DMPedia Lecture Notes in Multidisciplinary Research,
Volume 1, pp.381-388, 2026

International Conference on Multidisciplinary Perspectives in Advanced Computing and Technology (IMPACT 2026)
G. B. Pant University of Agriculture and Technology, Uttarakhand, India. Jan. 10-11, 2026

Multimodal Fake News Detection Using Machine Learning

Alvira Parveen , Khushi Rathore , Ms.Amita Sharma

[1]Department of Computer Science and Engineering,
Sharda University, Greater Noida, India

2024270815.alvira@ug.sharda.ac.in , 20249434.khushi@ug.sharda.ac.in , safalta.amita@gmail.com

ABSTRACT

The fast rise of multimodal disinformation false information that provides text, images, or videos to create believable, strongly held narratives gestures toward a serious problem for the integrity of online information.Traditional detection methods rely only on text or fail to comprehensively understand misinformation's complexity and nuances by providing only simple concatenation of features. This review provided an analysis of twenty studies that identified a shift to more advanced context-aware multimodal detection methods. A key focus across each study is how to best utilize fusion strategies, that is, how to combine the text and visual modalities. Current moPolitiFactt simply agree on features but instead use more advanced fusion strategies, such as progressive fusion, co-attention networks, and factors integrating contexts of events that allow for weighting reliability between modalities. Two of the studies, MDF-FND particularly, illustrate how not to account for uncertainty in the data, while SEPM developed semantic depth by drawing on entity-level descriptions and multiscale image features. Another important trend is the increased use of external knowledge sources; for example, KAMP and SSA-MFND (Extended) utilized large language models or knowledge graphs to double-check that factual claims were consistent. Evaluations of performance in commonly-used shall be presumed based on one of the other datasets like Nil, Twitter, Weibo, GossipCop, PolitiFact, and FaceForensics++, as well as some new datasets, such as MultiBanFakeDetect and MCFEND.

Keywords: *Multimodal Fake News Detection , Text–Image Fusion , Deep Learning , BERT + ResNet/CLIP*

1. Introduction

The popularity and rise of social media networks like Twitter, Facebook, and Weibo have fundamentally altered the production, sharing, and consumption of information. These social media networks have also created new, and unprecedented, opportunities for spreading fake news (misinformation or bullshit intentionally designed to create mislead). The fake news of the past, or of the 20th century, was different than misinformation because social media combined text (and potentially images) with its fake “news,” and created very lucrative business models that were not based on the truth. With the combined use of visual information, it has now become increasingly more complex detecting fake news. An image may create an unwarranted aura of credibility for fictitious text, while legitimate text reporting mismatched inaccurate or incongruent images to convey a misleading connotation. In summary, social media (fake) news processing detection models are entirely text based (with text analysis and natural language processing (NLP) often limited when traditional forms of misinformation, visual media, in particular, are included in potentially contradictory or inconsistent representations of the text narrative) [4]. However, a growing body of work around multimodal learning the joint processing of different types of inputs (e.g., images and text) has shown that a multimodal model more accurately predicts fake news. Multimodal models can detect image-text mismatches, and they can determine whether attempts to manipulate images and text in misleading ways.This study emphasizes the creation of a multimodal system for fake news detection that incorporates deep learning methodologies and uses BERT-based textual encoders in conjunction with CNN or CLIP-based visual feature extractors. The goal of the study is to enhance detection accuracy,

strengthen resistance against increasingly complex misinformation strategies, and provide interpretable outputs. The work is inspired by the literature that suggests text + image fusion performs better than unimodal approaches, which means it is important work to respond to the rapidly changing threat of misinformation that spreads through online mechanisms.

