Fine-Tuning Pre-Trained Transformers for Climate Claim Verification

Tsang Tsz Hin, Jason
The Chinese University of Hong Kong, 1155175030@link.cuhk.edu.hk

Abstract - **Misinformation and disinformation on the internet present a significant challenge in the context of climate change debate. The dissemination of false or misleading information can hinder public understanding and impede efforts to combat the growing issue of climate change. While social media platforms have implemented automatic fact-checking algorithms, existing models lack domain-specific training to effectively verify climate change-related information. As a remedy, a new fact-checking dataset is proposed that combines data from CLIMATE-FEVER with web-scraped information, resulting in a comprehensive dataset comprising 8,115 annotated claim-evidence pairs. The improved dataset is used to fine-tune a variety of pre-trained transformers for climate claim verification tasks. The best model, RoBERTa, achieved an accuracy of 0.7288 and F1-score of 0.7229, improving upon previously reported state-of-the-art (SoTA) F1-score of 0.7182.**

Keywords – Claim Verification; Climate Change; Transformer

INTRODUCTION

The growth of social media and the internet has encouraged the flow of information but has also become a breeding ground for misinformation and disinformation (Adams et al., 2023). A study by Ofcorm (2022) showed that one in three internet users is unaware that online content might be false or biased. Furthermore, it has been found that there is more misinformation on climate change, which poses threats to democratic values and the public understanding of various issues (Treen et al., 2020). Such misinformation can manipulate the understanding of climate change, polarizing the climate change debate (Cook, 2022). The spread of misinformation undermines public support and acts as a barrier to climate action (Maertens et al., 2020).

The topic of climate change fundamentally poses cognitive challenges to the public because it challenges many people's worldviews (Lewandowsky, 2021). The shift from the capitalist ideal of an unregulated free market to socialist climate mitigation interventions has brought about many oppositions. This has created rhetorical adversity where misinformation acts as a tool to distort the discussion of climate change (Lewandowsky, 2021).

Existing literature has provided mitigation methods to dispel misinformation and increase the public's climate change literacy. Research through experiments has shown that providing factual scientific evidence can help disconfirm bias and reduce climate change illiteracy (Ranney & Clark, 2016). Public education on this topic can also be an effective measure to debunk this misinformation and raise people's acceptance towards climate change (Lewandowsky, 2021). However, these education programs should not be the last resort to combat the problem of misinformation. The root of the problem still lies in the vast transmission of inaccurate or misleading claims on the internet and social media.

In recent years, due to the advances in Machine Learning and Natural Language Processing, researchers have been exploring the potential of automated fact-checking (Guo et al., 2021). The current algorithmic fact-checkers are not fully automated but rather act as tools to assist human fact-checkers. Currently, automatic fact-checkers are still in the early stages of development, with the ability to process a narrow range of simple verification tasks (Graves, 2018). With the rapid development of transformer-based models, many papers have explored this realm and demonstrated that transformer-based fact-checking models can be useful in evidence retrieval and claim classification (Elbassuoni, 2023; Soleimani et al., 2020).

Transformers have revolutionized the need to train a language model from scratch. These models are pre-trained with a large corpus of language data, capturing knowledge and encoding them into parameters (Han et al., 2021). The pre-trained model serves as the backbone for fine-tuning on specific tasks, making it computationally inexpensive and efficient. The focus of language model training has therefore shifted to collecting large and high-quality datasets for fine-tuning on different downstream tasks. In the fact-checking domain, one renowned paper is "FEVER: a large-scale dataset for Fact Extraction and Verification" by Thorne et al. (2018). The paper introduced a framework for creating a fact-checking dataset. This framework later inspired the introduction of CLIMATE-FEVER, a dataset for the Verification of Real-World Climate Claims (Diggelmann et al., 2020).

Previous research has experimented with fine-tuning pre-trained transformer models on the CLIMATE-FEVER dataset for climate claim verification. However, this dataset has the following flaws: 1) an imbalanced dataset, and 2) the absence of a pre-split dataset.

