International Journal of Scientific Modern Research and Technology
(Volume: 24, Issue: 01, Number: 11) Paper ID: IJSMRT - 26070111
ISSN: 2582-8150

Misinformation Detection on Social Media: From Machine Learning and Transformers to Multimodal, Graph-Based and Explainable AI

Ajay Dixit
Research Scholar
Department of IT, NIIST, Bhopal, Madhya Pradesh, India

Madhuvan Dixit
Head and Professor
Department of IT, NIIST, Madhya Pradesh, India

Abstract: The scale, velocity and multimodal nature of contemporary social media have made misinformation detection a difficult problem at the intersection of natural language processing, network science, computer vision, information retrieval and human-centered artificial intelligence. This critical review synthesizes the evolution of automated misinformation detection from handcrafted linguistic features and conventional machine learning to convolutional and recurrent neural networks, pretrained Transformers, graph neural networks (GNNs), multimodal fusion and emerging large-language-model-assisted verification. Particular attention is given to the distinction between content classification and evidence-grounded verification, because high benchmark accuracy does not necessarily imply factual reasoning or cross-domain robustness. The review compares commonly used datasets, feature families, architectures, fusion strategies, explainability methods and evaluation protocols, and identifies recurrent methodological weaknesses including label ambiguity, dataset leakage, source and temporal confounding, class imbalance, limited multilingual coverage, weak cross-dataset validation and over-reliance on accuracy. A literature-derived synthesis framework is presented in which contextual Transformer representations are complemented by CNN-based local features, BiLSTM sequential features and multi-head attention-based fusion, while calibrated prediction and post-hoc or intrinsic explanations support transparent decision making. The review argues that the next generation of misinformation detectors should move beyond closed-set binary classification toward temporally robust, multilingual, multimodal and evidence-aware systems that quantify uncertainty, retrieve external evidence, withstand adversarial paraphrasing and provide explanations that can be audited by human fact-checkers. A deployment-oriented evaluation ladder and research agenda are proposed to guide reproducible future studies.

Keywords: misinformation detection; fake news; social media; Transformer; BERT; CNN; BiLSTM; graph neural network; multimodal learning; explainable AI; large language model; fact checking.

How to cite this article: Ajay Dixit, Madhuvan Dixit. (2026). Misinformation Detection on Social Media: From Machine Learning and Transformers to Multimodal, Graph-Based and Explainable AI, International Journal of Scientific Modern Research and Technology (IJSMRT), ISSN: 2582-8150, Volume-24, Issue-01, Number-11, July- 2026, pp. 90 - 104, URL: https://www.ijsmrt.com/wp-content/uploads/2026/09/IJSMRT-26070111.pdf

IJSMRT-26070111

1. Introduction

Social media has reduced the cost and latency of information production and distribution to a historically unprecedented level. This democratization of communication supports rapid access to eyewitness reports, public-health information and civic participation, but it also creates an environment in which false, misleading, decontextualized or manipulated claims can circulate before professional verification is available. Misinformation is therefore not merely a content-quality problem: it is a socio-technical phenomenon shaped by platform design, human cognition, network structure and automated amplification. Influential studies have shown that false information can diffuse differently from verified information and that these differences cannot be attributed solely to bots [1,2].

 ISSN: 2582-8150

Terminology requires care. Misinformation usually denotes false or misleading information regardless of intent, whereas disinformation implies intentional deception. A rumor is information whose truth status is unresolved when it circulates, while the expression 'fake news' is often used operationally for labeled datasets even though it is conceptually imprecise. Satire, clickbait, propaganda and manipulated media can overlap with misinformation but are not equivalent. This distinction matters because a model trained on short political fact-check statements is solving a different problem from a model trained on conversational rumor threads, multimodal Reddit posts or source-credibility data [3,4].

Early computational approaches relied on lexical statistics, bag-of-words, TF-IDF, sentiment, readability, punctuation, source metadata and manually engineered credibility features. Such systems remain valuable as transparent baselines, but they struggle with context dependence and semantic ambiguity. Deep learning reduced dependence on manual feature engineering: CNNs learned local phrase patterns, recurrent neural networks modeled sequence order, and attention mechanisms emphasized informative portions of text. The Transformer subsequently replaced recurrence with self-attention and enabled large-scale contextual pretraining [9]. BERT demonstrated that a deeply bidirectional pretrained encoder can be fine-tuned for downstream classification tasks [10], while RoBERTa showed that pretraining data and optimization choices strongly influence results [11].

Misinformation detection has also moved beyond text-only classification. Social context, user behavior, reply structure and propagation networks provide signals that are naturally modeled as graphs. FakeNewsNet was specifically designed to combine news content with social and spatiotemporal context [7], while Fakeddit provides more than one million multimodal samples with several label granularities [8]. GNN surveys now treat graph construction and propagation modeling as central design choices [18,19]. Multimodal reviews likewise show increasing interest in text-image fusion, video and audio analysis, multilingually and explainability [20].