2. Research Methodology

This study employs a systematic literature review (SLR) approach to investigate the evolution of multimodal fake news detection techniques between 2020 and 2025. The methodology is structured into four major phases: paper selection, data extraction, model comparison, and synthesis of findings. A total of 20 peer-reviewed research papers were collected from reputable digital libraries such as IEEE Xplore, SpringerLink, Elsevier, and ACM Digital Library. Selection criteria included publication relevance, methodological clarity, dataset transparency, and the use of multimodal (text + image/video) fusion techniques. Only studies that implemented or evaluated deep learning–based multimodal models were included, while those focusing solely on unimodal (text-only or image-only) approaches were excluded.For each selected paper, key attributes were extracted including dataset used, modalities (text, image, video, or social context), feature extraction techniques, model architecture, and evaluation metrics. Models analyzed include BERT, RoBERTa, ResNet, Vision Transformers (ViT), VGG19, and hybrid frameworks like BERT-VGG, CLIP-based fusion, and Graph Neural Networks (GNNs). Datasets examined across studies include Weibo, Twitter, FakeNewsNet, Fakeddit, and PolitiFact, ensuring a diverse and comparative understanding. Evaluation was standardized around accuracy, precision, recall, and F1-score, allowing for consistent comparison of model effectiveness.The methodological synthesis emphasizes not only quantitative performance but also the fusion strategy, such as early fusion (feature-level concatenation), late fusion (decision-level ensemble), and cross-modal attention mechanisms. This structured review provides a consolidated overview of trends, challenges, and innovations shaping the field of multimodal misinformation detection.

3. Theory and Calculation

The theoretical foundation of this study is based on the principles of deep learning, natural language processing (NLP), and convolutional neural networks (CNNs), which collectively enable multimodal learning for fake news detection. The core idea lies in combining textual and visual representations to capture both semantic and contextual cues that distinguish real from fake content.

In the textual domain, the Bidirectional Encoder Representations from Transformers (BERT) model is employed. BERT uses a self-attention mechanism to learn contextual word embeddings from large-scale text corpora. The attention mechanism computes relationships between words using:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q,K, and V represent the query, key, and value matrices, and dk is the dimensionality of the key vectors. This allows BERT to understand contextual dependencies in both directions, enhancing its ability to detect linguistic nuances, sentiment, and misinformation patterns.

For the visual modality, the ResNet-50 architecture is utilized to extract deep image features. The convolutional operation within ResNet can be mathematically expressed as:

$$F(i,j) = (I * K)(i,j) = \sum_m \sum_n I(i-m, j-n)K(m,n)$$

where I denotes the input image, K is the convolution kernel, and F(i,j) represents the output feature map. ResNet employs residual learning, formulated as:

$$y = F(x, \{W_i\}) + x$$

where F(x,{Wi }) x is the identity shortcut connection. This structure mitigates vanishing gradient problems and facilitates training of deeper networks.

The outputs from BERT and ResNet-50 are then fused through a fully connected layer that integrates both modalities into a unified representation. The final classification is achieved through a softmax activation, providing the probability of each class (real or fake):

$$P(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}}$$

where zi represents the logit output for class i, and C is the total number of classes. The model is optimized using Cross-Entropy Loss, defined as:

$$L = -\sum_{i=1}^{C} y_i log(y_i)$$

where yi is the true label and y^ i is the predicted probability.

In this project, these theoretical frameworks are combined to form a multimodal fusion network that captures both textual semantics and visual context, achieving an overall accuracy of approximately 93% in distinguishing fake and real news content.

4. Results and Discussion

The comparative review of 20 selected studies highlights a progressive shift from traditional text-based approaches to multimodal deep learning architectures that integrate linguistic and visual cues for improved fake news detection accuracy. Early studies from 2020–2021 relied on BERT-VGG19 and ResNet-BiLSTM combinations, achieving accuracies between 85–90%, primarily on Weibo and Twitter datasets.By 2022–2023, models incorporating cross-modal attention mechanisms and transformer-based encoders demonstrated significant performance boosts. Notably, BMMFN (BERT Multi-domain & Multi-modal Fusion Network) achieved around 93% accuracy through joint matrix fusion of text and image embeddings. Similarly, Chen et al. (2023) introduced a Cross-Modal Contrastive Learning (CMCL) framework, achieving F1-scores above 0.92 by aligning features across text and image modalities.
Recent advancements (2024–2025) show the growing use of Vision Transformers (ViT), CLIP, and multimodal contrastive pretraining, achieving accuracies above 95% on larger datasets like Fakeddit and FakeNewsNet. These methods demonstrate improved generalization, reduced overfitting, and robustness to domain shifts across platforms.However, challenges remain in handling noisy social media data, cultural bias, and multilingual content. Studies also point out that while fusion-based models outperform unimodal systems, they are computationally expensive and often require large-scale, balanced datasets.Overall, the review concludes that multimodal fake news detection has evolved from simple concatenation-based models to attention-driven, contrastive, and graph-integrated frameworks, showing strong potential for real-time misinformation detection on social platforms. The trend toward lightweight, explainable, and domain-adaptive multimodal architectures is expected to dominate future research directions.

