Future Work

To address these limitations and build upon the current findings, several aspects for future work can be explored. Enhancing the evidence retrieval component by investigating more advanced techniques, such as Transformer-based models or other neural network architectures, could potentially lead to significant improvements in the relevance and quality of the retrieved evidence. These approaches may be better equipped to capture the contextual relationships between the claim and the evidence, thereby enhancing the overall performance of the system. Addressing the data imbalance issue is another crucial aspect that requires attention. Employing techniques such as data augmentation, class weighting, ensemble methods, or transfer learning could help mitigate the impact of the imbalanced dataset and improve the model's ability to learn from the available data effectively.

8 Conclusion

This project developed an automated fact-checking system for climate change claims, using a two-stage approach: TF-IDF-based evidence retrieval and Transformer-based claim classification. Our system classifies claims as SUPPORTS, REFUTES, NOT ENOUGH INFO, or DISPUTED by retrieving relevant evidence and assessing the claim's validity. Key findings highlight the crucial role of effective evidence retrieval. The TF-IDF method, while useful, faced limitations with the diverse range of evidence passage lengths and struggled to capture contextual relationships between claims and evidence. This underscores the need for more advanced retrieval techniques to improve the relevance and quality of retrieved evidence. Another significant challenge was the imbalanced nature of our training data, leading to biased model predictions and unstable performance. Despite these challenges, our system shows potential in reducing climate change misinformation. Future work should focus on integrating more sophisticated retrieval methods and addressing data imbalance to achieve higher accuracy and reliability.

References

P. Aggarwal, R. Kaur, and P. Aggarwal. 2023. Text classification framework for short text based on tfidf-fasttext. *Multimedia Tools and Applications*.

G. Mustafa, M. Usman, L. Yu, M. T. Afzal, M. Sulaiman, and A. Shahid. 2021. Multi-label classification of research articles using word2vec and identification of similarity threshold. *Scientific Reports*, 11(1).

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Team Contributions

Sunchuangyu Huang

As a member of the project team, my contributions focused on key aspects of the automated fact-checking system for climate change claims. In the system design ideation phase, I offered ideas on the two-stage approach and emphasized the importance of the evidence retrieval component, helping shape the decision-making process. During data preprocessing and analysis, I developed strategies to handle the diverse evidence passage lengths and imbalanced claim label distribution, ensuring the data was ready for further analysis and model training. I implemented the evidence retrieval logic using the TF-IDF method, designing and testing the retrieval algorithms, and identifying areas for improvement. Additionally, I contributed to writing and compiling the project report and played a key role in preparing and delivering the project presentation.

Wei Zhao

As a member of the project team, my contributions were primarily focused on building and optimizing the classification model for the automated fact-checking system. During the model development phase, I designed and implemented both Transformer and LSTM models, conducted training and evaluation to determine their performance on our dataset, and selected the best-performing model for the final system. I compared the models based on accuracy, loss, and their ability to handle the nuances of the claim-evidence relationships. In addition to the technical work , I also contributed to the project report by detailing the classification by explaining the rationale behind model selection, and providing a analysis of the model performances.

Xuan Wang

As a member of the project team, my contributions were pivotal in the testing and refinement phases of the evidence retrieval component for our automated fact-checking system targeting climate change claims. I assisted in reviewing and debugging the final code, ensuring the robustness and accuracy of our retrieval algorithms. I also conducted extensive literature research to support our project's framework and methodologies, contributing valuable insights and contemporary practices to our approach. In our project report, I focused on introducing the model and meticulously cited relevant literature, ensuring our findings and methodologies were well-supported and accurately represented. During our project presentations, I effort in discussing the conclusions drawn from our work and proposed potential avenues for future enhancements, highlighting how these could further elevate the system's effectiveness in real-world applications.