Unveiling Fake News with Adversarial Arguments Generated by Multimodal Large Language Models

Xiaofan Zheng
Xi'an Jiaotong University
zxf_xjtu@stu.xjtu.edu.cn

Minnan Luo*
Xi'an Jiaotong University
minnluo@xjtu.edu.cn

Xinghao Wang
Xi'an Jiaotong University
370300626@stu.xjtu.edu.cn

Abstract

In the era of social media, the proliferation of fake news has created an urgent need for more effective detection methods, particularly for multimodal content. The task of identifying fake news is highly challenging, as it requires broad background knowledge and understanding across various domains. Existing detection methods primarily rely on neural networks to learn latent feature representations, resulting in black-box classifications with limited real-world understanding. To address these limitations, we propose a novel approach that leverages Multimodal Large Language Models (MLLMs) for fake news detection. Our method introduces adversarial reasoning through debates from opposing perspectives. By harnessing the powerful capabilities of MLLMs in text generation and cross-modal reasoning, we guide these models to engage in multimodal debates, generating adversarial arguments based on contradictory evidence from both sides of the issue. We then utilize these arguments to learn reasonable thinking patterns, enabling better multimodal fusion and fine-tuning. This process effectively positions our model as a debate referee for adversarial inference. Extensive experiments conducted on four fake news detection datasets demonstrate that our proposed method significantly outperforms state-of-the-art approaches[1].

1 Introduction

The rise of social media has led to a proliferation of fake news, posing harmful effects on individuals and society (Fisher et al., 2016; Vosoughi et al., 2018). This is especially concerning when images and text are strategically combined to create seemingly credible but false information (Naeem and Bhatti, 2020). Relying solely on investigative journalism to detect and expose fake news manually is not a viable solution—this approach is both labor-intensive and time-consuming, which limits the scale of coverage and introduces delays in debunking misinformation. Therefore, there is a pressing need to develop automated methods capable of detecting fake news across multiple modalities in a timely manner (Tasnim et al., 2020). This will not only address emerging trends in misinformation but also enhance the accuracy and efficiency of rumor identification (Roth, 2022).

Previous studies have attempted to use pre-trained visual-language models for fake news detection by adding task-specific classification layers (Zhang et al., 2021; Kaliyar et al., 2021). These methods frame the problem as an end-to-end classification task, merely identifying surface signals conveyed by the combination of text and images, and attempting to detect fake news by checking for consistency between the two (Xue et al., 2021). However, many well-crafted fake news articles exhibit strong associations between images and text, and current models often fail to detect these (Shu et al., 2017). In the face of this fake news, leveraging the vast knowledge humans have acquired in real life and their critical thinking and vigilance towards biased, subjective, and inflammatory content in the news is a better approach to combating fake news. All of this relies on a deep understanding of fake news (Yang et al., 2022; Wang et al., 2024).

Inspired by the success of Multimodal Large Language Models (MLLMs) in combining reasoning with background knowledge at a cognitive level (Li et al., 2024), we aim to leverage MLLMs' capabilities in understanding and reasoning (Liu et al., 2024a). However, we observed that directly using the analysis generated by MLLMs could introduce bias into the model's predictions (Lin et al., 2024; Hu et al., 2024), preventing comprehensive thinking. To address this limitation, we propose

* Corresponding author: Minnan Luo, School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China.

[1]Our code is available at: https://github.com/qingpingwan/AAR

Proceedings of the 31st International Conference on Computational Linguistics, pages 7862–7869
January 19–24, 2025.

a new approach: the Adversarial Arguments Reasoning (AAR) Model. This model leverages multimodal adversarial reasoning knowledge extracted from MLLMs to assist in fake news detection.

In contrast to these traditional methods that rely solely on identifying superficial harmful signals, our approach involves creating logical reasoning between textual and visual information, simulating human-level thinking about the news, and incorporating necessary background knowledge. Unlike recognition-based detection models, we believe that a reasoning-driven method that integrates these elements will significantly improve fake news detection (Qi et al., 2024).

