IJISEM
INTERNATIONAL JOURNAL OF INNOVATIONS IN SCIENCE ENGINEERING AND MANAGEMENT

OPEN ACCESS

Volume: 5

Issue: 3

Month: September

Year: 2026

ISSN: 2583-7117

Published: 01.09.2026

Citation:

Ekta R. Kaneriya, Disha M. Ganatra "Beyond Single-Modality Detection: A Systematic Review of Multimodal AI for Social Media Cybercrime" International Journal of Innovations in Science Engineering and Management, vol.5, no.3, 2026, pp. 505- 509

DOI:

10.69968/ijisem.2026v5i3505-509

Beyond Single-Modality Detection: A Systematic Review of Multimodal AI for Social Media Cybercrime

Ekta R. Kaneriya[1], Disha M. Ganatra[2]

[1]Ph.D. Research Scholar, Department of Computer Science, Atmiya University, Rajkot, Gujarat
[2]Assistant Professor, Department of Computer Science, Atmiya University, Rajkot, Gujarat

Abstract

Social media platforms have emerged as a key battleground for cy-bercriminals, which leverages the speed, scale and interconnectedness of social media to orchestrate financial crimes such as phishing and OTP Fraud, identity-based crime such as deep-fake and account takeover, and social harms such as cyberbullying and coordinated influence campaigns. The threats are increasingly multifaceted and cross-platform, making it difficult for traditional detection solutions to stay ahead of the curve, as they are typically limited to a specific data type or a single platform. We review recent publications on detecting cybercrime in social networks ranging from misinformation detection, to machine learning and deep learning based classifiers, fraud and intrusion detection pipelines, to novel multimodal architectures for AI. The reviewed literature can be divided into three the-matic groups: (1) detection methods based on content and propagation patterns, (2) methods based on user behaviour and network structure, and (3) multimodal fusion models coupled with explainable AI mechanisms. We evaluate the data sets used, algorithm selection, reported performance, and inherent limitations of the underlying studies for each cluster. The analysis shows that there is a gap. Although there has been significant progress in the detection of individual tasks, existing systems are still highly specialised and platform-dependent, or type-dependent, or modal-ity dependent. Very few integrate text, image and behaviour into a single model. It also identifies a lack of attention to interpretability and cross-platform real time operation. To address these gaps, we suggest the necessity of developing a multimodal detection system that is uniform, adaptive and explainable, which can work in heterogeneous environments on social media platforms, and we propose some research directions for achieving this goal.

Keywords; Social Media Cybercrime, Multimodal AI, Cross-Platform Detec-tion

INTRODUCTION

Social media platforms like Facebook, Instagram, WhatsApp, X, and YouTube now shape almost every part of life, from staying in touch with family to conducting busi-ness persons. Naturally, that has also drawn the attention of cybercriminals looking for easy and high-volume targets. What began with simple phishing links and spam has grown into fake news, deepfake media, identity theft, cyberbullying, hate speech, fake accounts, cryptocurrency and romance scams, and digital payment fraud, many of which now combine text, images, and user behaviour, rendering single modality detec-tion systems inadequate.

Techniques from Machine learning, deep learning, and NLP all under the broader of umbrella of AI have played a major role in identifying these threats. Multimodal AI extends this further by combining multiple datatypes at once, which tends to catch more than any single method alone. Explainable AI (XAI) matters here too without it; even accurate model is hard for an analyst to trust. Yet most, published work still looks at just one type of cybercrime or platform, and struggle to keep pace with rapidly evolv-ing, cross platform attacks.

In this review, we look at how recent work has approached multimodal learning, adaptive detection, explainability, and cross-platform coverage for social media cyber-crime detection.

We weigh what each approach gets right, where it falls short, and what gaps still need to be filled.

Objectives:

Map the recent AI-driven techniques for detecting social media cybercrime. Evaluate existing multimodal and cross-platform detection approaches. Identify strengths and limitations of current methods. Highlight the research gaps to future multimodal, explainable detection systems.

