Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.38389

Multimodal Content Understanding for Misinformation Diffusion and User Cognitive Bias in Social Media Platforms

Shuyi Wang

School of Communication and Information, Nanyang Technological University, Singapore, Singapore
rara481846778@gmail.com

Abstract. The rapid growth of social media has transformed misinformation dissemination into a multimodal and network-driven phenomenon. This study proposes a multimodal analytical framework to investigate how textual claims, visual signals, emotional expressions, and user cognitive biases jointly influence misinformation diffusion on social media platforms. This paper establishes a data set with 4,000 public posts obtained from January to March 2026 on Sina Weibo platform as the study context. 2,000 posts are selected as misinformation posts based on their refutation record while 2,000 posts are chosen as factual posts based on their verification record. A multimodal detection model based on Chinese RoBERTa and ResNet-50 and repost-comment networks are built in the experiments, the results show that the multimodal model has AUC of 0.934 ± 0.012 and F1 of 0.887 ± 0.016 and is 0.092 more than that of the text-based methods. Misinformation posts have a greater diffusion intensity, since there is a 3.84 ± 1.26-level rise in the average depth of the reposts and an increase of 0.027 ± 0.009 in network density. Negativity bias, confirmation bias, and authority cues are significant predictors in the regression analysis.

Keywords: misinformation diffusion, multimodal content understanding, cognitive bias, social media, deep learning

1. Introduction

False information on social media is increasingly communicated through interconnected modalities, including text, images, emotions, comments, and platform interactions. Its credibility is often shaped by visual realism, emotional intensity, authority cues, and social exposure, while users' judgments are also influenced by familiarity, identity, and source credibility. This makes misinformation both a computational detection problem and a cognitive decision-making issue.

Multimodal approaches can improve misinformation detection [1], but existing research often focuses more on accuracy than on post-publication diffusion. Misinformation spread is also shaped by network structure, influential users, social reinforcement, repeated exposure, and cognitive biases such as confirmation, negativity, and authority bias [2].

Recent advances in multimodal deep learning and vision-language models provide new possibilities for detection [3]. Current methods combine language, image, and adaptive fusion techniques [4], while explainable models increasingly emphasize why content is identified as

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.38389

misleading [5]. This study therefore develops a multimodal framework on Sina Weibo to jointly examine misinformation detection, diffusion, and cognitive bias.

2. Literature review

2.1. Multimodal misinformation on social media

Increasingly, misinformation on social media takes the form of combined text-image content. Images can contribute to the perception of truthfulness, emotional impression, and context of posts that influence user decisions. According to recent studies on multimodal detection of misinformation, adding textual semantics to image representation improves performance compared to purely unimodal detection systems [6]. Recent studies have also explored large vision-language models for multimodal misinformation detection, demonstrating their potential to jointly reason over textual and visual evidence [7]. Beyond misinformation detection, recent research has also explored the use of large language models for misinformation correction and intervention on social media [8]. However, the presence of multimodal components creates difficulties for such content, as the image might be used in a variety of contexts, while its consistency does not guarantee accuracy.

2.2. Mechanisms of misinformation diffusion

Diffusion of misinformation in social media is affected by the properties of the post and structure of the social network. Emotional language, identity markers, and simple message structure favor quick spread of posts. The presence of influential users, interactive groups, and frequent repetition through reposts facilitates information diffusion. Current research highlights the necessity of taking into consideration the content features along with relational and temporal ones in diffusion analysis, as popularity alone cannot determine diffusion process [9].

2.3. Cognitive bias of users in the online environment

Online information analysis is not limited to rational assessment only. Confirmation bias contributes to acceptance of information congruent with existing belief system, while authority bias adds credibility to the well-known source of information. Negativity bias makes alarming and threatening information especially noticeable. Group identity bias can enhance acceptance of information posted in the community. Recent findings of psychological and computational research prove the role of such biases in user interaction with identical information structures [10].

3. Experimental methodology

3.1. Research design and dataset collection

This study applies a computational research design combining multimodal classification, social network analysis, and cognitive bias measurement. Sina Weibo is selected as the research platform because it contains public posts, repost relationships, comment interactions, image-based communication, and official rumor-refutation records.

The dataset includes 4,000 public posts collected between January 1 and March 31, 2026. The misinformation group contains 2,000 posts confirmed through official Weibo rumor-refutation information, while the control group contains 2,000 factual posts published by verified official media accounts during the same period. The sampling topics include public health, disasters,

DOI: 10.54254/2755-2721/2026.38389

consumer risks, social security, and public policy because these domains frequently generate high-impact misinformation.