The CLIMATE-FEVER dataset suffers from imbalanced data, with few refuted claim-evidence pairs. This lowers the usefulness of the model, as debunking false climate claims is a crucial task in claim verification. To address this issue, this paper combines web-scraped data to rectify the imbalance problem and introduces a larger and balanced dataset.

The second problem with the dataset is that the author did not pre-split it, making it difficult to compare research results. To address this problem, this paper pre-splits the improved dataset and made it publicly available to facilitate future research. Additionally, this paper fine-tuned popular open-sourced pre-trained transformer models, including BERT, RoBERTa, T5, XLNet, and GPT-2, enabling direct comparisons between them.

The contributions of this paper can be summarized in threefold. First, an improved climate claim-evidence pairs dataset is presented. Second, a variety of pre-trained models are fine-tuned on the improved dataset, highlighting differences in their performance in fact-checking tasks. Third, to foster research efforts in this domain, the model, training code, and dataset are made publicly available on GitHub and Hugging Face.

RELATED WORK

To this day, most research on claim verification models is based on the FEVER dataset (Bekoulis et al., 2020). The FEVER (Fact Extraction and VERification) dataset consists of 185,445 claims artificially generated by altering claims extracted from Wikipedia (Thorne et al., 2018). The dataset is the largest general corpus for claim verification training, and various models have been built on this dataset.

Finetuning on FEVER dataset for claim verification task was described as a Natural Language Inference problem by Soleimani et al. (2020), in which a language model receives inputs from the claim as the premise and the corresponding evidence as the hypothesis. Many models have adopted this fine-tuning objective as foundation, introducing transformers with more complex architectures to achieve state-of-the-art results (Dominik & Elliott, 2020; Zhao et al., 2020).

However, the claim verification model trained on FEVER faces difficulties when verifying real-world climate claims. The authors of the CLIMATE-FEVER dataset utilized the claim verification model built by Hanselowski et al. (2018), which was originally trained on the FEVER dataset, to test its effectiveness in verifying real-world climate claims. The results showed a dissatisfactory accuracy score of 0.3878 and an F1-score of 0.3285 (Diggelmann et al., 2020). The low accuracy score is argued to be caused by the difference between the artificial nature of the FEVER dataset and the real-world nature of the CLIMATE-FEVER dataset. This has raised the need for building a model based on real-world climate claim verification.

Wang et al. (2021) were one of the earliest research teams to develop a fact-checking system for climate change claims. They introduced a semi-supervised training method to fine-tune RoBERTa using the CLIMATE-FEVER dataset and achieved a state-of-the-art F1 score of 0.7182. However, the fact-checking pipeline in this paper adopted evidence retrieval from the open internet instead of a static in-house corpus. As a result, comparing and replicating the model's performance becomes challenging.

Webersinke et al. (2021) introduced ClimateBert, using DistilRoBERTa as a starting point to fine-tune on climate-related data, with a focus on handling climate-related tasks. When evaluating its performance on the CLIMATE-FEVER dataset for fact-checking, ClimateBert outperformed Wang et al. (2021) with an F1 score of 0.757. However, it's worth noting that the paper simplified the claim verification task into a binary classification problem, categorizing claims and evidence as "Supports" or "Refutes" while excluding the "Not Enough Information" class. This simplification may have contributed to the favorable results obtained.

Vaghefi et al. (2022) adopted a similar approach and developed climateGPT-2, which pre-trained a climate-related corpus using the GPT-2 model. The model has been proven to enhance the baseline GPT-2 model in the claim verification task on CLIMATE-FEVER dataset, with an increase in the F1 score from 0.67 to 0.72. However, the paper did not specify the details of the training process and whether any data preprocessing steps were performed.

RESEARCH GAP

This paper aims to address the existing research gap by focusing on two main aspects: 1) a standardized dataset and 2) a direct comparison between models. Currently, there is no pre-split dataset for CLIMATE-FEVER. Previous research, therefore, does not share the same text corpus for the train-test split. Without a standardized dataset, model performance comparison becomes difficult. Moreover, there is a lack of research on which pre-trained model is the most suitable model to fine-tune for claim verification tasks. To address this problem, this paper places special emphasis on the standardized control of the fact-checking dataset, along with a pre-split text corpus. This enables a comprehensive comparison of various pre-trained transformer models in the context of claim verification.

