Advancing Multimodal Fact-Checking Against Climate Misinformation: A Benchmark Dataset and Comparison of Lightweight Models

Omar El Baf, Quentin Senatore, Amira Mouakher, Laure Berti-Équille

UMR 228 Espace-Dev, Espace pour le Développement, Montpellier & Perpignan, France
IRD – Institut de Recherche pour le Développement, Marseille, France

Abstract

This paper proposes TIGER, a high-quality benchmark dataset to better evaluate multimodal fact-checking models that verify the veracity of claims combining texts and images on topics related to climate change. Unlike previous datasets, which are very unbalanced, do not focus on climate misinformation, and are often unimodal, TIGER includes curated claims and scripts to augment the dataset with information extracted from IPCC reports, claims generated by ChatGPT and related web-scraped images. We also propose M4FC, a set of lightweight MLP-based models with different textual and visual encoders and compare them against other ML models. Our models outperform strong baselines such as Random Forests and Gradient Boosting by up to $+1.5\%$ in accuracy and $+1.7\%$ in F1-score on the TIGER dataset. The key advantage of MLP-based models for multimodal fact-checking is simplicity and flexibility, as they achieve competitive performance with lower computational cost and carbon footprint compared to heavier architectures.

1 Introduction

Climate misinformation spans a wide range of subtopics, from scientific denialism and media manipulation to economic myths and distortions of climate justice, often blending complex narratives that are politically charged, emotionally appealing, or deceptively simple. As this misinformation increasingly appears in multiple modalities such as text, images, videos, and memes [Akhtar et al., 2023, Biamby et al., 2021], detection and countermeasure have become significantly more urgent and technically challenging [Akhtar et al., 2023, Wang et al., 2024a]. Experts have even rated dissemination of misinformation and disinformation as the number one risk facing the world[1], a threat that is only increasing with the proliferation of generative AI [Jin et al., 2024, Zhang et al., 2024]. Multimodal misinformation is particularly problematic, as it is not only perceived by humans as more credible, but it also tends to spread faster than its text-only counterparts, amplifying its social impact [Akhtar et al., 2023]. However, current efforts to build robust AI classifiers capable of identifying fake information are hindered by the lack of high-quality benchmark datasets that reflect the full spectrum and subtlety of real-world misinformation, especially with respect to topics related to climate change [Wang et al., 2024a]. Traditionally, automated fact-checking research has focused primarily on textual data, with many existing surveys focusing on this single modality [Akhtar et al., 2023]. Although significant progress has been made in verifying textual claims against textual sources –as evidenced by large-scale datasets such as FEVER[2], which contains $185,445$ claims manually verified against Wikipedia, the real world challenge of misinformation frequently involves a complex

[1] https://www.weforum.org/stories/2024/01/ai-disinformation-global-risks/
[2] https://fever.ai/dataset/fever.html

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2025.

interplay of modalities [Wang and Shu, 2023, Diggelmann et al., 2020, Akhtar et al., 2023, Biamby et al., 2021, Alam et al., 2021].

Existing datasets are often limited in scope and in at least four different ways:

1) Many prior datasets for the verification of multimodal claims, such as FAKENEWSNET [Shu et al., 2020], COSMOS [Aneja et al., 2021], INFOSURGEON [Fung et al., 2021], FACTIFY [Mishra et al., 2022], FAUXTOGRAPHY [Zlatkova et al., 2019], and MOCHEG [Yao et al., 2023], focus primarily on multimodal content but lack the critical component of multihop reasoning, where a system must integrate and interpret multiple pieces of evidence from various sources to reach a final verdict [Wang et al., 2024a];

2) Similarly, while datasets such as HOVER focus on multi-hop textual claims, they do not incorporate the multimodal aspect [Jiang et al., 2020];

3) Even datasets such as MM-CLAIMS, which include tweets and images on topics such as COVID-19, Climate Change and Technology, primarily address claim check-worthiness identification rather than complete verification that requires evidence retrieval and veracity classification [Cheema et al., 2022]; and

4) The annotation process for these datasets can be subjective, often leading to the moderation of the agreement scores of multiple annotators that highlights the inherent difficulty of the task.