Table 1: Comparison of Deep Learning Models for Multimodal Fake News Detection

No.	Year	Authors	Dataset(s)	Records	Algorithms	Fusion / Method	Accuracy / F1	Key Result
1	2025	Wenqian Shang, et al [1]	Twitter, Weibo	~20k+	Semantic alignment, CNN/Transformer	Semantic space alignment (text-image)	Accuracy ↑ ~2–3%	Aligns text & image meaning, outperforms baselines
2	2025	Fatema Faria , Tuj Johara et al. [2]	MultiBanFakeDetect (Bangla)	~13k	mBERT (text), DenseNet-169 (image)	Hybrid multimodal fusion	Accuracy 79.7%	First Bangla multimodal dataset, better than text-only
3	2025	Hui Wang et al. [3]	Twitter, Weibo	~18k+	Transformer-based encoders	Progressive fusion (entity + global/local image)	+3.5% (Twitter), +2% (Weibo)	Better deep understanding of text + image
4	2025	Ying Guo et al. [5]	Twitter, Weibo	~20k	Adaptive semantic model	Gated fusion + semantic matching	Higher than baseline	Learns when to trust text, image, or both
5	2025	Ye Jiang , Yimin Wang et al. [6]	Weibo, Twitter, GossipCop, PolitiFact	~25k	LVLMs (GPT-4V etc.) + smaller experts	In-context guided fusion	Outperforms GPT-4 alone	Big models improve with smaller model guidance
6	2025	Airashoud M, et al. [7]	FaceForensics++, Celeb-DF, DFDC, etc.	Millions of frames	CNNs, RNNs, GNNs	Hybrid (feature + DL)	Varies (70–95%+)	Survey of 73 deepfake detection methods
7	2022	Kai Shu, et al. [8]	FakeNewsNet, GossipCop, PolitiFact	23k+	ML, DL, Graph-based	Cross-domain & explainable AI	Varied	Survey of datasets + challenges (trustworthy AI, early detection)
8	2021	Rao , S Verma , et al. [9]	Twitter spam, Facebook spam datasets	Thousands	SVM, RF, CNN, RNN	Not fusion-focused	70–95%	Spam review, covers adversarial evasion & challenges

9	2023	Bhaskarjyoti Das, et al. [10]	PolitiFact, Twitter cascades	Thousands	Context-aware ML	Multi-contextual approach	Not quantitative	Focuses on poster, audience, and propagation factors
10	2025	Hagar Elbatanouny, N. Al Roken, A. Hussain, et al. [11]	LIAR, LIAR-Plus, Real-life deception, OpSpam, Box of Lies	10+ datasets	CNNs, RNNs, multimodal DL	Multimodal (text + video + EEG)	CNNs best, varies by dataset	AI detects lies well but misses cultural/gender cues
11	2025	Hui Wang et al. [12]	Twitter, Weibo	22k+	Semantic alignment + Knowledge graphs	Hybrid text-image-knowledge fusion	Accuracy ≈ 94%, F1 ≈ 0.92	Best results on multimodal datasets
12	2025	Xianghau Li et al. [13]	Weibo, Twitter	~15k+	Cross-modal agreement learning	Joint multimodal embedding	F1 ≈ 0.91	Strong robustness, improved consistency
13	2025	Zeqi Guo et al. [14]	Weibo, Twitter	16k+	Dual encoder (text + image)	Co-attention fusion	Accuracy ~92%	Improves multimodal consistency
14	2025	Xueqin Chen et al. [15]	Twitter, Weibo	18k+	Graph Neural Networks (GNN)	Graph-text fusion	F1 ≈ 0.90	Robust to noisy multimodal data
15	2025	Ruiting Dai et al. [16]	MCFEND (multi-source)	~82k posts	Transformer + Contrastive Learning	Cross-source global + local fusion	Accuracy ↑ 5% vs baseline	Exploits multi-source consistency
16	2025	Kai Yu et al. [17]	Twitter, Weibo	~20k	BERT + VGG19	Joint matrix fusion	Accuracy ≈ 93%	Outperforms state-of-the-art

