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

Omar El Baf, Quentin Senatore, Amira Mouakher and Laure Berti-Equille

IRD - Institut de recherche pour le développement, Marseille, France
UMR 228 Espace-Dev, Espace pour le développement, Perpignan, France

6 December 2025

Why Misinformation Matters?

"People are more likely to believe findings from a neuroscience study when the report is paired with a coloured image of a brain … the images provide a physical basis for thinking."
(McCabe & Castel, 2007)

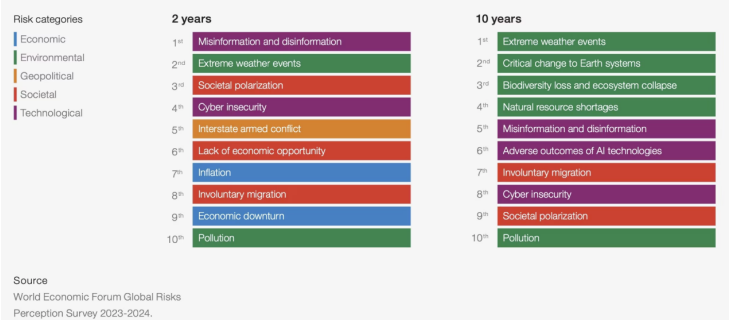

Figure 1: The proliferation of misinformation and disinformation

Source: https://www.weforum.org/stories/2024/01/ai-disinformation-global-risks/

Challenges in Multimodal Fact-Checking

- AI systems struggle to detect climate misinformation due to the lack of diverse, high-quality multimodal benchmark datasets.

Limitations of existing datasets

- **Missing multi-hop reasoning:** Focus on multimodal content but lack multi-step evidence integration
- **No multimodal integration:** Support reasoning but only for textual data
- **Limited verification scope:** Address check-worthiness, not full claim verification
- **Subjective annotation:** Inconsistent labeling due to annotator perceptive disagreement

⇒ Hard to evaluate or compare models fairly for multimodal climate fact-checking

Climate Fact-Checking Datasets

Dataset	Modality	Size	Key Features / Notes
CLIMATE-FEVER (2020)	Text	1,535 claims	Extends FEVER; uses Wikipedia evidence; suffers from class imbalance and low-quality claims.
CLIMATEX (2023)	Text	8,094 statements	Extracted from IPCC reports; annotated with graded certainty; evaluates model confidence in scientific assertions.
MM-CLAIMS (2022)	Text + Image	86k tweets (3.4k labeled)	Combines textual and visual features; demonstrates improved claim detection.
Bai et al. (2024)	Text + Image	49,316 tweets	Climate tweets with paired images; explores reasoning over contradictory cross-modal signals.
MULTICLIMATE (2024)	Video + Text	4,209 (from 100 videos)	Video-based stance detection; includes transcripts and frames; introduces temporal reasoning challenges.

Table 1: Summary of existing climate-related fact-checking datasets

Critical Gaps

- **Limited reasoning depth:** Few datasets require multi-hop or cross-modal inference.
- **Quality issues:** Imbalanced labels and noisy or weakly verified claims.
- **Lack of scale and diversity:** Most datasets remain narrow in topic or platform coverage.
- **Few temporal datasets:** Limited resources for video-based or evolving misinformation.

⇒ Need for comprehensive **multimodal climate datasets**
⇒ Demand for **efficient models** capable of robust multimodal fact-checking

