Discourse-Driven Detection of Multimodal Misinformation

Adrián Girón[1], Javier Huertas-Tato[1], and David Camacho[1]

Department of Computer Systems Engineering, Universidad Politécnica de Madrid, Madrid, Spain
{adrian.giron,javier.huertas.tato,david.camacho}@upm.es

Abstract. We introduce **LANTERN** (Large Vision-Language Model for Nested Thread Representation), a lightweight framework for detecting multimodal misinformation in online social networks. LANTERN incorporates the discourse surrounding a post by transforming nested comment threads into structured text that preserves reply depth, authorship, and engagement scores. This context is jointly processed with visual embeddings by a compact Large Vision-Language Model (LVLM) adapted with Low-Rank Adaptation (LoRA). On Fakeddit, LANTERN reaches 97.16% accuracy for binary classification, 97.05% for three classes, and 93.72% for six classes using only one quarter of the training data. The results show that structured social context and lightweight adaptation provide an effective and scalable approach to misinformation detection.

Keywords: Misinformation detection · Multimodal learning · Vision-language models · Social networks · LoRA

1 Introduction

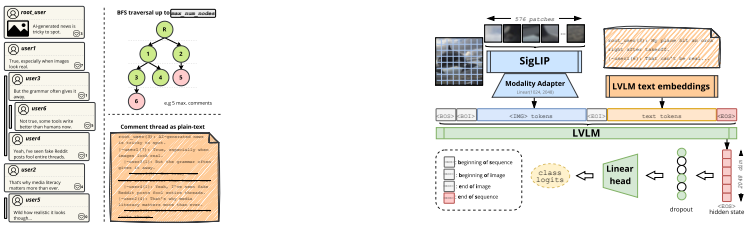
Misinformation in online social networks (OSNs) is rarely expressed through a single modality. Posts combine images, captions, and user interactions, whose reactions can reveal inconsistencies, skepticism, or patterns associated with deceptive content. Traditional systems commonly encode image and text separately and fuse their representations later [2], but do not fully exploit the conversational structure around a post.

This paper summarizes LANTERN [1], which directly uses a compact LVLM as a classifier. Its main hypothesis is that post veracity is encoded not only in the original content, but also in the surrounding discourse. The method converts nested comment trees into structured textual sequences, combines them with visual tokens in a unified LVLM input, and fine-tunes only lightweight LoRA adapters. The implementation is publicly available.[1]

2 LANTERN

Structured discourse representation. For each post, the caption is treated as the root of a comment tree and replies are recursively connected to their parents. A

[1] https://github.com/adgiz05/LVLM-fake-posts-detection



(a) Structured comment-tree representation.

(b) LVLM classification architecture.

Fig. 1. LANTERN preprocessing and classification pipeline.

breadth-first search prunes the tree to the first N nodes, prioritizing interactions close to the original post. Each comment is truncated to 20 words and represented as `author(score): text`. Tabulation encodes reply depth, allowing the model to infer interaction structure without a separate graph neural network. Figure 1(a) illustrates this representation, which integrates semantic and social signals in the same token stream.

Lightweight LVLM classification. LANTERN fine-tunes the open-source Janus-Pro-1B LVLM [3]. Its SigLIP visual encoder [4] represents each image with 576 tokens, followed by the caption and structured comments. The final end-of-sequence hidden state provides a multimodal representation of 2048 dimensions for a dropout layer and a linear classification head. Separate heads support binary, three-way, and six-way detection, as shown in Fig. 1(b).

Instead of updating all model parameters, LoRA adapters are trained while the backbone remains frozen. Training uses a batch size of 2, eight gradient-accumulation steps, AdamW, and at most five epochs on one NVIDIA RTX 3090 GPU with 24 GB of VRAM.

3 Evaluation Summary

The method is evaluated on Fakeddit [2], a Reddit dataset with nearly 700,000 multimodal posts containing images, captions, comments, and metadata. It provides binary, three-way, and six-way labels: true, satire, misleading content, imposter content, false connection, and manipulated content. To the authors' knowledge, it is the only public benchmark combining multimodal posts with structured comment threads.

Ablation findings. An ablation study on 15,000 training samples examines the number of comments, LoRA rank, and textual format. Table 1(a) shows that author identity improves the text-only baseline and engagement score provides a further gain. Five comments are generally sufficient; ten add limited value and may introduce redundant or off-topic replies. The selected configuration

Table 1. Ablation and scaled accuracy results on Fakeddit.

(a) Binary ablation using 15k samples.

Format	#C	Acc.	F1
Text	5	94.37	92.61
Author + text	5	94.72	93.05
Author + score + text	5	**95.04**	**93.47**

(b) Accuracy using 150k samples.

Approach	2-way	3-way	6-way
Nakamura et al.	89.09	88.90	85.88
Girón et al.	95.06	95.09	93.71
LANTERN	**97.16**	**97.05**	**93.72**

uses five comments, rank 8, and `author(score): text`. The modest rank is sufficient because the LVLM already contains useful cross-modal representations from pre-training.

Scaled performance and analysis. The best configuration is trained on 150,000 posts, approximately one quarter of the complete training set. Table 1(b) shows that LANTERN improves the strongest comparable binary and three-way results while matching the best six-way performance.

Training normally reaches its strongest checkpoint within two to four epochs. Runs with 15,000 examples take six to eight hours, whereas 150,000-sample runs require three to four days. Fine-tuning remains necessary: a preliminary few-shot generative experiment yields 14.64% invalid outputs and 39.52% accuracy. Treating the LVLM as a classifier avoids invalid free-form responses and provides stable predictions.

The fine-grained task reveals limitations. Satire and misleading content are frequently confused with true posts because subtle cases may require external knowledge, while manipulated content remains difficult when visual artifacts are weak. The perfect result for the imposter class should be investigated to determine whether it reflects robust separation or dataset-specific correlations.

4 Conclusions

LANTERN demonstrates that discourse is valuable for multimodal misinformation detection. Encoding a pruned comment tree as structured text enables a compact LVLM to reason jointly over images, captions, and user reactions without a complex fusion pipeline. Combined with LoRA, the method reaches state-of-the-art results with limited resources and a fraction of the available data. Future work may integrate graph-aware architectures and external grounding for subtle or knowledge-dependent categories.

Acknowledgments. Supported by MARTINI (PCI2022-134990-2; MCIN/AEI/10.13039/501100011033; EU NextGenerationEU/PRTR), IBERIFIER Plus (DIGITAL-2023-DEPLOY-04-EDMO-HUBS 101158511), and EMIF/MuseAI, managed by the Calouste Gulbenkian Foundation.

Disclosure of Interests. The authors have no competing interests relevant to this article.

References

1. Girón, A., Huertas-Tato, J., Camacho, D.: Discourse-driven detection of multimodal misinformation. IEEE Trans. Emerg. Topics Comput. Intell. **10**(2), 1389–1398 (2026)
2. Nakamura, K., Levy, S., Wang, W.Y.: Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In: Proc. LREC 2020, pp. 6149–6157 (2020)
3. Chen, X., et al.: Janus-Pro: Unified multimodal understanding and generation with data and model scaling. arXiv:2501.17811 (2025)
4. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. arXiv:2303.15343 (2023)
5. Girón, A., Huertas-Tato, J., Camacho, D.: Multimodal visio-lingual content analysis to detect fake content on Reddit. In: IDEAL 2024, pp. 143–154. Springer (2025)