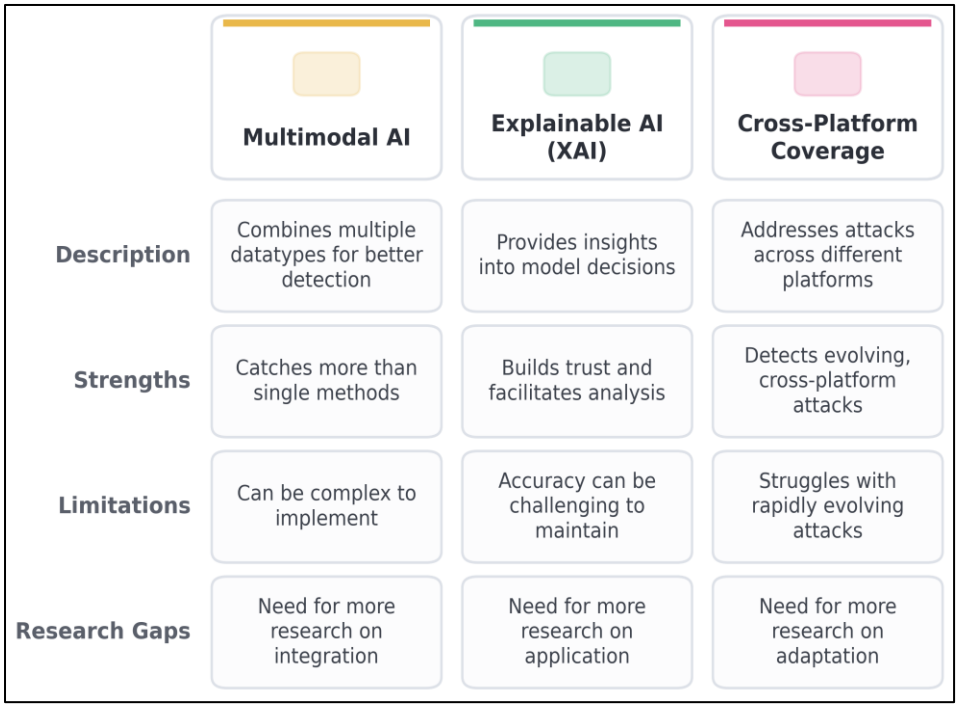

Figure 1. sums up why we structured the review this way — each of these three areas keeps surfacing as both a strength and a weak point across the papers we looked at. (fig-ure created using Napkin AI)

LITERATURE REVIEW

Content- and Propagation-Based Misinformation Detection

Multimodal fake news detection has moved fast in the last two years. Nasser et al. [1] reviewed 121 studies published between 2018 and 2025 and found that trans-former models and recurrent networks now dominate the field, with Twitter and Weibo remaining the two most common testbeds. Within this space, contrastive learning has become a popular tool for aligning text and image representations. A recent study com-bining contrastive learning with optimal transport [2] showed that this pairing improves cross-modal alignment better than earlier fusion strategies. GAMED [3] took a different route, using. A multi-expert decoupling architecture that separates knowledge-driven and content-driven signals are separated before merging, and this separation is what lets the modal generalise across different types of fake news. We find this decoupling idea particularly useful because most fusion models We reviewed skip this step entirely. UNITE-FND [4] approached the same problem from another angle entirely: rather than fusing modalities directly, it translates visual scenes into text and lets a language model reason over both, sidestepping some of the alignment problems that plague image-text fusion. A broader survey [5] reached a similar conclusion across all these approaches: multimodal methods consistently outperform text-only baselines, but performance still varies a lot depending on how well the fusion architecture handles missing or noisy modalities.

Behavioural and Network-Based Cybercrime Detection

All cybercrime signals don't necessarily originate from content. Much of it is related to the behavior and interconnection of accounts. In MSM-BD [6] the authors solved the problem of bot detection by fusing heterogeneous data (profile images, tweet text, and account statistics) with

a cross-modal attention mechanism, and evaluated the network on TwiBot-22 benchmark. The results indicated that it is still difficult to detect such bots using only textual information, but some of this gap can be bridged by behavioural and visual information. Fraud detection has been along similar lines. In a recent work on graph fraud detection using LLMs [7] social platforms and e-commerce systems were represented as multimodal graphs where entities such as users, products and content are represented both structure and semantically. By infusing large language models into this graph structure, the system acquired a reasoning ability about fraud patterns that are often not captured by traditional graph neural networks, especially for fake reviews and AI-generated deceptive content.