To address these gaps, we propose TIGER (Text-Image for fact-checkinG Evaluation Resource), a balanced high-quality text-image dataset to benchmark multimodal fact checking models. This resource aims to enable a more reliable and extensive evaluation of fact-checking models for multimodal evidence retrieval, veracity classification, and fine-grained stance assessment on climate change-related topics. In addition, we compared six deep learning models and propose M4FC (MLP Model for MultiModal Fact Checking), a lightweight model based on MLP for multimodal fact checking. Not only does our model show better performance in terms of accuracy and F1 in both unimodal and multimodal datasets, but it also has the lowest carbon footprint and training time. Using MLP blocks, it can be extended to several modalities other than text and images. The code, data and experimental results are available at: `https://github.com/LaureBerti/TIGER_M4FC`

2 Related Work

Several climate-related datasets have been proposed to support automated fact-checking and stance detection. Among text-only resources, CLIMATE-FEVER [Diggelmann et al., 2020] extends the FEVER framework by providing $1,535$ climate claims with Wikipedia-based evidence. While it remains a widely used benchmark, it suffers from class imbalance, dominated by 'Not Enough Info' labels, and includes a number of low-quality claims. More recently, CLIMATEX [Lacombe et al., 2023] introduced $8,094$ statements from IPCC reports[3], annotated with graded certainty levels, thereby enabling evaluation of models' ability to capture confidence in scientific assertions.

Beyond textual resources, multimodal corpora have also emerged. MM-CLAIMS [Cheema et al., 2022] contains around $86,000$ tweets (with $3,400$ labeled), demonstrating that combining textual and visual features improves claim detection. Building on this direction, Bai et al. [Bai et al., 2024] introduced a multimodal corpus of climate-related tweets with paired images, highlighting the challenges of reasoning over potentially contradictory cross-modal signals. Extending to temporal modalities, MULTICLIMATE [Wang et al., 2024b] provides video-based stance detection data, with transcripts and frames illustrating the added complexity of temporal dynamics.

Despite these advances, several limitations remain. Modality coverage is uneven: while text-image datasets are increasingly available, the integration of richer multimodal contexts with fine-grained reasoning requirements is still limited. Label granularity is also constrained, with few resources offering balanced veracity categories alongside structured misinformation taxonomies. In addition, data quality issues, such as excessive reliance on 'not Enough Info' labels or low-information claims, reduce the reliability of existing benchmarks, while limited interoperability across datasets hinders systematic evaluation. These challenges underscore the need for a balanced, extensible, and taxonomically rich multimodal dataset such as TIGER, and for lightweight but effective multimodal verification architectures such as M4FC.

