

Challenges and Breakthroughs of LLMs in Detecting False Information on Social Media: From DELL to Human-Machine Collaboration

Yonghua Hong¹, Weiting Wang^{2, *}

¹ Software College of Henan University of Science and Technology, Henan University of Science and Technology, Henan, China

² Qingdao University Drake University Joint College, Qingdao University, Shandong, China

* Corresponding Author Email: wangweiting@qdu.edu.cn

Abstract. Due to the rise of large language models (LLMs) in today's society, the spread of false information has become even more rampant, and the detection and intervention of false information are quite challenging. This paper proposes to use LLMs to detect and intervene in false information on social media. More specifically, the DELL framework based on LLMs provides a systematic multi-stage governance solution. This framework maintains strong generalization performance and offers a clear - structured, operable, and interpretable technical path for applying large language models to false information governance. At the same time, it copes with complex scenarios by integrating multi-modal detection. We also explore the role of graph neural networks in rumor detection and combine graph neural networks with LLMs to form a joint framework for collaborative detection and intervention. This paper provides a comprehensive review of the detection and intervention of false information by LLMs, which can help other researchers better understand the current mainstream methods for detecting and intervening in false information.

Keywords: Large language model, false information, social media.

1. Introduction

With the rapid development of the Internet, social media has become the core carrier of information dissemination. However, with the widespread application of large language models, the generation and spread of multimodal false information have shown a trend of scale and concealment, seriously threatening the security of the public information ecosystem. Traditional detection methods are facing bottlenecks. Against this backdrop, research on the detection and intervention of false information by LLMs on social media is extremely important and urgently needed.

Generally speaking, prior work in analyzing and detecting false information using LLMs was only limited to relying on a single part. For example, relying on a single dataset or a single evaluation metric. In previous papers, most of the research focused on plain text, and only a few focused on multimodality and images. For example, one study employed the DELL framework to detect image-based misinformation [1]. The article illustrates how LLMs leverage their zero-shot capability to fact-check false information directly [2]. Papageorgiou et al also reports that employing this framework can boost the macro-F1 score by up to 16.8%. However, the data in this paper is relatively single, lacking research and analysis on multimodal data. In the identification of manipulators, the VigDet framework is used in Zhang Y et al [3]. to detect abnormal accounts. When detecting a single abnormal account, it has a high accuracy rate on Twitter. Meanwhile, the survey results cited in this paper use the model to discover cross-platform coordinated account groups for collaborative detection. In addition, most papers mainly focus on text-level evaluation, lacking analysis of other evaluation indicators. At the same time, LLMs also have dual roles and risks. It is necessary to conduct in-depth analysis and research on the dual roles of LLMs, which are both powerful tools for generating false information and potential detection tools. Jin Y et al [4]. points out that false information undermines trust in the digital ecosystem. Huang H et al [5]. emphasizes that false threat intelligence triggers data pollution attacks and false security alerts. In terms of modeling and quantifying the spread, a complex social network model is constructed to quantitatively evaluate the

spread dynamics, node influence and uncertainty of the content generated by LLM. Experiments usually construct causal models to identify factors that have a causal effect on users' susceptibility to false information and social manipulation. In response to the problem of the spread of false information generated by large - language models on social media, the MM - Soc benchmark is proposed to evaluate the performance of multimodal large models in social media tasks, covering 10 tasks [6]. Through testing 4 types of open - source MLLMs, it is found that the zero - shot performance is close to random guessing, far lower than that of fine - tuned models; increasing the model size can improve the instruction - following ability, but the task performance is unbalanced.

This article intends to provide a comprehensive review of LLMs in the detection and intervention of false information on social media. The remaining structure of the article includes Section 2: Method; Section 3: Discussion (Changes and Future Prospects); Section 4: Conclusion.

2. The Application of LLMs in the Direction and Intervention of False Information on Social Media

2.1. Introduction to Large Language Models

Large language models are a type of language model trained through self-supervised learning on massive amounts of text data. They are primarily used for natural language processing tasks, particularly excelling in text generation. As a manifestation of complex systems, LLMs model the statistical patterns and semantic structures of language to achieve multi-level understanding and generation of human language [7].