DATASET OVERVIEW

This paper fine-tuned the pre-trained transformer based on two sources of data: 1) the CLIMATE-FEVER dataset and 2) web-scraped data from Climate Feedback.

The CLIMATE-FEVER dataset is published by Diggelmann et al. in 2020. The CLIMATE-FEVER dataset is a collection of 1,535 real-world claims related to climate change**,** each accompanied by 5 manually annotated evidence sentences, resulted in a database of 7,675 annotated claim-evidence pairs (Diggelmann et al., 2020).

The real-world climate claims are retrieved from the internet, sourced equally from scientifically-informed and climate change skeptics/deniers sources (Diggelmann et al., 2020). The claims are further filtered by climate scientists through majority voting, resulting in 1535 verifiable climate claims for the dataset (Diggelmann et al., 2020).

The CLIMATE-FEVER dataset differs significantly from the FEVER dataset in terms of claim complexity. CLIMATE-FEVER focuses on collecting real-world claims, which are often more detailed than the simple, self-contained claims found in the FEVER dataset. This makes CLIMATE-FEVER a more suitable dataset for training climate claim verification models, as most misinformation in the climate domain is based on real-world comments. For example, one of the climate claims in the FEVER dataset is:

“The Gray wolf is threatened by global warming.”

On the other hand, the CLIMATE-FEVER claims are shown to be more complex:

“Carbon emissions are declining in most rich nations and have been declining in Britain, Germany, and France since the mid-1970s.”

As for the corresponding evidence in the CLIMATE-FEVER dataset, it is automatically retrieved from Wikipedia for each given claim (Diggelmann et al., 2020). Each claim is paired with evidence and annotated as SUPPORTS, REFUTES, or NOT_ENOUGH_INFO, as presented in Table 1.

Claim:	The Great Barrier Reef is experiencing the most widespread bleaching ever recorded
Evidence #1	
Evidence Label:	SUPPORTS
Evidence Sentence:	
“A March 2016 report stated that coral bleaching was more widespread than previously thought, seriously affecting the northern parts of the reef as a result of warming ocean temperatures.”	
Evidence #2	
Evidence Label:	NOT_ENOUGH_INFO
Evidence Sentence:	
“The decline in the quality of water over the past 150 years (due to development) has contributed to coral bleaching, algal blooms, and pesticide pollution.”	

Table 1: Example of annotated claim and evidence pairs in CLIMATE-FEVER dataset

Another part of the dataset consists of web-scraped data from Climate Feedback. Climate Feedback is a fact-checking website that enables qualified scientists to review and verify climate claims. According to Climate Feedback, individuals must hold a PhD in the climate domain and have at least one published article in a scientific journal within the last five years to become one of their reviewers (About Us – Climate Feedback, 2023). The website has gained recognition as a highly respected and influential fact-checking tool for combating climate denial and misinformation (Nuccitelli, 2021).

Claim:	We do not know if CO2 is the cause of global warming
Verdict:	Inaccurate
Source:	Holman W. Jenkins, The Wall Street Journal, 3 Nov. 2023
Details:	
“Inaccurate: Evidence has allowed scientists to conclude, beyond reasonable doubt, that elevated atmospheric-CO2 from human emissions is the main driver of global warming. Scientists have quantified how much CO2 has strengthened the greenhouse effect by absorbing and radiating heat and this warming effect matches the observed global temperatures increase.”	
Key Take Away:	
“We know that CO2 causes global warming through the greenhouse effect based on overwhelming evidence from data collected over decades of investigation. There is international scientific consensus that elevated atmospheric-CO2 from human emissions is not just a cause of global warming, it is the leading cause of global warming.”	