[3]https://www.ipcc.ch/reports/

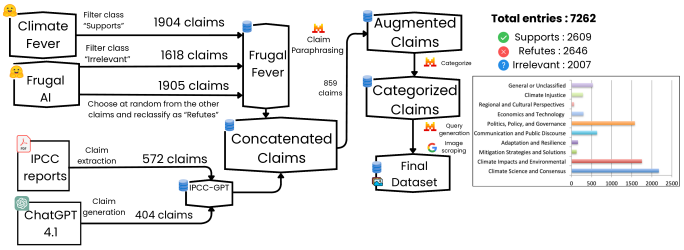

Figure 1: TIGER: Multimodal dataset creation workflow

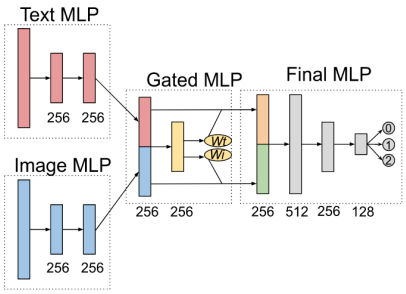

Figure 2: Architecture of M4FC

3 Our solutions: TIGER and M4FC

TIGER. The dataset construction process combines multiple sources to ensure diversity and balance between claim categories. It has been created by filtering low-quality entries from two existing unimodal datasets from the literature, namely CLIMATE FEVER and FRUGAL AI[4] whose claims have been curated and augmented with a richer taxonomy and additional claims extracted from a recent curated corpus of IPCC reports, and synthetic claims generated by ChatGPT, as illustrated in Figure 1. This dataset is extensible as it comes with scripts that can extend static and textual claims with new paraphrased claims and image pairing. Textual claims are categorized and augmented with image query-based scraping using Mistral.

To improve class balance, particularly between 'Refutes' and 'Irrelevant', we applied paraphrasing using the Mistral model, ensuring that each class contained at least $2,000$ claims. Subsequently, the Mistral model was used to categorize all claims into ten distinct thematic categories related to climate change. In addition, for each claim, we generated three search queries intended for image retrieval: a relevant query designed to support the claim, a neutral query that provided contextual imagery without directly reinforcing the claim, and an irrelevant query, often resulting in unrelated or humorous content. Finally, we performed large-scale image scraping from Google Images. For each claim, one of the three generated queries was selected with a predefined probability, entered into the Google Images search engine, and the corresponding image was downloaded. The resulting dataset has a total of $7,262$ claims with $2,609$ claims supporting true facts, $2,646$ claims refuting truth (fake news) and $2,007$ irrelevant claims. Scripts are provided to regularly extended the current version of the dataset with new sources of claims and images.

M4FC. Our model leverages MLPs for multimodal fact-checking to learn non-linear interactions across diverse feature spaces. When combining textual and visual signals, the MLP blocks, illustrated in Figure 2, can capture complex correlations without relying on heavy architectural assumptions, making the architecture lightweight, flexible, modular, and extensible. Its simplicity allows efficient training and quick inference while still leveraging multimodal embeddings from powerful upstream models. This makes our model a practical yet effective choice for integrating multimodal claims in fact-checking pipelines compared to the few existing computationally expensive models for fact checking such as VERACITY [Curtis et al., 2025] or MULTICLIMATE [Wang et al., 2024b].

4 Experimental Results

All experiments were conducted on a Dell Precision 7680 workstation with an Intel Core i9-13950HX processor (24 cores) and 16 GB of RAM. A five-fold cross-validation was applied, with results averaged over 50 iterations and rounded to the nearest hundredth.

As highlighted in Table 1, the comparative evaluation between the TIGER dataset and CLIMATE FEVER reveals clear differences in predictive performance and model stability. On CLIMATE FEVER, accuracy remains below 70% with notably low F1-scores, indicating that this dataset is more challenging and possibly affected by noise or weaker signal-to-label alignment. Interestingly, training and inference times are extremely short on this dataset for the tree-based and MLP models (training under 20 s, inference nearly instant), whereas M4FC models require around a minute.

[4]https://huggingface.co/frugal-ai-challenge

Table 1: Experimental results across datasets, encodings, and models.