The core principle of LLMs is based on deep neural networks, especially the Transformer architecture. This architecture uses the self-attention mechanism to enable parallel processing of sequential data and capture long-range dependencies, effectively learning grammatical rules, semantic relationships, and contextual logic in language. Through pre-training on large-scale corpora—such as books, web pages, and dialogue texts—the models master vocabulary, syntax, and common expressions, allowing them to perform various downstream tasks like question answering, summarization, translation, and reasoning.

LLMs possess two key capabilities: First, language understanding, which involves recognizing the emotional tone, intent, and logical relationships in text, and even handling complex linguistic phenomena like metaphor and irony. Second, content generation, where they can produce coherent, reasonable, and diverse text outputs based on user instructions or contextual cues, including but not limited to writing articles, generating code, composing poetry, and even engaging in multimodal fusion generation.

Currently, influential large language models include OpenAI's ChatGPT, DeepSeek's model of the same name, and Anthropic's Claude. These models excel not only in general natural language tasks but are also widely used in specific applications such as intelligent assistants, content creation, machine translation, educational support, and data analysis. In social media environments, LLMs offer new technical approaches for identifying and intervening in misinformation due to their powerful text processing and generation capabilities.

2.2. False Information Detection Based on the DELL Framework

In addressing the challenge of false information on social media, the DELL framework (Detection–Explanation–Localization–Learning) based on large language models provides a systematic multi-stage governance solution. The core innovation of this framework lies in its modular, interpretable, and self-evolving capabilities, which not only identify false content but also clarify the basis for judgment, locate key dissemination paths, and enable continuous model optimization [8].

Specifically, the framework executes four stages in sequence: First, the Detection stage uses LLMs to perform joint semantic analysis of multimodal information such as text and images, identifying logical contradictions, emotional manipulation, or factual biases. Next, the Explanation stage

generates easy-to-understand explanations for why certain content is deemed false, improving user acceptance of model decisions. In the Localization stage, the system employs graph neural network technology to explore communities and key propagation nodes in social networks, identifying potential large-scale diffusion paths to provide targets for targeted intervention. Finally, the Learning stage incorporates user feedback and new samples to achieve continuous correction of model errors and domain adaptation.

Research shows that the DELL framework has achieved significant results on multiple public benchmarks, including the Twitter rumor database and the Weibo false news dataset, with detection accuracy exceeding 89%. It particularly maintains strong generalization performance in few-shot and cross-domain scenarios. This framework offers a clearly structured, actionable, and interpretable technical path for applying LLMs to false information governance.

2.3. Application in Multimodal False Information Detection

As false information production techniques evolve, single-modality text-based rumors have gradually been replaced by “multimodal” content that integrates text, images, audio, and even video. Such information is more deceptive—for example, a manipulated news screenshot or a real video with misleading commentary can serve as carriers of false information. The application of LLMs in this field is expanding from pure text analysis to multimodal understanding.

Taking multimodal large models like GPT-4V as an example, their core application lies in achieving joint semantic understanding and consistency verification across different modalities of information. Specifically, the model first uses a visual encoder to extract visual features from images while leveraging its powerful language capabilities to parse textual content. Then, through a cross-modal attention mechanism, the model deeply analyzes the correspondence between text and images, identifying several common fraud patterns: First, text-image mismatch, such as pairing an unrelated old image of a disaster with text about a recent event; second, image tampering, such as detecting illogical Photoshop traces, inconsistent lighting, or superimposed text; third, semantic misleading, where the image and text are each authentic but produce ambiguous or deceptive conclusions when combined [9].

Research indicates that this multimodal fusion detection method effectively handles complex scenarios that are difficult for pure text models to address. In a benchmark test targeting social media image-text rumors, the introduction of multimodal analysis improved detection accuracy by approximately 15%, significantly reducing the false positive rate.

2.4. Joint Propagation Intervention Based on Graph Neural Networks and LLMs

The harm of false information lies not only in its content but also in its rapid diffusion within social networks. Therefore, effective intervention must combine content analysis with propagation network analysis. A current mainstream innovation involves integrating LLMs, which deeply understand content, with Graph Neural Networks (GNNs), which excel at mining network structural features, to construct a collaborative detection and intervention framework.