The newest stage of the field includes large language models (LLMs), retrieval-augmented generation (RAG), multimodal foundation models and explanation-oriented systems. These technologies create opportunities for zero-shot transfer, claim decomposition, evidence retrieval and natural-language justification. However, they also introduce hallucination, prompt sensitivity, computational cost and calibration concerns. A fluent explanation can be persuasive without being faithful to the model or grounded in reliable evidence. Consequently, progress should be assessed not only by in-domain accuracy but also by temporal robustness, cross-domain generalization, calibration, adversarial resistance, evidence quality and human auditability.

This review contributes a unified, deployment-oriented perspective. It first organizes the field by the type of information available to the model—content, context, propagation, multimodal and external evidence. It then traces the methodological progression from classical machine learning through deep neural networks, Transformers, GNNs and LLM-assisted verification. Common benchmark datasets and evaluation practices are critically examined, with emphasis on leakage and confounding. Finally, a literature-derived Transformer-CNN-BiLSTM-attention fusion architecture is proposed as a research template, and a future agenda is developed around multilinguality, multimodal evidence, uncertainty estimation, continual learning and human-centered evaluation.

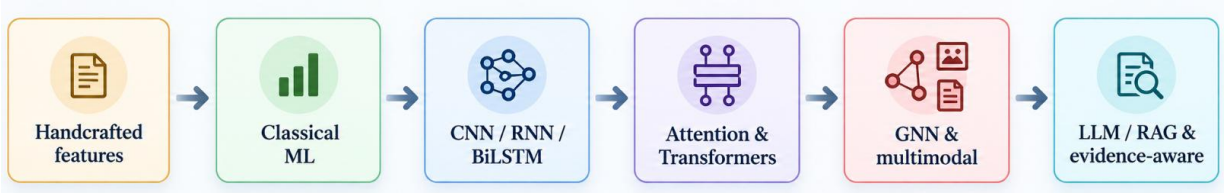

Figure 1. Evolution of computational misinformation detection from handcrafted features to evidence-aware foundation-model pipelines.

 ISSN: 2582-8150

2. Problem Definition and Terminology

A scientifically useful definition of the task must state what is being predicted, at what unit of analysis and under what evidence conditions. Post-level classification assigns a label to an individual post; event-level detection aggregates several posts about the same event; source-level assessment estimates credibility of publishers or accounts; rumor verification predicts the status of a conversational claim; and evidence-based fact checking assesses whether external evidence supports, refutes or leaves a claim unresolved. These tasks should not be merged into a single leaderboard because their labels, inputs and error costs differ.

Table 1. Key terminology and its implications for experimental design.

Term	Operational meaning	Typical labels	Methodological caution
Misinformation	False or misleading content, irrespective of intent	true/false; reliable/unreliable	Truth status may be uncertain or time-dependent
Disinformation	False or misleading content produced or spread intentionally	disinformation/non-disinformation	Intent is difficult to infer from content alone
Rumor	Unverified claim circulating socially	true/false/unverified/non-rumor	Ground truth may be assigned after circulation
Fake news	Dataset-oriented label for fabricated or misleading news	fake/real or fine-grained truthfulness	Definitions differ substantially across datasets
Fact checking	Evidence-based assessment of a claim	supported/refuted/not enough information	Requires reliable retrieval and provenance

3. Review Methodology

This manuscript is a structured critical review rather than a meta-analysis. The literature synthesis was organized around peer-reviewed studies and authoritative benchmark papers that address automated misinformation, fake-news or rumor detection on social-media or closely related online-news data. The review emphasizes methodological development, dataset design, evaluation validity and deployment readiness rather than pooling heterogeneous accuracy values.

Search strategy

A reproducible search can be executed in Scopus, Web of Science Core Collection, IEEE Xplore, ACM Digital Library, ScienceDirect and SpringerLink using a query of the form: ('misinformation' OR 'disinformation' OR 'fake news' OR 'rumour' OR 'rumor') AND ('social media' OR 'social network' OR 'microblog*') AND ('detection' OR 'classification' OR 'verification') AND ('machine learning' OR 'deep learning' OR 'transformer' OR 'BERT' OR 'graph neural network' OR 'multimodal' OR 'large language model'). Database-specific syntax should be documented together with the exact final search date. Backward and forward citation chasing is useful for benchmark datasets and seminal architectures.

Eligibility and critical appraisal

Eligible evidence includes full peer-reviewed journal or conference studies that evaluate computational detection or provide systematic reviews of relevant methods. Purely opinion-based commentary, work without sufficient methodological detail and studies unrelated to veracity or credibility should be excluded. For empirical papers, critical appraisal should record dataset provenance, labeling method, duplicate control, train/validation/test separation, class distribution, baseline adequacy, hyperparameter reporting, ablation analysis, code availability, temporal or cross-domain testing and explanation evaluation. These criteria are important because headline performance can be inflated by source leakage, duplicate content, topic overlap or random splits that do not mimic

deployment. Recent quality-stratified synthesis has specifically highlighted reproducibility and generalizability weaknesses in rumor-detection literature [25].