Table 2: Example of a claim review extracted from Climate Feedback

Climate Feedback reviews each claim in detail, providing explanations and summarizations to verify the claims. For example, in Table 2, the review cited a claim from a Wall Street Journal article, fact-checking the statement "We do not know if CO2 is the cause of global warming." In the details section, the reviewer points out that the claim is inaccurate and provides evidence to refute it. The key takeaway section summarizes the claim review.

This paper considers the "Details" and "Key Takeaway" sections of Climate Feedback as evidence since they provide reasons for refuting or supporting the claims. The verdict is treated as the label for each claim-evidence pair, determining whether the claim is labeled as SUPPORTS or REFUTES. In comparison to CLIMATE-FEVER, Climate Feedback offers more comprehensive and detailed evidence. Although there may be slight differences in the nature of claim-evidence pairs between the two sources, it is believed to be beneficial to incorporate Climate Feedback's data into the claim verification model training. To accomplish this, web scraping was employed to extract and format the data in the format of CLIMATE-FEVER. The resulting structured data from web scraping is presented in Table 3.

Claim:	We do not know if CO2 is the cause of global warming
Evidence #1	
Evidence Label:	REFUTES
Evidence Sentence:	
“Evidence has allowed scientists to conclude, beyond reasonable doubt, that elevated atmospheric-CO2 from human emissions is the main driver of global warming. Scientists have quantified how much CO2 has strengthened the greenhouse effect by absorbing and radiating heat and this warming effect matches the observed global temperatures increase”	
Evidence #2	
Evidence Label:	REFUTES
Evidence Sentence:	
“We know that CO2 causes global warming through the greenhouse effect based on overwhelming evidence from data collected over decades of investigation. There is international scientific consensus that elevated atmospheric-CO2 from human emissions is not just a cause of global warming, it is the leading cause of global warming.”	

Table 3: Structured claim-evidence pairs from Climate Feedback

DATA PREPROCESSING

The CLIMATE-FEVER dataset consists of 1,535 claims, each of which is paired with 5 pieces of evidence, resulting in a total of 7,675 claim-evidence pairs. However, the dataset is found to be imbalanced as the evidence labels are dominated by "NOT_ENOUGH_INFO," with 4,930 occurrences, while the "SUPPORTS" and "REFUTES" labels only have 1,943 and 802 occurrences respectively. This imbalance is primarily attributed to the evidence retrieval process, as the evidence is algorithmically retrieved from Wikipedia. The limited source of evidence may have caused the retrieval process to retrieve unnecessary information, leading to the dominance of the "NOT_ENOUGH_INFO" class.

The Climate Feedback web-scraped dataset consists of 168 instances with 440 claim-evidence pairs. The dominant class in this dataset is "REFUTES," with 406 claim-evidence pairs, while the "SUPPORTS" label has 34 pairs. Adding this dataset can lead to an increase in the "REFUTES" class by over 50%, resulting in a more balanced dataset. Moreover, the increase in refuted claim-evidence labels can be benefit for training the model to differentiate between true and false information, leading to improved performance (Bekoulis et al., 2020).

For the train-test split, this paper uses a similar approach to Webersinke et al. (2021) and Vaghefi et al. (2022). The CLIMATE-FEVER dataset is split randomly, with 10% allocated for validation. The remaining dataset is divided into 10% for testing and 80% for training data. To maintain data integrity and facilitate comparison with the aforementioned studies, the Climate Feedback data was incorporated into the training data only after the train-test split. The resulting dataset has been made available on Hugging Face under the name "climate-fever-plus" for future research and replication purposes[1].

METHODOLOGY

The objective of this paper is to fine-tune transformers for classifying an evidence sentence and labeling it as SUPPORTS, REFUTES, or NOT_ENOUGH_INFO for a given claim. To achieve this, this paper adopts the training methodology from the works of Webersinke et al., 2021, and Wang et al., 2021 for easier comparison. For concatenating claims with evidence, a [SEP] token is used to separate them, and then feeding the concatenated sentence into various pre-trained transformer models with their corresponding model tokenizers.