Dataset	Encoding			Models	Accuracy	F1-score	Time (s)			Carbon Footprint (g)
	Encoder	Visual	Textual				Encoding	Training	Inference	
CLIMATE FEVER extended with images	Jina-Clip v2	EVA02-L14	Jina XLM-RoBERTa	MLP	63.5 ± 1.68	51.08 ± 2.16	17160	**7.42**	0.03	**0.51**
				Random Forest	69.38 ± 2.78	**56.07± 4.88**	17160	17.89	0.02	0.93
				Gradient Boosting	**70.17 ± 2.77**	54.91 ± 4.64	17160	12.93	**0.004**	1.21
	CLIP	ViT-B/16	CLIP Transformer	M4FC	64.87 ± 0.68	35.92 ± 6.93	10080	64.39	0.18	7.09
	CLIP	ResNet-50	CLIP Transformer	M4FC	65.26 ± 0.68	41.28 ± 7.61	**9420**	39.98	0.17	19.87
	CLIP	ResNet-50x4	CLIP Transformer	M4FC	65.15 ± 0.79	42.69 ± 6.82	19500	63.73	0.18	13.41
TIGER	Jina-Clip v2	EVA02-L14	Jina XLM-RoBERTa	MLP	83.14 ± 1.10	82.73 ± 1.16	19020	**9.97**	0.04	**0.56**
				Random Forest	83.33 ± 3.10	82.67 ± 3.27	19020	20.97	0.004	1.03
				Gradient Boosting	83.48 ± 0.78	82.91 ± 0.80	19020	38.06	0.006	1.34
	CLIP	ViT-B/16	CLIP Transformer	M4FC	84.41 ± 0.82	83.96 ± 0.82	11100	175.22	1.13	7.85
	CLIP	ResNet-50	CLIP Transformer	M4FC	**84.84 ± 0.77**	84.39 ± 0.82	**10440**	493.20	2.45	21.99
	CLIP	ResNet-50x4	CLIP Transformer	M4FC	**84.84 ± 0.78**	**84.40 ± 0.80**	22500	332.53	**1.15**	14.84

While this efficiency is appealing, it comes at the cost of predictive reliability: rapid training on CLIMATE FEVER appears to reflect underfitting, highlighting its limitations as a benchmark for robust multimodal verification.

In contrast, the TIGER dataset consistently yields substantially stronger and more stable results. Across both Jina-Clip v2 and CLIP-based encoders, accuracies exceed 83% with balanced F1-scores, demonstrating the robustness of the dataset in supporting reliable model training and evaluation. The higher training times observed with CLIP encoders reflect the richer cross-modal alignments they capture, but this cost translates directly into stronger performance and greater stability.

Another notable trend is the comparison between M4FC and classical tree-based models. While Random Forest and Gradient Boosting achieve competitive accuracy, particularly on CLIMATE FEVER, M4FC matches or outperforms them on TIGER, with higher stability and lower variance across cross-validation folds. M4FC also provides smoother learning curves and more consistent generalization, particularly when paired with CLIP encoders. This supports our claim that lightweight MLP blocks, when combined with strong pretrained representations, are both sufficient and effective for multimodal fact-checking.

Finally, the two best-performing configurations, namely CLIP with ResNet-50 and CLIP with ResNet-50x4, achieve identical predictive performance with an accuracy of 84.84% and F1-scores of approximately 84.4. However, while ResNet-50 requires shorter encoding time, it incurs substantially higher training time and carbon footprint (21.99 g) compared to ResNet-50x4 (14.84 g). This indicates that, when efficiency and sustainability are taken into account, ResNet-50x4 constitutes the more favorable option, offering equivalent predictive accuracy at a lower environmental cost.

5 Conclusion & Perspectives

We introduced TIGER, a balanced and extensible multimodal dataset for climate fact-checking, and M4FC, a family of lightweight MLP-based models that achieve strong and stable performance while remaining computationally efficient. Our results show that TIGER enables more reliable benchmarking than existing datasets, and that simple architectures can serve as effective baselines for multimodal misinformation detection. By releasing both resources, we aim to foster a shared benchmarking culture that encourages collaboration and accelerates progress in multimodal fact-checking against climate misinformation.

References

Mubashara Akhtar, Michael Schlichtkrull, Zhijiang Guo, Oana Cocarascu, Elena Simperl, and Andreas Vlachos. Multimodal automated fact-checking: A survey. *arXiv preprint arXiv:2305.13507*, 2023.