Each post record includes text content, image information, hashtags, publishing time, repost count, comment count, like count, user verification status, and top public comments. Posts without images, duplicate content, commercial advertisements, and incomplete interaction records are removed. The final dataset contains 4,000 posts, 9,742 images, 126,830 comments, and 1.84 million repost interaction records. All user identifiers are anonymized before analysis.

3.2. Features of posts and bias variables

The feature construction framework integrates four dimensions: textual features, visual features, interaction features, and cognitive bias variables. Textual features are extracted from post descriptions, hashtags, and emotional expressions. Chinese RoBERTa is used to generate semantic representations, while SnowNLP calculates sentiment intensity. Additional linguistic variables include certainty expression, emotional polarity, topic concentration, claim intensity, and misleading linguistic patterns.

Visual features are extracted from attached images using ResNet-50. The image encoder captures object categories, scene information, facial presence, warning symbols, medical scenes, disaster scenes, screenshots, and embedded text regions. Since misinformation frequently uses emotionally powerful images, visual sensationalism is measured through object intensity, threat-related objects, and emotional scene probability.

Interaction features include repost speed, comment density, user verification status, and network position. Cognitive bias variables are measured from the top 30 comments of each post. Confirmation bias is identified when comments support information consistent with users' existing beliefs. Authority bias is measured through references to experts, institutions, celebrities, or official-like accounts. Negativity bias is calculated through fear, threat, and crisis-related expressions. Group identity bias is identified through collective identity language and community defense patterns.

The multimodal representation is generated through late fusion. For a post (i), the integrated representation is defined as shown in Equation (1):

$$Z_i = \sigma\left(W_t X_i^{\text{text}} + W_v X_i^{\text{visual}} + W_m X_i^{\text{meta}} + W_b X_i^{\text{bias}} + b\right) \tag{1}$$

where X_i^{text} represents textual features extracted from Chinese RoBERTa, X_i^{visual} represents image features from ResNet-50, X_i^{meta} represents interaction characteristics, and X_i^{bias} represents cognitive bias features. W_t, W_v, W_m, and W_b are learnable parameters, and b is the bias term. The fusion layer produces a unified multimodal representation for misinformation classification. Figure 1 presents the proposed multimodal framework that integrates textual semantics, visual representations, interaction characteristics, and cognitive bias features.

DOI: 10.54254/2755-2721/2026.38389

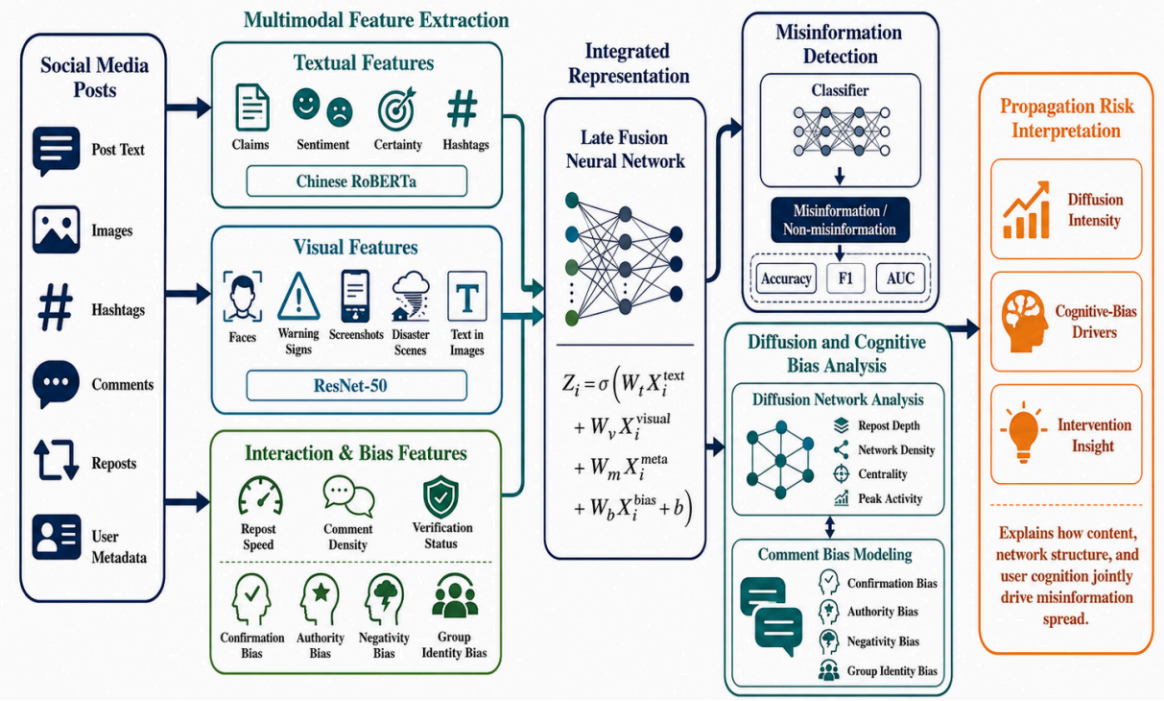


Figure 1. Multimodal framework for misinformation detection, diffusion analysis, and cognitive bias modeling