Our approach involves two key phases: 1). Adversarial Argument Extraction: In the first phase, we fine-tune a smaller language model by integrating both language and visual features. This allows the model to learn how to fuse cross-modal image and text data. At the same time, we selectively extract adversarial multimodal reasoning knowledge from MLLMs. This equips our framework with cognitive reasoning capabilities, enabling it to predict the authenticity of news more effectively. 2). Authenticity Judgment: In the second phase, the fine-tuned small language model uses the extracted reasoning knowledge to make a final judgment on the authenticity of the news. By incorporating multimodal reasoning knowledge, we enhance the detection process, revealing the truth behind the news more effectively.

Our contributions can be summarized as follows:

- To the best of our knowledge, we are the first to explicitly utilize adversarial arguments generated by MLLMs to assist in multimodal fake news detection.
- We propose a novel reasoning-based framework that fine-tunes a small language model by integrating multimodal reasoning knowledge extracted from MLLMs. This promotes better multimodal fusion and enables lightweight fine-tuning for fake news detection.
- Through extensive experiments on classical datasets, we demonstrate that the AAR model outperforms existing methods in terms of both accuracy and practicality.
- We have open-sourced the adversarial arguments generated by MLLMs. By incorporating these analyses, we have built an MLLMs-enhanced dataset, which serves as an open-source resource for further research.

2 Related Work

2.1 Multimodal Fake News Detection

Previous research on multimodal fake news detection mainly follows two approaches. The first is a content-based method, which simply utilizes the image and text information of the news (Wu et al., 2021), employing multimodal networks such as BERT, ViT, etc. (Dosovitskiy et al., 2021; Devlin et al., 2019), for classification. However, the downside of these models is their lack of understanding of the real world (Hu et al., 2023a), performing well only on the current datasets (Mu et al., 2024), and being limited to identifying surface signals in the news (Zhu et al., 2022). The second approach is based on social context, where additional information such as the news dissemination tree, comments, and the identity of the publisher is used to assist in the judgment (Ma et al., 2015; Cui et al., 2022). However, not all news articles can access such information, and the high latency of this method hinders the timely detection of fake news (Hu et al., 2022).

2.2 Multimodal Large Language Models

Recently, MLLMs have demonstrated outstanding capabilities across various downstream tasks (Li et al., 2024). However, the enormous training costs and dataset limitations prevent MLLMs from being trained on specific tasks (Xuan et al., 2024), limiting their application in the domain of fake news detection (Liu et al., 2024a). At the same time, some experimental studies suggest that there exists a gap between MLLMs and fine-tuned smaller models (Hu et al., 2024), but they also point out the tremendous potential of MLLMs in fake news detection (Liu et al., 2024b). Our approach leverages MLLMs to simulate adversarial reasoning between opposing viewpoints in the debate process, extracting multimodal reasoning knowledge from these adversarial arguments to assist in fake news detection.

3 Methodology

For a given multimodal news sample, denoted as $\mathcal{N} = \{\mathcal{X}, \mathcal{U}, \mathcal{Z}\}$, where $\mathcal{X}$ represents the image associated with the text $\mathcal{U}$, and $\mathcal{Z}$ denotes the label of the news.

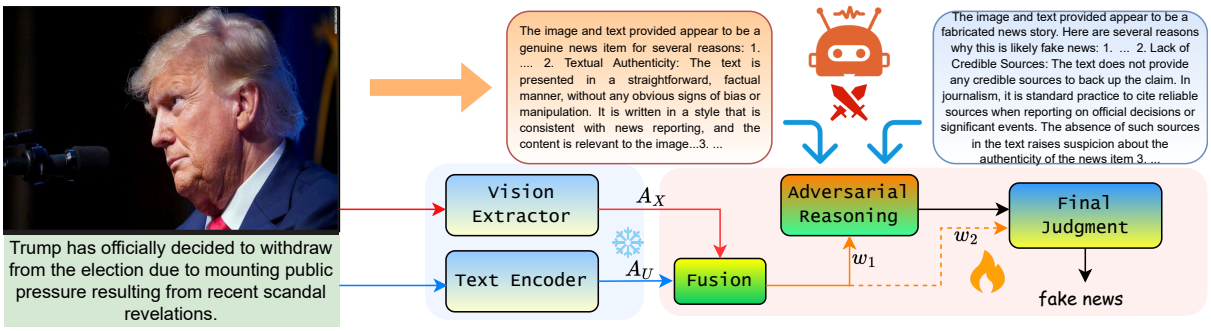

Figure 1: Illustration of AAR model.