To the best of current knowledge, no research has included different pre-trained transformers for a direct model comparison specifically on the CLIMATE-FEVER dataset, let alone this larger and more balanced dataset. To bridge this research gap and test which model can outperform in claim verification tasks, this paper tested a total of 5 different popular open-source pre-trained base models from Hugging Face: BERT, RoBERTa, T5, XLNet, and GPT-2. This not only allow a direct comparison between different models, but it also enables the comparison of the nature of transformers, determining whether a decoder-only, encoder-only, or decoder-encoder model performs best in this task.

Hyperparameter	Value
Training Batch Size	16
Eval Batch Size	32
Number of Epochs	10
Warmup Steps	500
Learning Rate	1e-5
Patience	10
Weight Decay	0.01
Optimizer	AdamW
Gradient Accumulation Steps	1
Mixed Precision Training	True
Learning Rate Scheduler	Linear

Table 4: Hyperparameters used for model training

[1] https://huggingface.co/datasets/Jasontth/climate_fever_plus

	Architecture	Number of Parameters	Pre-training Objective	Val Loss	Test Accuracy	Test F1
BERT-base	Encoder-Only	110M	Masked Language Modelling + Next Sentence Prediction	0.826	0.657	0.614
RoBERTa-base	Encoder-Only	125M	Masked Language Modelling	**0.752**	**0.729**	**0.723**
XLNet-base	Decoder-Only	110M	Permutation Language Model	0.773	0.662	0.668
GPT-2	Decoder-Only	117M	Causal Language Modeling	0.826	0.652	0.605
T5-base	Encoder-Decoder	223M	Text-to-Text	0.818	0.648	0.602

Table 5: Model performance evaluated based on average validation loss, test accuracy, and average F1 test score

For simplification, the fine-tuning process is standardized with the same hyperparameter, as detailed in Table 4. The model is initialized with pre-trained weights and fine-tuned using the AdamW optimizer and a linear learning rate scheduler to optimize the model's parameters. During training, the model is fine-tuned on the training dataset to minimize the cross-entropy loss. The training process continues until the maximum number of epochs is reached or early stopping is triggered. The complete training code is available on GitHub[2].

Performance Analysis

Table 5 summarizes the performance of the fine-tuned pretrained transformer model for label prediction. The best model, RoBERTa, achieved the lowest loss of 0.7521, the highest accuracy of 0.7288, and an average F1 score of 0.7229. This model is stored and made publicly available on Hugging Face[3].

The model outperformed the accuracy of 0.3878 and F1 score of 0.3285 reported in the paper by Diggelmann et al. (2020). The RoBERTa model in this paper showed approximately two-fold increase in these metrics, demonstrating outstanding performance.

Previously, the state-of-the-art (SoTA) research score was achieved by Wang et al. (2021) through their fact-checking pipeline, which demonstrated a SoTA F1 score of 0.7182 based on evaluation results. The model described in that paper did not utilize a static local corpus for evidence retrieval; instead, it retrieved evidence from the open internet. In contrast, this paper relies solely on an improved dataset and a static local corpus for claim-evidence pairs, resulting in an impressive F1 score of 0.7229 on the test data. Notably, the test data in this paper are unseen claim-evidence pairs, thereby increasing the reliability of the results. These findings indicate an improvement over previous research without accessing the open internet.

When comparing it to the Climate-GPT model by Vaghefi et al. (2022), their results showed a validation loss of 0.83 and an F1 score of 0.72. The RoBERTa model in this paper also outperforms their model with a lower validation loss and a similar F1 score. However, it is important to note that the average F1 score mentioned here is derived from the test dataset, not the validation set, which arguably indicates better performance.

Although the ClimateBert model by Webersinke et al. (2021) showed an F1 score as high as 0.757, it is arguable that this result was achieved by filtering instances labeled as "NOT_ENOUGH_INFO." This filtering simplified the complexity of the claim verification task, reducing it to a binary classification problem. Therefore, due to the difference in the dataset, direct model comparison is not possible.