Multimodal Fusion and Explainable AI

Simultaneously merge video, audio and text, and detection has pushed us beyond the fusion of text and images. Even though the real world is creating increasingly multimodal deepfakes, most of the digital forensic methods surveyed by Qureshi et al. [8] still consider video, audio and text as individual challenges. We confirmed this with real test data in the Deepfake-Eval-2024 benchmark [9]: Models trained on older academic datasets struggled significantly on deep-fakes that were actually being spread on social media in 2024, primarily due to their lack of representation of newer generation methods such as diffusion-based synthesis.

Hate speech detection has run into a related problem. Two shared-task systems, ARC-NLP [10] and the MHS-STMA framework [11], both found that combining text with embedded images substantially improves detection, especially for memes and propaganda images where the harmful meaning only emerges from the combination of the two.

A separate study on federated cross-modal graph transformers [12] pushed the cross-platform angle further, modelling text, image, audio, and social graph structure together in a decentralised, privacy-preserving setting, and found that adversarial training was necessary to stop attackers from exploiting the federated structure itself. Joshi et al. [13] looked specifically at explainable misinformation detection across multiple platforms and found that very few systems generate explanations that generalise once a model is retrained for a new platform, which points to explainability being treated as an after-thought rather than a design constraint.

Research Gap

In all these studies, there is a clear leap forward [1] and [8] in individual parts of a problem, but the pieces are not yet put together. Content-based multi-modal models [1, 3, 4, 5] are good at detecting manipulated text-image pairs, but they were not designed for behavioral characteristics such as posting frequency or account age. Most behavioral and graph-based models [6] [7] effectively detect bots and fraud rings, but largely leave out the content they produce. Explainability work [13], [14] is progressing but nearly all studies evaluate it for a single modality or a single platform, in a real, mixed-platform environment, it remains to be seen how well these explanations will transfer. There are no systems reviewed that integrate content, behavior and cross-platform operation under one banner of explainability in one system. Having such a gap would involve, at least, pairing up fusion model with a behavioral signal source such as [6] and an explainability layer that current systems lack for the survivors that need to be held onto for a new platform as demonstrated by [13].

METHODOLOGY

The scope of this review was limited to recent high-quality research related to the use of multimodal AI systems in the detection of cybercrime on social media. The papers were mainly collected from the arXiv and IEEE arXiv websites, A strong preference for papers published between 2023 and 2025 to maintain the review's freshness in the context of the rapid development of this field, and a preference for articles available on the Internet, such as those found in and on ACM digital libraries, ScienceDirect, Springer and Frontiers.

The following terms were used to search for multimodal fake news detection, multimodal deepfake detection, multimodal hate speech detection, multimodal bot and fraud detection, and explainable multimodal AI. Only studies that examined two or more data mo-dalities (text, image, video, audio, behavioural/graph features) used in the specific context of a social media cybercrime problem were included. The following requirements were used to select papers: purely editorial papers or single-modality studies were excluded.

The modalities combined in each selected paper, its fusion strategy, the datasets and benchmarks employed, the reported performance and any explicit mention of the concept of explainability or cross-platform generalisation were analysed in each selected paper. This analysis is used to create the three thematic clusters found in the literature review.