Our core idea is to enhance the reasoning ability of the model in fake news detection by introducing adversarial arguments and dialectical thinking. This process involves analyzing the same news from different standpoints and seeking reasons within these conflicting viewpoints that can contribute to the final judgment. We leverage Multimodal Large Language Models (MLLMs) to generate reasons supporting both the real and fake stances of the news, which serve as additional knowledge input for the model.

Next, the image and text are separately fed into the frozen image and text encoders, which are derived from a pre-trained CLIP model (Xu et al., 2023). These encoders extract image features $\mathbf{A_x}$ and text features $\mathbf{A_u}$, mapping them into the same high-dimensional space. Since these encoders are pre-trained on large-scale datasets and aligned, no additional fine-tuning is required.

We then fuse the image and text features through a multi-head self-attention mechanism to achieve the fusion module and make the first prediction based on the fused tensor:

$$\begin{aligned} \mathbf{R_1} &= \text{self-attention}(\mathbf{A_x}) \\ \mathbf{R_2} &= \text{self-attention}(\mathbf{A_u}) \\ \mathbf{G} &= \alpha \cdot \mathbf{R_1} + \beta \cdot \mathbf{R_2} \\ \mathbf{z_1} &= \text{Judge}(\mathbf{W_2} \cdot \mathbf{G}) \end{aligned} \tag{1}$$

Here, α and β are trainable weighting parameters used to control the influence of image and text features during the fusion; $\mathbf{W_2}$ is a trainable weight matrix used to project the fused features $\mathbf{G}$; and Judge is a classifier implemented through a feed-forward neural network and activation function, outputting the first prediction $\mathbf{z_1}$.

Next, the adversarial arguments generated by MLLMs are concatenated together and transformed into a tensor representation $\mathbf{A_d}$ via the text encoder.

In the Adversarial Reasoning Module, the model uses a multi-head cross-attention mechanism for reasoning based on the original news features $\mathbf{G}$ and makes a second prediction:

$$\begin{aligned} \mathbf{L_g} &= \text{attention}(\mathbf{A_d}, \mathbf{W_1} \cdot \mathbf{G}, \mathbf{W_1} \cdot \mathbf{G}) \\ \mathbf{z_2} &= \text{Judge}(\mathbf{L_g}) \end{aligned} \tag{2}$$

Here, $\mathbf{W_1}$ is another trainable weight matrix used to project the fused features $\mathbf{G}$; the attention mechanism attention operates on the adversarial arguments $\mathbf{A_d}$ and the original fused news features $\mathbf{G}$, extracting relevant information from the adversarial arguments.

The final loss function is computed using the cross-entropy function, measuring the differences between the news label $\mathcal{Z}$ and the model's two predictions z_1 and z_2:

$$\begin{aligned} \mathcal{L}_g = \text{BCE}(\mathcal{Z}, \mathbf{z_1}) \quad \mathcal{L}_d = \text{BCE}(\mathcal{Z}, \mathbf{z_2}) \\ \mathcal{L} = \gamma_1 \cdot \mathcal{L}_g + \gamma_2 \cdot \mathcal{L}_d \end{aligned} \tag{3}$$

Where γ_1 and γ_2 are hyperparameters used to balance the two loss functions, and BCE is the binary cross-entropy loss function.

More details about prompt design can be found in Appendix §A.