Model Comparison

RoBERTa and BERT are both encoder-only transformer model which is featured with its bidirectional attention mechanism. RoBERTa is essentially an improved version of BERT. It is pre-trained with a larger dataset and a larger batch size, resulting in longer training (Liu et al., 2019). It has also eliminated the Next Sentence Prediction (NSP) pre-training objective and adopted a dynamic masking pattern for model training. Additionally, it has achieved state-of-the-art results on benchmarks such as GLUE, RACE, and SQuAD, demonstrating its improved performance over BERT. This improvement in performance is also evident in this paper, where the F1 score of RoBERTa surpassed that of the BERT model by over 10%.

[2] https://github.com/Jasontth/Climate-Claim-Verification
[3] https://huggingface.co/Jasontth/climate-fever-plus-RoBERTa

XLNet and GPT-2 are both decoder-only transformer models. The main difference between an encoder and a decoder-only transformer lies in the fact that the decoder transformer adopts a masked multi-head self-attention mechanism to train its autoregressive property (Yang et al., 2019; Radford et al., 2019). This autoregressive property typically works well in generative tasks. However, the task in this paper is claim verification, which might not be suitable for the autoregressive property of these transformers, leading to lower accuracy and F1 scores compared to the RoBERTa model.

T5 model is the only encoder-decoder model trained in this paper. The model is pre-trained to handle all NLP problems in a text-to-text format, where both the input and output are always in text format (Raffel et al., 2019). The encoder-decoder architecture allows it to combine the strengths of both the encoder and decoder, and this complexity enables it to understand the intricacies of input sequences and generate corresponding output sequences. However, in this paper, despite the T5-base model having the largest number of parameters, the results are not outstanding and underperformed the accuracy and F1 scores achieved by encoder-only and decoder-only transformer models.

IMPLICATIONS AND FUTURE RESEARCH

Recent developments of transformer models have largely focused on decoder-only language models such as GPT-4, Gemini, and Llama. In comparison, encoder-only transformers have received less attention. Many believe that decoder-only models will ultimately dominate due to their superior zero-shot generalization performance and the ability to fine-tune for any downstream task (T. J. Wang et al., 2022). However, the generality and creativity of decoder models have raised questions about their reliability and accuracy. Some research has shown that models like GPT may not be able to produce trustworthy output in certain domains (Bhattacharyya et al., 2023; Farhat et al., 2023; Alkaissi & McFarlane, 2023).

Claim verification, unlike text generation and question answering tasks, requires a more rigorous and strict process. A claim must be backed by reliable evidence without additional creativity or over-interpretation. Everything must be based on facts. This paper shows that in the claim verification task, RoBERTa, an encoder-only model, arguably has better performance. Future research should emphasize the ability of encoder-only models and explore whether they have better capabilities for claim verification tasks compared to decoder-only or encoder-decoder models.

The outstanding performance of the model described in this paper can also contribute to the larger dataset provided in this study. The paper web scraped claim-evidence pairs and improved the original CLIMATE-FEVER dataset, resulting in a more balanced class distribution. Specifically, the "REFUTES" class was increased by over 50%, making the dataset more useful for fact-checking. Increasing the number of refuted claims has been reviewed as beneficial in claim verification models, which aligns with the findings of this paper where the best model outperformed previous research (Bekoulis et al., 2020). Future research should focus on continuing to improve the dataset and collect more claim-evidence pairs. Data augmentation techniques should also be explored and adopted to expand and balance the dataset.

This paper primarily focuses on claim verification tasks, trained by fine-tuning based on customized claim-evidence pairs. However, the static nature of claim-evidence pairs is not realistic enough for full fact-checking automation. As depicted in Figure 1, given a claim, the document needs to be retrieved from the internet and filtered down to sentences as evidence. These pieces of evidence are then fed into a claim verification model for label classification. This paper is, therefore, a good starting point for future research in developing a fact-checking pipeline, as the claim verification model is readily available and can be incorporated into future studies.