In this joint framework, LLMs and GNNs each have distinct roles and complement each other. The LLM acts as a “content analyst,” responsible for in-depth interpretation of individual pieces of information, determining their authenticity, and generating analysis summaries or credibility scores. The GNN serves as a “network structure analyst,” abstracting the social platform into a graph structure where users are nodes, and relationships such as follows, reposts, and comments are edges. By analyzing information propagation paths, diffusion speed, and patterns, the GNN can accurately identify key nodes (such as influential bloggers or automated bot accounts) that promote rumor spread and vulnerable communities [10].

The integration point between the two is that the LLM generates high-quality feature representations for each piece of propagated content, serving as initial features for nodes or edges in the GNN. The GNN then uses these features for network-level embedding learning, ultimately outputting not only a judgment on the authenticity of specific information but also an assessment of

the overall propagation situation. Based on this, intervention strategies can be more precise—for example, prioritizing the delivery of debunking information generated by the LLM to followers of key propagation nodes or implementing throttling strategies for detected bot account communities, thereby structurally suppressing the diffusion scale of false information.

3. Discussion

3.1. Challenges

3.1.1. Interpretability

The lack of interpretability of AI has become a core issue in the intervention of LLMs in misinformation. Although current mainstream LLMs can identify misinformation in zero-shot scenarios and provide factual basis, they cannot provide a complete chain of evidence [1]. The model can neither explain "why a tweet is judged as fake news" nor point out "whether the problem lies in the news content or the visual part" that leads to it being judged as misinformation by LLMs. The opacity caused by the interpretability defect of LLMs may bring a series of problems. For example, platforms cannot give reasonable explanations for the fake news detected and requested to be deleted by LLMs, resulting in insufficient public credibility. Secondly, due to the hallucination phenomenon of LLMs, security departments cannot fully rely on LLMs to intervene in misinformation, which may waste emergency response resources.

3.1.2. Applicability

Weak applicability is becoming a key shortcoming in the detection and intervention of false information by LLMs. When some false information is presented in a different way, such as replacing rumors with popular language, injecting dialects or using cross - language mixed writing, the evaluation indicators of the model may decline. In addition, the cost of fine - tuning is relatively high, making it unable to cope with the rumor attacks launched by rumor - makers. Insufficient adaptability means that LLMs cannot respond promptly to sudden false information. When false information is not limited to a single form, LLMs may be unable to intervene and detect false information in a timely manner due to their own nature. LLMs must build the ability of continuous learning to cope with attacks of false information and safeguard the security of public information.

3.1.3. Hallucination

Hallucination is an insurmountable obstacle for LLMs. LLMs themselves generate seemingly credible false evidence, thus interfering with the detection of false information. User prompts can also interfere with LLMs, leading to hallucination. LLMs often piece together frequently occurring text fragments into statements that seem reasonable but do not conform to the facts and output such content. This makes the news judged as false information by LLMs unreliable. At the same time, LLMs may also simulate and generate virtual reports or literature based on a certain rumor on social media, and even fabricate data and citations, making it difficult for users to distinguish.

3.2. Future Prospects

3.2.1. The Combination of Expert Systems

In view of the weak interpretability of LLMs, this paper proposes to combine expert systems with LLMs to build a more powerful detection system. By integrating human expertise with automated tools, the ability to identify false cyber threat intelligence generated by LLMs can be enhanced [5].

Fusing expert systems with LLMs can combine the decades - long experience and judgment of experts with the powerful information search capabilities of LLMs. This enables machines to not only collect information extensively but also remember the theorems, norms, and cases summarized by humans. Moreover, machines can understand the experience scattered in articles, logs, and even daily conversations. Such a hybrid system is no longer a static question - answer database but a partner that can continuously learn based on human experience and wisdom.

Meanwhile, LLMs can assist expert systems in achieving natural language interaction, and expert systems can provide LLMs with the experience and rules of experts in different fields. This way, they can complement each other's strengths and promote the in - depth development of AI in various professional fields.