The TIGER Dataset

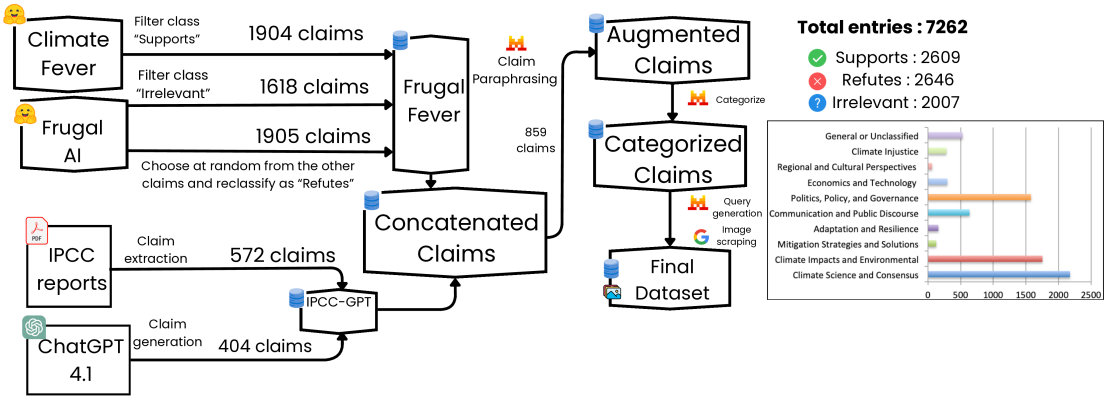

Figure 2: TIGER: Multimodal dataset creation workflow

The M4FC Model

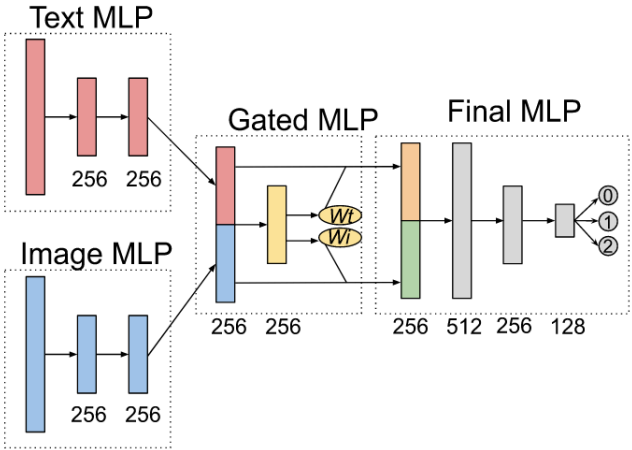

Figure 3: Architecture of M4FC

Results on CLIMATE FEVER (extended with images)

Encoding			Model	Accuracy	F1-score	Time (s)			CO_2 (g)
Encoder	Visual	Textual				Encoding	Training	Inference	
Jina-Clip v2	EVA02-L14	Jina XLM-RoBERTa	MLP	63.50 ± 1.68	51.08 ± 2.16	17 160	**7.42**	0.03	**0.51**
			Random Forest	69.38 ± 2.78	**56.07** ± **4.88**	17 160	17.89	0.02	0.93
			Gradient Boosting	**70.17** ± **2.77**	54.91 ± 4.64	17 160	12.93	**0.004**	1.21
CLIP	ViT-B/16	CLIP Transformer	M4FC	64.87 ± 0.68	35.92 ± 6.93	10 080	64.39	0.18	7.09
CLIP	ResNet-50	CLIP Transformer	M4FC	65.26 ± 0.68	41.28 ± 7.61	**9 420**	39.98	0.17	19.87
CLIP	ResNet-50x4	CLIP Transformer	M4FC	65.15 ± 0.79	42.69 ± 6.82	19 500	63.73	0.18	13.41

- Accuracy remains below 70%, with low F1-scores ⇒ **dataset is challenging and noisy**.
- Tree-based and MLP models train extremely fast (< 20s), but likely **underfit**.
- M4FC requires longer training (≈1 min) but shows higher stability.
- Indicates weaker signal-to-label alignment in CLIMATE FEVER.

Results on TIGER

Encoding			Model	Accuracy	F1-score	Time (s)			CO_2 (g)
Encoder	Visual	Textual				Encoding	Training	Inference	
Jina-Clip v2	EVA02-L14	Jina XLM-RoBERTa	MLP	83.14 ± 1.10	82.73 ± 1.16	19 020	**9.97**	0.04	**0.56**
			Random Forest	83.33 ± 3.10	82.67 ± 3.27	19 020	20.97	0.004	1.03
			Gradient Boosting	83.48 ± 0.78	82.91 ± 0.80	19 020	38.06	0.006	1.34
CLIP	ViT-B/16	CLIP Transformer	M4FC	84.41 ± 0.82	83.96 ± 0.82	11 100	175.22	1.13	7.85
CLIP	ResNet-50	CLIP Transformer	M4FC	**84.84 ± 0.77**	84.39 ± 0.82	**10 440**	493.20	2.45	21.99
CLIP	ResNet-50x4	CLIP Transformer	M4FC	**84.84 ± 0.78**	**84.40 ± 0.80**	22 500	332.53	**1.15**	14.84

- All models achieve strong, stable performance (>83% accuracy, balanced F1-scores).
- Higher training times for CLIP encoders reflect richer cross-modal alignments.
- M4FC matches or outperforms tree-based models with smoother generalization.
- ResNet-50x4 achieves same accuracy as ResNet-50 but with **lower carbon footprint**.

Conclusion

- TIGER: A balanced, extensible multimodal dataset for climate fact-checking.
- M4FC: Lightweight MLP-based models with strong, stable, and efficient performance.
- TIGER enables more reliable benchmarking than existing datasets.
- Simple architectures can be effective baselines for multimodal misinformation detection.
- Released both resources to promote collaboration and accelerate progress in climate fact-checking.

Thank you for your attention !

https://github.com/LaureBerti/TIGER_M4FC