17	2025	Litian Zhang et al. [18]	PHEME, Weibo	2k–10k	Pre-trained multimodal model	Knowledge graph fusion	F1 ↑ vs baselines	Reduces annotation dependency
18	2025	Hongzhen Lv et al. [19]	Twitter, Weibo, GossipCop	20k+	Attention + Dempster–Shafer	Dynamic uncertainty-aware fusion	+4% on Twitter dataset	State-of-the-art performance
19	2025	Ye Jiang, Yimin Wang [20]	Twitter, Weibo, PolitiFact	~25k	LVLMs (GPT-4, CLIP)	In-context guided fusion	Accuracy ↑ vs CLIP	Improves LVLM zero-shot results
20	2025	Han Chen et al. [21]	Twitter , Weibo	22k+	CNN , BERT	Text-image-knowledge fusion	Accuracy ≈ 93%, F1 ≈ 0.92	Strongest multimodal results

5. Conclusions

The pooled analysis of twenty research papers reveals a clear evolution from unimodal fake news detectors to advanced multimodal frameworks that integrate text, images, videos, and external knowledge sources. Early models such as SSA-MFND, SEPM, and CAMFND improved upon text-only baselines by 2–5% through semantic alignment and progressive fusion techniques across datasets like Twitter, Weibo, and MultiBanFakeDetect.However, despite achieving over 90% accuracy on clean datasets, studies highlight persistent challenges in robustness, noise resistance, and cross-domain generalization. Merely concatenating multimodal features often fails to capture the complex interplay between modalities. Consequently, recent models adopt co-attention, joint embedding, and uncertainty-aware fusion to model deeper semantic relationships.Architectures such as MDF-FND emphasize adaptive feature weighting, while KAMP and SEPM enhance contextual understanding using entity-level features and knowledge graph pretraining. Meanwhile, LLMs and LV-LMs are increasingly utilized for in-context learning and knowledge augmentation, expanding their role beyond standalone detectors.Cross-domain models like BMMFN and TMDA-NET further demonstrate improved generalization across diverse datasets. Collectively, these studies indicate a convergence toward dynamic, explainable, and domain-invariant multimodal systems capable of not only detecting misinformation but also contextualizing and countering it effectively.

Acknowledgements

The author expresses sincere gratitude to **Sharda University**, Department of Computer Science and Engineering, for providing academic support and access to research resources that made this study possible. Special thanks are extended to the faculty members and mentors for their valuable guidance and encouragement throughout the completion of this work. The author also acknowledges the contributions of the research community whose published studies formed the foundation of this review on deep learning-based multi-modal fake news detection.

References

[1]. Shang, W. *et al.* (2026) ‘Semantic space aligned multimodal fake news detection’, *Information*

Fusion, 125, p. 103469. doi:10.1016/j.inffus.2025.103469.

[2]. Faria, F.T. *et al.* (2025) 'MultiBanFakeDetect: Integrating Advanced Fusion Techniques for multimodal detection of Bangla fake news in under-resourced contexts', *International Journal of Information Management Data Insights*, 5(2), p. 100347. doi:10.1016/j.jjimei.2025.100347.

[3]. Wang, H. *et al.* (2025) 'SEPM: Multiscale semantic enhancement-progressive multimodal fusion network for fake news detection', *Expert Systems with Applications*, 283, p. 127741. doi:10.1016/j.eswa.2025.127741.