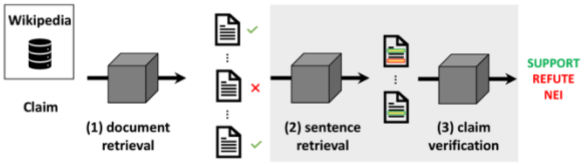

Fig 1: A three-step pipeline model for fact-checking task

CONCLUSION

Combating misinformation and disinformation in the climate debate is a challenging topic. NLP researchers and climate scientists desperately need reliable large language models (LLM) to automate the fact-checking process for climate claims. This paper contributes a larger and more balanced publicly available dataset that combines CLIMATE-FEVER data with web-scraped data. The dataset resulted in over a 50% increase in refuted claim-evidence pairs. A variety of pre-trained transformer models are fine-tuned on this improved dataset. The RoBERTa model is reported to be the best model in this paper, achieving a test accuracy of 0.7288 and an F1-score of 0.7229. This result improved upon the previously reported state-of-the-art (SoTA) F1 score of 0.7182. Moreover, this paper also discusses and argues that encoder-only models potentially represent the best architecture for claim verification tasks. Future research will focus on building a pipeline for fact-checking tasks, incorporating evidence retrieval with the current model presented in this paper. Ultimately, the goal of this paper is to encourage further research on algorithm development for fact-checking in the climate domain and to call for a joint effort in building a larger fact-checking dataset.

References

About us – Climate Feedback. (2023). Climate Feedback. Retrieved December 23, 2023, from https://climatefeedback.org/About/

Adams, Z., Osman, M., Bechlivanidis, C., & Meder, B. (2023). (Why) is misinformation a problem? *Perspectives on Psychological Science*, *18*(6), 1436–1463. https://doi.org/10.1177/17456916221141344

Alkaissi, H., & McFarlane, S. I. (2023). Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus*. https://doi.org/10.7759/cureus.35179

Bekoulis, G., Papagiannopoulou, C., & Deligiannis, N. (2020). A review on fact extraction and verification. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2010.03001

Bhattacharyya, M., Miller, V. M., Bhattacharyya, D., & Miller, L. E. (2023). High rates of fabricated and inaccurate references in ChatGPT-Generated medical content. *Cureus*. https://doi.org/10.7759/cureus.39238

Cook, J. (2022). Understanding and countering misinformation about climate change. In *IGI Global eBooks* (pp. 1633–1658). https://doi.org/10.4018/978-1-6684-3686-8.ch081

Diggelmann, T., Boyd-Graber, J., Bulian, J., Ciaramita, M., & Leippold, M. (2020). CLIMATE-FEVER: a dataset for verification of Real-World climate claims. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2012.00614

Dominik, S., & Elliott, A. (2020). e-fever: Explanations and summaries for automated fact checking. *Proceedings of the 2020 Truth and Trust Online (TTO 2020)*, 32–43. https://doi.org/10.3929/ethz-b-000453826

Elbassuoni, S. (2023, September 7). *FACTFORMER: A TRANSFORMER BASED FACT CHECKER*. http://hdl.handle.net/10938/24149

Farhat, F., Sohail, S. S., & Madsen, D. Ø. (2023a). How trustworthy is ChatGPT? The case of bibliometric analyses. *Cogent Engineering*, *10*(1). https://doi.org/10.1080/23311916.2023.2222988

Graves, D. B. (2018). Understanding the promise and limits of automated fact-checking. *The Reuters Institute*. https://doi.org/10.60625/risj-nqnx-bg89

Guo, Z., Schlichtkrull, M. S., & Vlachos, A. (2021). A survey on Automated Fact-Checking. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2108.11896

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Zhang, L., Han, W., Huang, M., Jin, Q., Lan, Y., Liu, Y., Liu, Z., Lu, Z., Qiu, X., Song, R., Tang, J., Wen, J., . . . Zhu, J. (2021). Pre-trained models: Past, present and future. *AI Open*, *2*, 225–250. https://doi.org/10.1016/j.aiopen.2021.08.002