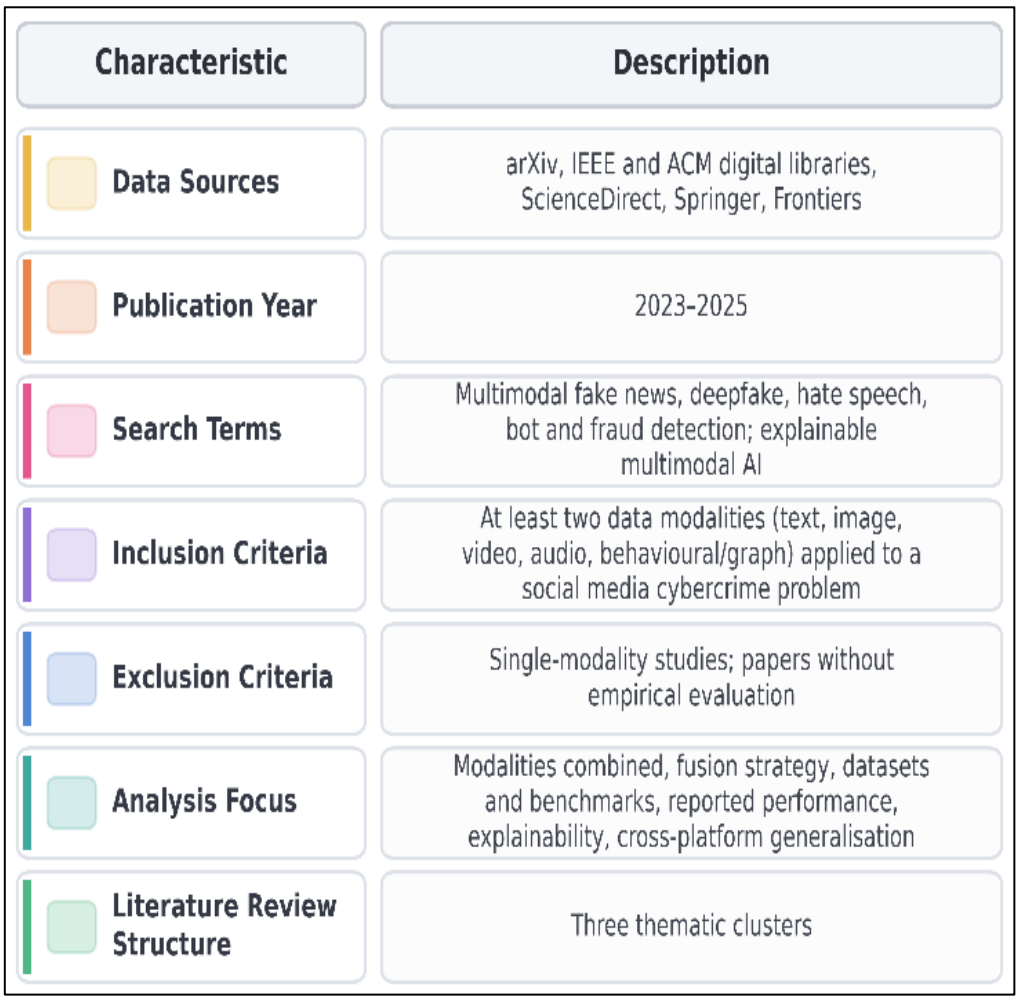

Figure 2 How the papers for this review were selected — where we looked, the time window we used, the search terms, and the criteria for including or exclud-ing a study.

Figure created using Napkin AI

RESULTS AND DISCUSSION

One thing is evident across the literature that has been reviewed: the combination of modalities is almost always more effective than using one type of modality alone. All studies that compared the accuracy of multimodal to unimodal baselines, whether they involved fake news [1] [4] and hate speech [10] [11] or bot and fraud detection [6] [7] showed a significant increase in accuracy. However, the gain is highly dependent on the fusion's implementation. Simple feature concatenation was outperformed by contrastive and attention-based fusion methods consistently [2] indicating the importance of the way modalities are fused, as well as the modalities themselves.

One area where the gap between the "academic benchmarking" and the "real world" is largest is deepfake detection. The models that performed well on the older data failed significantly on the real deepfakes found on social media [9] – telling message number two: performance on benchmarks doesn't mean performance in the real world, not least because of the continuous improvement of generation techniques.

Cross-platform detection remains the weakest area overall. Only the federated graph transformer study [12] attempted to operate across a genuinely decentralised, multi-platform setting, and even that was tested on synthetic rather than live cross-platform data. This reinforces the research gap identified earlier: an approach that fuses content, behaviour, and explainability while genuinely working across platforms hasn't been demonstrated yet, and building one is the natural next step for this field.