[4]. Singh, R., Suyal, H., & Digra, M. (2024, December). An approach to detect fake news based on machine learning. In *2024 Eighth International Conference on Parallel, Distributed and Grid Computing (PDGC)* (pp. 239-243). IEEE.

[5]. Guo, Y. *et al.* (2025) 'CAMFND: Cross-modal adaptive-aware learning for multimodal fake news detection', *Pattern Recognition Letters*, 195, pp. 1–7. doi:10.1016/j.patrec.2025.02.035.

[6]. Jiang, Y. and Wang, Y. (2025) 'IMFND: In-context multimodal fake news detection with large visual-language models', *Knowledge-Based Systems*, 325, p. 113880. doi:10.1016/j.knosys.2025.113880.

[7]. Alrashoud, M. (2025) 'Deepfake video detection methods, approaches, and challenges', *Alexandria Engineering Journal*, 125, pp. 265–277. doi:10.1016/j.aej.2025.04.007.

[8]. Shu, K. (2022) 'Combating disinformation on social media: A computational perspective', *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 2(1), p. 100035. doi:10.1016/j.tbench.2022.100035.

[9]. Rao, S., Verma, A.K. and Bhatia, T. (2021) 'A review on Social Spam Detection: Challenges, open issues, and Future Directions', *Expert Systems with Applications*, 186, p. 115742. doi:10.1016/j.eswa.2021.115742.

[10]. Das, B. and TSB, S. (2023) 'Multi-contextual learning in disinformation research: A review of challenges, approaches, and opportunities', *Online Social Networks and Media*, 34–35, p. 100247. doi:10.1016/j.osnem.2023.100247.

[11]. Elbatanouny, H. *et al.* (2025) 'A comprehensive analysis of Deception Detection Techniques Leveraging Machine Learning', *Expert Systems with Applications*, 283, p. 127601. doi:10.1016/j.eswa.2025.127601.

[12]. Wang, H. *et al.* (2025a) 'SEPM: Multiscale semantic enhancement-progressive multimodal fusion network for fake news detection', *Expert Systems with Applications*, 283, p. 127741. doi:10.1016/j.eswa.2025.127741.

[13]. Li, Xianghua *et al.* (2025) 'A survey of Multimodal Fake News Detection: A cross-modal interaction perspective', *IEEE Transactions on Emerging Topics in Computational Intelligence*, 9(4), pp. 2658–2675. doi:10.1109/tetci.2025.3543389.

[14]. Zeqi, G. *et al.* (2025) 'Task-oriented multi-domain adversarial network for fake news detection', *Applied Soft Computing*, 177, p. 113227. doi:10.1016/j.asoc.2025.113227.

[15]. Chen, X. *et al.* (2025) 'Enhancing text-centric fake news detection via external knowledge distillation from LLMS', *Neural Networks*, 187, p. 107377. doi:10.1016/j.neunet.2025.107377.

[16]. Dai, R. *et al.* (2025) 'A unified cross-source context enhancement model for multi-source fake news detection', *Knowledge-Based Systems*, 324, p. 113867. doi:10.1016/j.knosys.2025.113867.

[17]. Yu, K., Jiao, S. and Ma, Z. (2025) 'Fake news detection based on Bert multi-domain and Multi-modal fusion network', *Computer Vision and Image Understanding*, 252, p. 104301. doi:10.1016/j.cviu.2025.104301.

[18]. Zhang, L. *et al.* (2024) *Knowledge-aware multimodal pre-training for fake news detection* [Preprint]. doi:10.2139/ssrn.4805672.

[19]. Lv, H. *et al.* (2025) *MDF-FND: A dynamic fusion model for multimodal fake news detection* [Preprint]. doi:10.2139/ssrn.5081768.

[20]. Jiang, Y. and Wang, Y. (2025) 'IMFND: In-context multimodal fake news detection with large visual-language models', *Knowledge-Based Systems*, 325, p. 113880. doi:10.1016/j.knosys.2025.113880.

[21]. Chen, H. *et al.* (2025) 'Multi-modal robustness fake news detection with cross-modal and Propagation Network Contrastive Learning', *Knowledge-Based Systems*, 309, p. 112800. doi:10.1016/j.knosys.2024.112800.