Hanselowski, A., Zhang, H., Li, Z., Sorokin, D., Schiller, B., Schulz, C., & Gurevych, I. (2018). UKP-Athene: Multi-Sentence Textual Entailment for Claim Verification. *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. https://doi.org/10.18653/v1/w18-5516

Lewandowsky, S. (2021). Climate change disinformation and how to combat it. *Annual Review of Public Health*, *42*(1), 1–21. https://doi.org/10.1146/annurev-publhealth-090419-102409

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). ROBERTA: A robustly optimized BERT pretraining approach. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1907.11692

Maertens, R., Anseel, F., & Van Der Linden, S. (2020). Combatting climate change misinformation: Evidence for longevity of inoculation and consensus messaging effects. *Journal of Environmental Psychology*, *70*, 101455. https://doi.org/10.1016/j.jenvp.2020.101455

Mukherjee, M., & Hellendoorn, V. J. (2023). Stack Over-Flowing with Results: The Case for Domain-Specific Pre-Training Over One-Size-Fits-All Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2306.03268

Naseer, M., Windiatmaja, J. H., Asvial, M., & Sari, R. F. (2022). ROBERTAENS: Deep Bidirectional Encoder Ensemble Model for Fact Verification. *Big Data and Cognitive Computing*, *6*(2), 33. https://doi.org/10.3390/bdcc6020033

Nuccitelli, D. (2021, August 25). New study uncovers the “keystone domino” strategy of climate denial. *The Guardian*. https://www.theguardian.com/environment/climate-consensus-97-per-cent/2017/nov/29/new-study-uncovers-the-keystone-domino-strategy-of-climate-denial

Ofcom. (2022, March 29). *The genuine article? One in three internet users fail to question misinformation*. Retrieved December 13, 2023, from https://www.ofcom.org.uk/news-centre/2022/one-in-three-internet-users-fail-to-question-misinformation

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8).

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1910.10683

Ranney, M., & Clark, D. (2016). Climate change Conceptual change: Scientific information can transform attitudes. *Topics in Cognitive Science*, *8*(1), 49–75. https://doi.org/10.1111/tops.12187

Soleimani, A., Monz, C., & Worring, M. (2020). BERT for evidence retrieval and claim verification. In *Lecture Notes in Computer Science* (pp. 359–366). https://doi.org/10.1007/978-3-030-45442-5_45

Thorne, J. H., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: a large-scale dataset for Fact Extraction and VERification. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1803.05355

Treen, K. M. D., Williams, H. T. P., & O’Neill, S. (2020). Online misinformation about climate change. *WIREs Climate Change*, *11*(5). https://doi.org/10.1002/wcc.665

Vaghefi, S., Muccione, V., Huggel, C., Khashehchi, H., & Leippold, M. (2022, December 9). *Deep Climate Change: A Dataset and Adaptive domain pre-trained Language Models for Climate Change Related Tasks.* Climate Change AI. https://www.climatechange.ai/papers/neurips2022/27

Wang, G., Chillrud, L., & McKeown, K. (2021). Evidence based Automatic Fact-Checking for Climate Change Misinformation. *International Workshop on Social Sensing on the International AAAI Conference on Web and Social Media*.

Wang, T. J., Roberts, A., Hesslow, D., Scao, T. L., Chung, H. W., Beltagy, I., Launay, J., & Raffel, C. (2022). What language model architecture and pretraining objective work best for Zero-Shot generalization? *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2204.05832

Webersinke, N., Kraus, M., Bingler, J. A., & Leippold, M. (2021). ClimateBert: A pretrained language model for Climate-Related Text. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2110.12010

Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., & Le, Q. V. (2019). XLNET: Generalized Autoregressive Pretraining for Language Understanding. *arXiv (Cornell University), 32,* 5753–5763. https://arxiv.org/pdf/1906.08237

Zhao, C., Xiong, C., Rosset, C., Song, X., Bennett, P., & Tiwary, S. (2020). Transformer-XH: Multi-Evidence Reasoning with eXtra Hop Attention. *International Conference on Learning Representations*. https://www.openreview.net/pdf?id=r1eIiCNYwS