3.3. Model and software, metrics of performance and comparison

The multimodal detection model is implemented using PyTorch 2.2. Chinese RoBERTa is used as the textual encoder and ResNet-50 is used as the visual encoder. Instead of directly combining raw features, the model applies feature normalization before fusion to reduce scale differences between language and image representations. A fully connected fusion layer then learns the contribution of different modalities. Python 3.11 is used for preprocessing, including Chinese word segmentation through jieba, sentiment analysis through SnowNLP, and image preparation through OpenCV. The dataset is divided into training, validation, and testing subsets using a 70%/15%/15% split. Model performance is evaluated through accuracy, precision, recall, F1-score, and AUC.

Baseline comparisons include text-only RoBERTa, image-only ResNet-50, early fusion, and the proposed late-fusion multimodal model. Statistical analysis of diffusion and cognitive bias differences is performed using SPSS 29. Independent-sample tests compare misinformation and factual posts, while regression analysis examines whether cognitive bias variables predict diffusion outcomes.

4. Experimental procedure

4.1. Data acquisition and processing

Data acquisition is conducted through Octoparse 8 using predefined search terms related to public health, disaster events, consumer risks, social security, and public policy. The crawler collects post text, image links, hashtags, timestamps, repost numbers, comment numbers, likes, verification status, and public comments.

DOI: 10.54254/2755-2721/2026.38389

The preprocessing pipeline contains four technical stages. First, duplicate detection is conducted using text similarity and image hashing. Posts with cosine similarity above 0.95 or identical image fingerprints are removed. Second, text cleaning removes URLs, excessive symbols, advertisements, and irrelevant fragments. Third, image processing uses OpenCV to resize images, normalize pixel values, and remove corrupted files. Fourth, comment preprocessing identifies bias-related expressions through sentiment scoring and semantic matching.

The diffusion network is constructed by treating users as nodes and repost/comment behaviors as directed edges. The adjacency matrix is defined as shown in Equation (2):

$$A_{ij}(t) = \sum_{k=1}^{N} w_k I(u_i \to u_j)_k exp(-\frac{t-t_k}{\tau}) \tag{2}$$

where $A_{ij}(t)$ represents the interaction strength between users i and j, w_k represents interaction weight, $I(u_i \to u_j)_k$ indicates whether a repost or comment relationship exists, t_k represents interaction delay, and τ controls temporal decay.

4.2. Training and evaluating the multimodal model

The training process uses the training dataset to optimize textual and visual encoders simultaneously. Chinese RoBERTa converts text sequences into semantic vectors, while ResNet-50 extracts image-level representations. During training, dropout regularization and weighted cross-entropy loss are applied because misinformation and factual posts may contain different linguistic distributions. The model learns whether image-text combinations provide consistent evidence or misleading associations. For example, a disaster image combined with unrelated location information may receive a higher misinformation probability even if individual elements appear realistic. Validation performance is monitored after each epoch, and the model with the highest validation AUC is selected for final testing. The baseline text-only model uses identical training settings without visual features. Comparing these models reveals the contribution of image information to misinformation detection. Attention visualization is further applied to identify whether the model focuses on emotionally strong words, misleading image regions, or authority-related signals.

4.3. Diffusion and cognitive bias analysis

Diffusion analysis constructs repost-comment networks for misinformation and factual posts separately. Gephi 0.10 is used to calculate network density, average degree, modularity, betweenness centrality, and community structure. Diffusion speed is measured by time required to reach 50% of total repost volume. Diffusion breadth is measured through unique user participation and community coverage. Cognitive bias analysis is conducted through comment coding. The top 30 comments of each post are classified into confirmation bias, authority bias, negativity bias, and group identity bias. Two independent coders evaluate comment samples, and Cohen's kappa is calculated to ensure coding reliability. Regression models examine whether content emotion, visual sensationalism, authority cues, and cognitive bias variables predict diffusion outcomes. The dependent variables include repost volume, diffusion speed, comment quantity, and network centrality. This approach connects individual cognitive reactions with large-scale propagation patterns.