4 Experiments

4.1 Experimental Setup

Our experiments are conducted on four real-world multimodal datasets: Weibo (Cao et al., 2017), Twitter (Boididou et al., 2015), Pheme (Kochkina et al., 2018), and MR2-en (Hu et al., 2023b). To validate the effectiveness of our model, we selected the following methods for comparison: **MLLMs** (Li et al., 2024): Directly using MLLMs to make final judgments based on adversarial arguments. We also compared our approach with

Models	Pheme		Weibo		Twitter		MR2-en	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
MLLMs (Li et al., 2024)	0.574	0.623	0.641	0.602	0.559	0.618	0.533	0.584
Bert (Devlin et al., 2019)	0.735	0.728	0.761	0.755	0.734	0.722	0.682	0.709
ViT (Dosovitskiy et al., 2021)	0.692	0.646	0.709	0.680	0.673	0.631	0.625	0.646
CAFE (Chen et al., 2022)	0.861	0.779	0.840	0.842	0.821	0.814	0.837	0.795
MMCAN (Zheng et al., 2022)	0.907	0.893	0.916	0.919	0.882	0.894	0.912	0.890
COOLANT (Wang et al., 2023)	0.915	0.910	0.911	0.908	0.902	0.895	0.901	0.886
AAR	**0.928**	**0.923**	**0.931**	**0.914**	**0.924**	**0.919**	**0.913**	**0.907**

Table 1: Performance comparison between AAR model and other models. The best results are in bold.

Models	Pheme	Weibo	Twitter	MR2-en
AAR	0.928	0.931	0.924	0.913
-w/o Fusion	0.743	0.752	0.696	0.703
-w/o AR	0.831	0.860	0.817	0.855
-w/o FI	0.824	0.848	0.839	0.831
-w/o FT	0.803	0.773	0.752	0.760
-w/o AD	0.852	0.846	0.854	0.872

Table 2: Ablation studies on our proposed model.

five advanced methods: **Bert** (Devlin et al., 2019), **ViT** (Dosovitskiy et al., 2021), **CAFE** (Chen et al., 2022), **COOLANT** (Wang et al., 2023), and **MMCAN** (Zheng et al., 2022).

Detailed information on the datasets and baseline methods are presented in Appendix §A.

4.2 Training Settings

Our LLVMs use LLaVA-v1.6-mistral-7b (also known as LLaVa-Next)[2]. The CLIP model uses MetaCLIP-b16[3] for English datasets and Alibaba's Chinese CLIP for Chinese datasets[4]. The MetaCLIP model is open-source and user-friendly, making it easier to integrate into existing frameworks. Due to LLaVA's weak Chinese generation capability, we set LLaVA to output in English even for Chinese datasets, then translate to Chinese using the DeepL API[5].

We use the AdamW optimizer with an initial learning rate of 2e-5, training for 40 epochs with early stopping to prevent overfitting. Our batch size is set to 32, with 8 attention heads (H = 8). The weight hyperparameters γ_1 and γ_2 are set to 0.5 and 0.5, respectively. To mitigate overfitting, we applied a dropout rate of 0.5 and a weight decay rate of 0.01. The model's prediction z_1 is only used during the training of the fusion module, while the final prediction output during testing is z_2.

All code is implemented in PyTorch and runs on NVIDIA V100 32G GPUs. Evaluation metrics include accuracy and F1 score.

4.3 Performance of AAR Model

Table 1 presents a detailed comparison of the performance of our proposed Adversarial Arguments Reasoning (AAR) model with state-of-the-art baseline methods across four datasets. By thoroughly analyzing the experimental results, we derive the following key insights: Our AAR model significantly outperforms baseline methods on all datasets, demonstrating the effectiveness and robustness of our proposed approach. Notably, the accuracy improvements on the Weibo and Twitter datasets exceed 2%, which is a substantial gain in the challenging domain of fake news detection. Compared to directly using Multimodal Large Language Models (MLLMs) for fake news detection, our lightweight model, trained additionally for the specific task, performs remarkably better. This finding underscores the necessity of training the AAR model for specific tasks and provides valuable methodological guidance for similar tasks in the future. In terms of modality, multimodal detection methods consistently outperform unimodal methods, highlighting the importance of integrating both image and text information for reasoning. It is noteworthy that text-based unimodal methods generally outperform image-based unimodal methods in our experiments. This observation indicates that, in the current datasets, textual information carries richer and more critical semantic cues, providing the model with more context and reasoning signals.