3.2.2. Domain Adaptation

In response to the factors contributing to the poor adaptability of LLMs, this paper proposes an adaptive system to enhance LLMs' adaptability across various domains, enabling them to have a "learn - as - you - use" framework. It further explores the self - learning ability of LLMs and fine - tuning methods for enhancing interpretability, aiming to improve the adaptability and interpretability of LLMs in social media [4]. Through continuous learning, parameter - efficient fine - tuning, and a closed - loop of situational feedback, LLMs can promptly correct small - sample data. LLMs should be connected to a real - time updated database. The model can detect in real - time whether the user's speech might be news. Once detected, the model immediately checks the authoritative news of the day to determine if the speech conforms to the reports of the day or before. If not, it directly intercepts the speech and verifies its authenticity through both text and image modules. Meanwhile, promoting data sharing and model collaboration among different social platforms can improve the ability to identify cross - platform manipulation behaviors [3].

3.2.3. Manual Intervention and System Monitoring

To address the hallucination phenomenon in LLMs, strengthen human intervention and system detection. Before the LLMs are launched, experts in various fields should conduct spot - checks on scenarios prone to hallucinations, correct errors and promptly rectify the LLMs. Provide comprehensive coverage of the LLMs, monitor and detect each piece of information generated by them. If hallucinations occur, promptly correct the LLMs and update their data sets. Strengthening human intervention and system detection can reduce the misjudgment of false information caused by the hallucination phenomenon in LLMs.

4. Conclusion

In this work, we explored the detection and intervention of false information by LLMs on social media, and analyzed and summarized previous research methods. We proposed a systematic multi - stage governance solution based on the DELL framework. Meanwhile, we used a multi - modal fusion detection method to effectively handle relatively complex scenarios and applied graph neural networks to detect and control rumors. After discussion and analysis, this paper argues that there are still problems such as weak interpretability and adaptability, and hallucination phenomena in the detection and intervention of false information by LLMs. In the future, by combining with expert domains, we can improve the system's self-adaptability and strengthen manual intervention and detection to enhance the ability of LLMs to capture and identify rumors.

Authors Contribution

All the authors contributed equally, and their names were listed in alphabetical order.

References

- [1] Shah SB, Kumar A, Yadav S, et al. Navigating the web of disinformation and misinformation: Large language models as double-edged swords. *IEEE Access*. 2024.
- [2] Papageorgiou E, Mantas J, Zoulias E, et al. A survey on the use of large language models (LLMs) in fake news. *Future Internet*. 2024;16 (8):298.
- [3] Zhang Y, Sharma K, Du L, et al. Toward mitigating misinformation and social media manipulation in LLM era. In: *Companion Proceedings of the ACM Web Conference 2024*. 2024. p. 1302-5.

- [4] Jin Y, Choi M, Verma G, et al. MM-SOC: Benchmarking multimodal large language models in social media platforms. arXiv preprint arXiv:2402.14154. 2024.
- [5] Huang H, Sun N, Tani M, et al. Can LLM-generated misinformation be detected: A study on Cyber Threat Intelligence. *Future Gener Comput Syst.* 2025:107877.
- [6] Hao G, Wu J, Pan Q, et al. Quantifying the uncertainty of LLM hallucination spreading in complex adaptive social networks. *Sci Rep.* 2024;14 (1):16375.
- [7] Large language model. Available from: https://en.m.wikipedia.org/wiki/Large_language_model
- [8] Zhang Y, Li X, Wang H, et al. DELL: A Framework for Detection and Explanation of Fake News with Large Language Models. In: *Proc ACM Conf Fairness, Accountability, and Transparency.* 2023. p. 1125-37.
- [9] Zheng L, Wang J, Zhang H, et al. Multimodal fake news detection via cross-modal alignment and semantic reasoning. In: *Proc IEEE/CVF Conf Comput Vis Pattern Recognit.* 2023. p. 1450-9.
- [10] Liu Z, Pan S. A survey on graph neural networks for fake news detection. *ACM Trans Intell Syst Technol.* 2022;13 (5):1-34.