Giscard Biamby, Grace Luo, Trevor Darrell, and Anna Rohrbach. Twitter-comms: Detecting climate, covid, and military multimodal misinformation. *arXiv preprint arXiv:2112.08594*, 2021.

Haoran Wang, Aman Rangapur, Xiongxiao Xu, Yueqing Liang, Haroon Gharwi, Carl Yang, and Kai Shu. Piecing it all together: Verifying multi-hop multimodal claims. *arXiv preprint arXiv:2411.09547*, 2024a.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. Agentreview: Exploring peer review dynamics with llm agents. *arXiv preprint arXiv:2406.12708*, 2024.

Qihui Zhang, Chujie Gao, Dongping Chen, Yue Huang, Yixin Huang, Zhenyang Sun, Shilin Zhang, Weiye Li, Zhengyan Fu, Yao Wan, et al. Llm-as-a-coauthor: Can mixed human-written and machine-generated text be detected? *arXiv preprint arXiv:2401.05952*, 2024.

Haoran Wang and Kai Shu. Explainable claim verification via knowledge-grounded reasoning with large language models. *arXiv preprint arXiv:2310.05253*, 2023.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*, 2020.

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. A survey on multimodal disinformation detection. *arXiv preprint arXiv:2103.12541*, 2021.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188, 2020.

Shivangi Aneja, Chris Bregler, and Matthias Nießner. COSMOS: Catching out-of-context misinformation with self-supervised learning. *arXiv preprint arXiv:2101.06278*, 2021.

Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen McKeown, Mohit Bansal, and Avi Sil. InfoSurgeon: Cross-media fine-grained information consistency checking for fake news detection. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1683–1698, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.133.

Shreyash Mishra, S Suryavardan, Amrit Bhaskar, Parul Chopra, Aishwarya N Reganti, Parth Patwa, Amitava Das, Tanmoy Chakraborty, Amit P Sheth, Asif Ekbal, et al. FACTIFY: A multi-modal fact verification dataset. In *Proceedings of the DE-FACTIFY@ AAAI*, 2022.

Dimitrina Zlatkova, Preslav Nakov, and Ivan Koychev. Fact-checking meets fauxtography: Verifying claims about images. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2099–2108, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1216. URL `https://aclanthology.org/D19-1216/`.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2733–2743, 2023.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. HoVer: A dataset for many-hop fact extraction and claim verification. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.309. URL `https://aclanthology.org/2020.findings-emnlp.309/`.

Gullal S Cheema, Sherzod Hakimov, Abdul Sittar, Eric Müller-Budack, Christian Otto, and Ralph Ewerth. MM-claims: A dataset for multimodal claim detection in social media. *arXiv preprint arXiv:2205.01989*, 2022.

Romain Lacombe, Kerrie Wu, and Eddie Dilworth. ClimateX: Do llms accurately assess human expert confidence in climate statements? *arXiv preprint arXiv:2311.17107*, 2023.

Nan Bai, Ricardo da Silva Torres, Anna Fensel, Tamara Metze, and Art Dewulf. Inferring climate change stances from multimodal tweets. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2467–2471, 2024.

Jiawen Wang, Longfei Zuo, Siyao Peng, and Barbara Plank. MultiClimate: Multimodal stance detection on climate change videos. In Daryna Dementieva, Oana Ignat, Zhijing Jin, Rada Mihalcea, Giorgio Piatti, Joel Tetreault, Steven Wilson, and Jieyu Zhao, editors, *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 315–326, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.nlp4pi-1.27. URL https://aclanthology.org/2024.nlp4pi-1.27/.

Taylor Lynn Curtis, Maximilian Puelma Touzel, William Garneau, Manon Gruaz, Mike Pinder, Li Wei Wang, Sukanya Krishna, Luda Cohen, Jean-François Godbout, Reihaneh Rabbany, et al. Veracity: An open-source ai fact-checking system. *arXiv preprint arXiv:2506.15794*, 2025.