[2]https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf

[3]https://huggingface.co/facebook/metaclip-b16-fullcc2.5b

[4]https://huggingface.co/OFA-Sys/chinese-clip-vit-base-patch16

[5]https://www.deepl.com/en/pro-api

4.4 Ablation Study

In the ablation study, to further investigate the effectiveness of each component of the AAR model, we perform a quantitative analysis by removing each component and comparing it with the following variants: **w/o Fusion**: This variant removes the fusion module, relying solely on the output of MLLMs for prediction. **w/o AR**: This variant excludes the Adversarial Reasoning component, using only the fusion output for prediction. **w/o AD**: This variant omits the step where MLLMs generate adversarial arguments, instead of having MLLMs directly evaluate the news. **w/o FI**: This variant excludes the image content, meaning no visual information is incorporated during the fusion process. **w/o FT**: This variant excludes textual content, meaning the news text is not utilized in the fusion process. This structured comparison allows us to better understand the contribution of each component to the overall model performance.

The results show that the Fusion variant performs the worst, indicating that Adversarial Reasoning over MLLM-generated arguments requires the assistance of original image and text features from the news. The results of the FI and FT variants also confirm that both image and text modalities significantly contribute to the final judgment. The AR variant demonstrates the effectiveness of the arguments generated by MLLMs, whose rich knowledge and reasoning processes simulate the human process of thinking and judging the truthfulness of news. The AD variant proves that adversarial outputs with opposing viewpoints better leverage the inherent knowledge and capabilities of MLLMs compared to directly allowing MLLMs to analyze and evaluate the news without a prior stance.

5 Conclusion

Our proposed AAR model leverages the broad world knowledge and understanding capabilities of MLLMs, which have been trained on vast amounts of data. By generating adversarial arguments to analyze news content, it uncovers potential contradictions and uncertainties, assisting in determining the truthfulness of news. Extensive experiments on multiple datasets have shown that compared to previous methods, it achieves better accuracy and generalization, offering an insightful solution for utilizing the powerful capabilities of MLLMs in fake news detection systems.

6 Discussion and Limitations

Traditional fake news detection methods overly rely on the surface-level consistency between text and images. However, when faced with more sophisticated fake news, where the association between text and images is stronger, these methods often fall short (Mu et al., 2024).

More effective detection methods should emulate the way humans assess the truthfulness of news: not only focusing on surface-level information but also utilizing rich background knowledge and critically analyzing the news content (Hu et al., 2023a). This involves detecting biases, subjectivity, and sensationalism in the news. Additionally, multimodal reasoning is required to deeply integrate text and image information, followed by logical reasoning to make the final decision.

We acknowledge certain limitations in our research. For instance, our method involves multiple calls to MLLMs, which may result in higher costs compared to traditional methods. Furthermore, due to cost and API limitations, we only tested with the most accessible open-source model, LLaVA-v1.6-mistral-7b, and have not yet evaluated other MLLMs. Additionally, there is room for improvement in our model architecture. For example, the fusion module could be enhanced by exploring more advanced fusion methods.

Acknowledgment

This work was supported by the National Nature Science Foundation of China (No. 62192781, No. 62272374), the Natural Science Foundation of Shaanxi Province (2024JC-JCQN-62), the National Nature Science Foundation of China (No. 62202367, No. 62250009, No. 62137002), the Key Research and Development Project in Shaanxi Province No. 2022GXLH-01-03, Project of China Knowledge Center for Engineering Science and Technology, and Project of Chinese academy of engineering "The Online and Offline Mixed Educational Service System for 'The Belt and Road' Training in MOOC China". We would like to express our gratitude for the support of K. C. Wong Education Foundation. We also appreciate the reviewers and chairs for their constructive feedback.

Lastly, we would like to thank all LUD lab members for fostering a collaborative research environment.