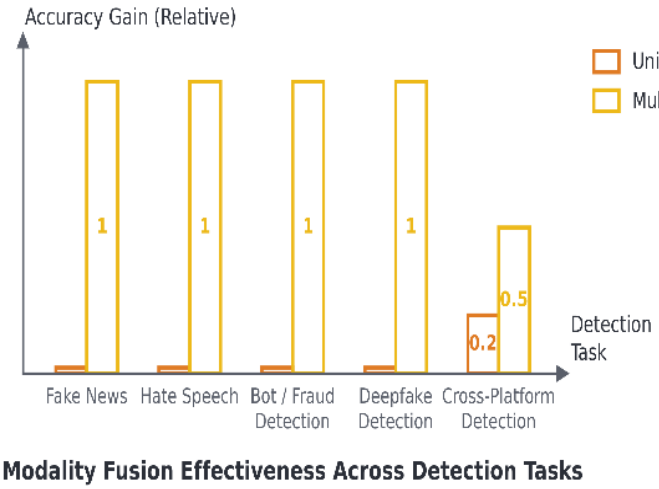

Figure 3. Multimodal approaches beat unimodal ones across every task we looked at, though the gap is smaller for cross-platform detection than for the others. (figure created using Napkin AI)

CONCLUSION

This review looked at how AI has been applied to cybercrime detection on social media, covering machine learning basics, fake news detection, bot and fraudulent account de-tection, multimodal learning, and explainable AI.

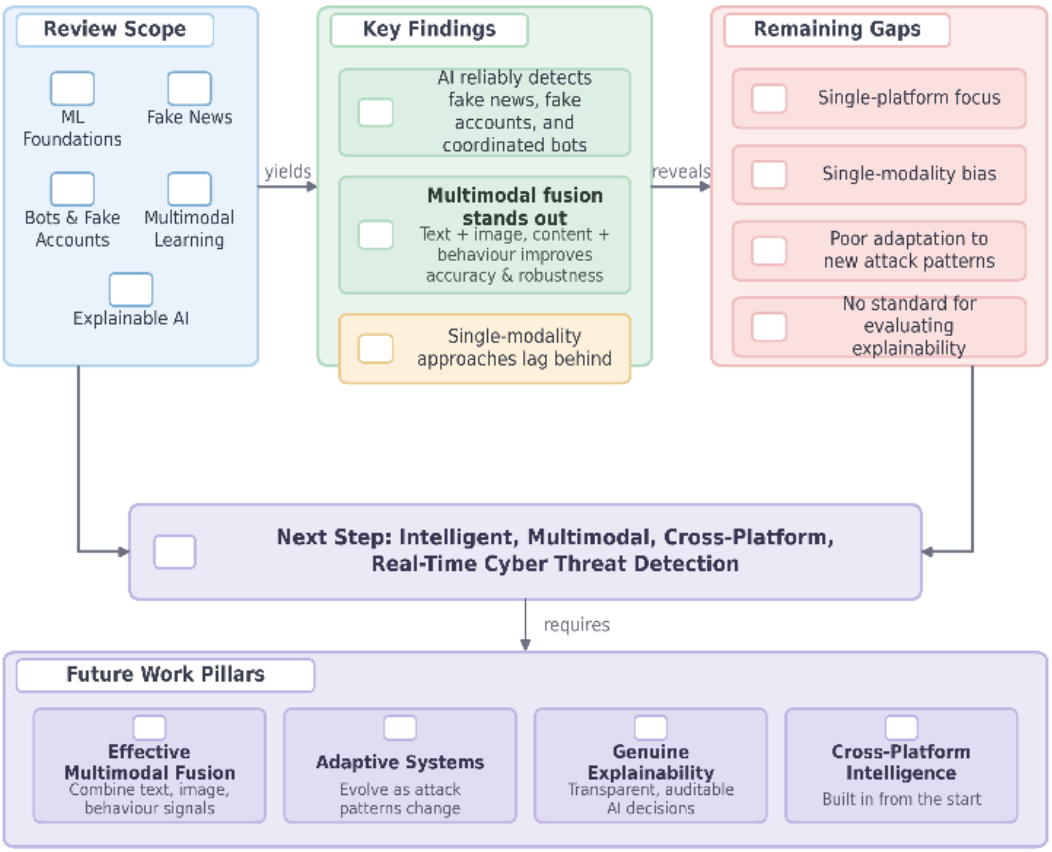

Figure 4 Where this review started, what it found, the gaps that are still open, and the direction future work needs to take. (Figure created using Erase.io)