DOI: 10.54254/2755-2721/2026.38389

5. Experimental results

5.1. Performance of the multimodal misinformation detection model

The multimodal model achieves the strongest classification performance because image and text features provide complementary information. The proposed model improves detection accuracy, recall, and AUC compared with unimodal baselines (see Table 1).

Table 1. Performance comparison of misinformation detection models

Model	Accuracy	Precision	Recall	F1-score	AUC
Text-only RoBERTa	0.812 ± 0.018	0.798 ± 0.021	0.784 ± 0.024	0.791 ± 0.019	0.842 ± 0.017
Image-only ResNet-50	0.763 ± 0.022	0.741 ± 0.025	0.752 ± 0.021	0.746 ± 0.023	0.801 ± 0.019
Early fusion model	0.846 ± 0.017*	0.831 ± 0.020*	0.824 ± 0.018*	0.827 ± 0.019*	0.879 ± 0.014*
Proposed multimodal model	0.902 ± 0.013***	0.891 ± 0.015***	0.883 ± 0.017***	0.887 ± 0.016***	0.934 ± 0.012***

*Note: * $p < 0.05$; *** $p < 0.001$.*

5.2. Diffusion differences between misinformation and non-misinformation posts

Misinformation posts demonstrate significantly stronger diffusion characteristics. Compared with factual posts, misinformation achieves greater repost depth, higher interaction density, and faster peak activity. Average repost depth increases by 3.84 ± 1.26 levels*, while network density increases by 0.027 ± 0.009*. The average time to reach peak diffusion decreases by 6.42 ± 2.11 hours.

5.3. Relationship between user cognitive bias and misinformation spread

Regression analysis shows that negativity bias, confirmation bias, and authority cues significantly predict misinformation diffusion. Negativity bias has the strongest effect on repost volume (β = 0.428, $p < 0.001$), followed by confirmation bias (β = 0.317, $p < 0.001$) and authority bias (β = 0.286, $p < 0.01$). Group identity bias mainly affects comment clustering and community concentration. Posts containing high emotional intensity and strong identity signals achieve faster diffusion peaks and larger interaction networks.

6. Conclusion

The current work offers a multimodal approach framework for the study of misinformation and user cognitive bias on social media platforms. Through the combination of textual semantics, visual information, interaction networks, and cognitive bias factors, the framework is able to conceptualize the phenomenon of misinformation as one that involves all the three aspects of content properties, social structure, and user psychology. Compared with the unimodal approach, the multimodal approach has proven its superior effectiveness in terms of detection while the network analysis has revealed that the process of misinformation occurs through more complex and dense interaction networks. The analysis of user cognitive bias has revealed that such factors as emotional intensity, confirmation bias, and authority significantly impact user behavior.

DOI: 10.54254/2755-2721/2026.38389

References

[1] Da'u, A., & Nafea, I. (2025). A systematic review of multimodal fake news detection on social media using deep learning models. *Results in Engineering, 26*, 104752.
[2] Liu, M., Yan, K., Liu, Y., Fu, R., Wen, Z., Liu, X., & Li, C. (2024). MisD-MoE: A multimodal misinformation detection framework with adaptive feature selection. *Proceedings of the 4th NeurIPS Efficient Natural Language and Speech Processing Workshop*, 114–122.
[3] Jin, Y., Choi, M., Verma, G., Wang, J., & Kumar, S. (2024). MM-SOC: Benchmarking multimodal large language models in social media platforms. *Findings of the Association for Computational Linguistics: ACL 2024*, 5745–5761.
[4] Zeng, F., Li, W., Gao, W., & Pang, Y. (2024). Multimodal misinformation detection by learning from synthetic data with multimodal LLMs. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 10467–10484.
[5] Qi, P., Yan, Z., Hsu, W., & Lee, M. L. (2024). SNIFFER: Multimodal large language model for explainable out-of-context misinformation detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13052–13062.
[6] Hu, W., Wang, Y., Jia, Y., Liao, Q., & Zhou, B. (2024). A multi-modal prompt learning framework for early detection of fake news. *Proceedings of the International AAAI Conference on Web and Social Media, 18*(1), 651–662.
[7] Tahmasebi, S., Müller-Budack, E., & Ewerth, R. (2024). Multimodal misinformation detection using large vision-language models. arXiv preprint arXiv:2407.14321.
[8] Zhou, X., Sharma, A., Zhang, A. X., & Althoff, T. (2024). Correcting misinformation on social media with a large language model. arXiv preprint arXiv:2403.11169.
[9] Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2020). FakeNewsNet: A data repository with news content, social context, and dynamic information. *Big Data, 8*(3), 171–188.
[10] Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences, 25*(5), 388–402.