References

Christina Boididou, Symeon Papadopoulos, Duc-Tien Dang-Nguyen, Giulia Boato, M. Riegler, S. Middleton, Andreas Petlund, and Yiannis Kompatsiaris. 2015. Verifying multimedia use at mediaeval 2016. In *MediaEval Benchmarking Initiative for Multimedia Evaluation*.

Qi Cao, Huawei Shen, Keting Cen, Wentao Ouyang, and Xueqi Cheng. 2017. Deephawkes: Bridging the gap between prediction and understanding of information cascades. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 1149–1158, New York, NY, USA. Association for Computing Machinery.

Yixuan Chen et al. 2022. Cross-modal ambiguity learning for multimodal fake news detection. In *The ACM Web Conference*, pages 2897–2905.

Jian Cui, Kwanwoo Kim, Seung Ho Na, and Seungwon Shin. 2022. Meta-path-based fake news detection leveraging multi-level social context information. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 325–334. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. ACL.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Marc Fisher, John Woodrow Cox, and Peter Hermann. 2016. Pizzagate: From rumor, to hashtag, to gunfire in dc. *The Washington Post*.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20):22105–22113.

Beizhe Hu, Qiang Sheng, Juan Cao, Yongchun Zhu, Danding Wang, Zhengjia Wang, and Zhiwei Jin. 2023a. Learn over past, evolve for future: Forecasting temporal trends for fake news detection. *Preprint*, arXiv:2306.14728.

Linmei Hu, Siqi Wei, Ziwang Zhao, and Bin Wu. 2022. Deep learning for fake news detection: A comprehensive survey. *AI Open*, 3:133–155.

Xuming Hu, Zhijiang Guo, Junzhe Chen, Lijie Wen, and Philip S. Yu. 2023b. Mr2: A benchmark for multimodal retrieval-augmented rumor detection in social media. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2021. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia tools and applications*, 80(8):11765–11788.

Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3402–3413, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024. Llava-next: Stronger llms supercharge multimodal capabilities in the wild.

Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. 2024. Towards explainable harmful meme detection through multimodal debate between large language models. *arXiv preprint arXiv:2401.13298*.

Hui Liu, Wenya Wang, Haoru Li, and Haoliang Li. 2024a. Teller: A trustworthy framework for explainable, generalizable and controllable fake news detection. *arXiv preprint arXiv:2402.07776*.

Qiang Liu, Xiang Tao, Junfei Wu, Shu Wu, and Liang Wang. 2024b. Can large language models detect rumors on social media? *arXiv preprint arXiv:2402.03916*.

Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. 2015. Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, pages 1751–1754.

Yida Mu, Xingyi Song, Kalina Bontcheva, and Nikolaos Aletras. 2024. Examining the limitations of computational rumor detection models trained on static datasets. *Preprint*, arXiv:2309.11576.

Salman Bin Naeem and Rubina Bhatti. 2020. The COVID-19 'infodemic': a new front for information professionals. *Health Information & Libraries Journal*, 37(3):233–239.

Peng Qi, Zehong Yan, Wynne Hsu, and Mong Li Lee. 2024. SNIFFER: Multimodal large language model for explainable out-of-context misinformation detection. *arXiv preprint arXiv:2403.03170*.

Yoel Roth. 2022. The vast majority of content we take action on for misinformation is identified proactively. `https://twitter.com/yoyoel/status/1483094057471524867`. Accessed: 2023-08-13.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19:22–36.

Samia Tasnim, Md Mahbub Hossain, and Hoimonty Mazumder. 2020. Impact of rumors and misinformation on covid-19 in social media. *Journal of Preventive Medicine and Public Health*, 53(3):171–174.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Bo Wang, Jing Ma, Hongzhan Lin, Zhiwei Yang, Ruichao Yang, Yuan Tian, and Yi Chang. 2024. Explainable fake news detection with large language model via defense among competing wisdom. In *Proceedings of the ACM on Web Conference 2024*, page 2452–2463, New York, NY, USA. Association for Computing Machinery.

Longzheng Wang, Chuang Zhang, Hongbo Xu, Yongxiu Xu, Xiaohan Xu, and Siqi Wang. 2023. Cross-modal contrastive learning for multimodal fake news detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23. ACM.