Across the studies we examined, AI methods consistently proved effective at catching fake news, fake accounts, and coor-dinated bot activity. Multimodal fusion stood out in particular. Combining multiple types of information like text with images, or content with behaviour, tended to improve both accuracy and robustness. Single-modality approaches simply couldn't match that. Real gaps remain, though. Most studies focus on just one platform. Most models still work with only one type of data. Few systems adapt well as new attack patterns emerge, and there is no agreed-upon way to evaluate how explainable a model actually is. These gaps point toward a fairly clear next step: building an intelligent, multimodal system that can detect a range of cyber threats in real time, across platforms rather than within just one.

Getting there will take more work on several fronts. multimodal data needs to fused more effectively, system needs to adapt as attacks evolve, AI decisions need to be gen-uinely explainable, cross-platform intelligence needs to be built in from the start.

REFERENCES

[1] M. Nasser *et al.*, "A systematic review of multimodal fake news detection on social media using deep learning models," *Results in Engineering*, vol. 26, p. 104752, 2025, doi: https://doi.org/10.1016/j.rineng.2025.104752.

[2] X. Shen, M. Huang, Z. Hu, S. Cai, and T. Zhou, "Multimodal Fake News De-tection with Contrastive Learning and Optimal Transport," *Front. Comput. Sci.*, vol. Volume 6-2024, 2024, doi: 10.3389/fcomp.2024.1473457.

[3] L. Shen *et al.*, "GAMED: Knowledge Adaptive Multi-Experts Decoupling for Multimodal Fake News Detection," 2024.

[4] Mukherjee and S. Ghosh, "UNITE-FND: Reframing Multimodal Fake News Detection through Unimodal Scene Translation," 2025.

[5] J. Lv, Y. Gao, L. Li, L. Shi, and S. Li, "Multi-modal fake news detection: A comprehensive survey on deep learning technology, advances, and challenges," *Journal of King Saud University Computer and Information Sciences*, vol. 37, Jul. 2025, doi: 10.1007/s44443-025-00317-7.

[6] T. Wu, Z. Ma, Y. Cui, Z. Zhou, and E. Wang, "MSM-BD: Multimodal Social Media Bot Detection Using Heterogeneous Information," 2025.

[7] T. Huang, Y. Wang, Q. Li, C. He, and J. Gao, "Can LLMs Find Fraudsters? Multi-level LLM Enhanced Graph Fraud Detection," in *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)*, New York, NY, USA: ACM, 2025, pp. 1530–1538. doi: 10.1145/3746027.3755245.

[8] S. M. Qureshi, A. Saeed, S. H. Almotiri, F. Ahmad, and M. A. Al Ghamdi, "Deepfake forensics: a survey of digital forensic methods for multimodal deep-fake identification on social media," *PeerJ Comput. Sci.*, vol. 10, p. e2037, 2024, doi: 10.7717/peerj-cs.2037.

[9] N. A. Chandra *et al.*, "Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024," 2025.

[10] U. Sahin, I. E. Kucukkaya, O. Ozcelik, and C. Toraman, "ARC-NLP at Multi-modal Hate Speech Event Detection 2023: Multimodal Methods Boosted by Ensemble Learning, Syntactical and Entity Features," 2023.

[11] Chhabra and D. K. Vishwakarma, "MHS-STMA: Multimodal Hate Speech Detection via Scalable Transformer-Based Multilevel Attention Framework," 2024.

[12] D. Premkumar and S. K. Nachimuthu, "Privacy-preserving cyberthreat detec-tion in decentralized social media with federated cross-modal graph transform-ers," *Sci. Rep.*, vol. 16, no. 1, p. 3608, 2026, doi: 10.1038/s41598-025-33596-1.

[13] G. Joshi *et al.*, "Explainable Misinformation Detection Across Multiple Social Media Platforms," 2022.

[14] R. Jadhav, V. Meshram, A. Bhosle, K. Patil, S. Dash, and S. Jadhav, "Explain-able multilingual and multimodal fake-news detection: toward robust and trust-worthy AI for combating misinformation," *Front. Artif. Intell.*, vol. Volume 8-2025, 2025, doi: 10.3389/frai.2025.1690616.