Yang Wu, Pengwei Zhan, Yunjian Zhang, LiMing Wang, and Zhen Xu. 2021. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the Association for Computational Linguistics*, pages 2560–2569.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, and Vasu Sharma. 2023. Demystifying clip data.

Keyang Xuan, Li Yi, Fan Yang, Ruochen Wu, Yi R Fung, and Heng Ji. 2024. LEMMA: Towards lvlm-enhanced multimodal misinformation detection with external knowledge augmentation. *arXiv preprint arXiv:2402.11943*.

Junxiao Xue et al. 2021. Detecting fake news by exploring the consistency of multimodal data. *Inf. Process. Manag.*, page 102610.

Zhiwei Yang, Jing Ma, Hechang Chen, Hongzhan Lin, Ziyang Luo, and Yi Chang. 2022. A coarse-to-fine cascaded evidence-distillation neural network for explainable fake news detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2608–2621, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining dual emotion for fake news detection. In *Proceedings of the web conference 2021*, pages 3465–3476. ACM.

Jiaqi Zheng et al. 2022. MFAN: multi-modal feature-enhanced attention networks for rumor detection. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 2413–2419.

Yongchun Zhu, Qiang Sheng, Juan Cao, Shuokai Li, Danding Wang, and Fuzhen Zhuang. 2022. Generalizing to the future: Mitigating entity bias in fake news detection. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

A Implementation Details

A.1 Datasets

The statistics of the datasets are shown in Table 3, divided into English and Chinese datasets. The English datasets are Pheme, Twitter, and MR2-en. The Pheme dataset is collected based on five breaking news events, with each event containing a set of posts, including a large amount of text and images. The Twitter dataset is released by MediaEval, aiming to detect fake multimedia content on social media. MR2-en is an English dataset crawled from Twitter and verified using the Google Fact Check Tools API. The Chinese dataset is Weibo, where all the fake news was crawled from May 2012 to January 2016, and it was collected and verified by Xinhua News Agency and Weibo. We preprocess and split the datasets following the methods of MCAN.

A.2 Baseline

- **BERT** (Devlin et al., 2019): Only uses textual content for classification.
- **ViT** (Dosovitskiy et al., 2021): Only uses visual information for classification.
- **CAFE** (Chen et al., 2022): Learns cross-modal fusion and adaptively aggregates multimodal and unimodal features for fake news detection.
- **COOLANT** (Wang et al., 2023): A cross-modal contrastive learning framework that learns cross-modal correlation through an attention-guided module to effectively detect fake news.
- **MMCAN** (Zheng et al., 2022): Designs a co-attention mechanism that is image-text matching-aware, capturing the alignment between images and text for better multimodal fusion in fake news detection.

Datasets	# of Real News	# of Fake News
Weibo	4779	4749
Twitter	6026	7898
Pheme	1972	3670
MR2-en	2318	1418

Table 3: The Statistics of Four Benchmark Datasets.

A.3 Adversarial Arguments Generation Prompts

"[INST] <image> news content: <text> Analyze the given news image and text to determine why this is likely genuine news. Provide specific analyses to support your conclusion that this news item is authentic.[/INST]"

"[INST] <image> news content: <text> Analyze the given news image and text to determine why this is likely fake news. Provide specific analyses to support your conclusion that this news item is not authentic.[/INST]"

B Ethics and Broader Impact

Our Adversarial Arguments Reasoning (AAR) model for fake news detection presents both promising opportunities and ethical challenges. The research offers substantial benefits, including reducing the spread of harmful fake news, providing a more nuanced approach to news authenticity, and creating a scalable alternative to manual fact-checking. However, we recognize critical ethical concerns such as the potential for algorithmic bias, the risk of misinterpreting complex information, and the possibility of technological solutions being misused for narrative control. To mitigate these risks, we commit to continuous model auditing, ensuring transparency in our decision-making processes, and promoting responsible AI development. Our approach is fundamentally designed not as an absolute arbiter of truth, but as a tool to enhance critical thinking and media literacy.