Because a formal PRISMA flow requires verifiable record counts from an actual database export, this article does not invent identification, screening or exclusion counts. If a target journal requires a systematic review label, the final authors should execute the above search, archive the records and add a PRISMA 2020 flow diagram with exact counts and reasons for full-text exclusion. The current review remains camera-ready as a critical/narrative review without such fabricated counts.

4. Taxonomy of Detection Signals

The most useful taxonomy begins with the evidence available to the model rather than the model name. Five broad information sources recur across the literature: content, context, propagation, multimodal signals and external evidence. Each source supports different inferences and fails under different conditions.

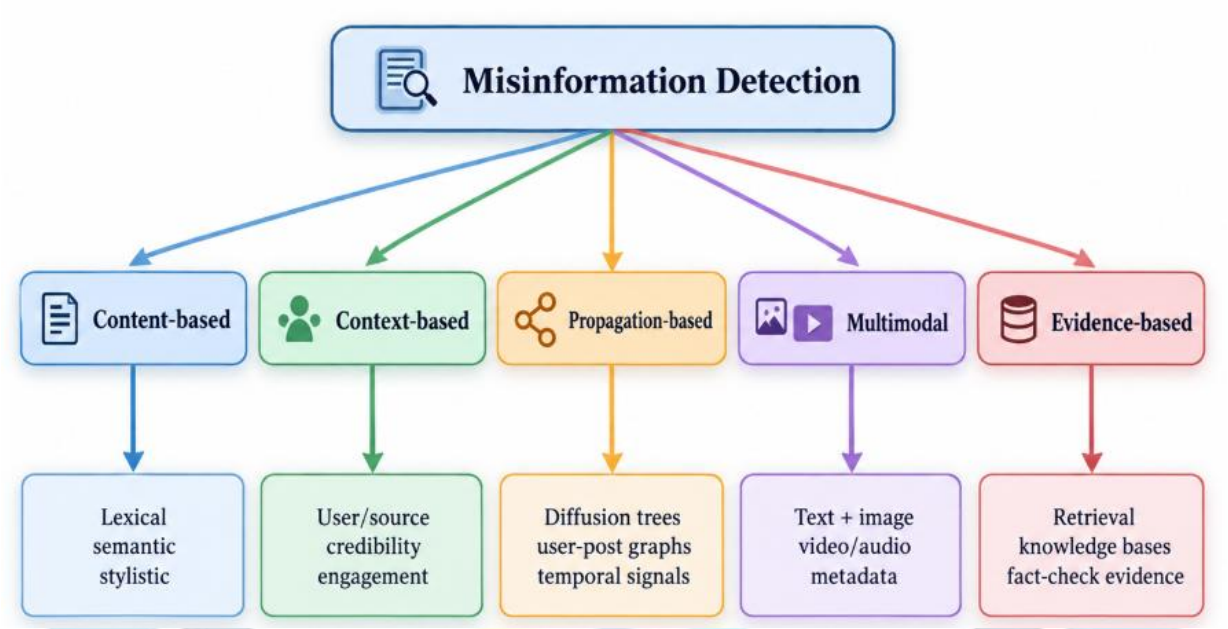

Figure 2. Taxonomy of information sources used for automated misinformation detection.

Content-based signals

Content-based methods analyze the linguistic, semantic, stylistic or visual properties of the item itself. Classical textual features include TF-IDF, character and word n-grams, readability, punctuation, lexical diversity, sentiment and topic distributions. Neural encoders learn representations directly from tokens. Their major advantage is that they can operate immediately after publication, before social reactions accumulate. Their limitation is epistemic: a false claim can be linguistically indistinguishable from a true claim unless the system has access to relevant world knowledge or evidence.

Context and source signals

Context-based methods use publisher, author or user information, engagement statistics, reply content, account histories and community interactions. Early credibility research on Twitter showed that message-level and user-level cues can support automatic credibility assessment [5]. Context often improves accuracy, but it is vulnerable to shortcut learning. If one source appears mainly in one class, a classifier may learn source identity instead of transferable misinformation characteristics.

 ISSN: 2582-8150

Propagation and graph signals

Propagation-based systems model repost chains, conversation trees or heterogeneous user-post-source networks. These signals capture how claims diffuse, which communities amplify them and how interactions evolve over time. GNNs are well suited to non-Euclidean relational data, but strong propagation cues may only become available after substantial diffusion. Therefore early detection should be evaluated as a function of elapsed time or observed interactions, not only on complete propagation graphs.

Multimodal signals

Multimodal systems combine text with images, video, audio or metadata. The core challenge is often cross-modal consistency rather than independent modality classification. A genuine photograph can be paired with a false caption, or an old video can be presented as a new event. Effective fusion must therefore determine when modalities support one another, conflict or are missing. Recent systematic synthesis confirms that Transformer, recurrent and convolutional architectures dominate much of the multimodal literature, while multilinguality, explainability and real-time operation remain open needs [20].

External evidence

Evidence-based systems retrieve documents, knowledge-base facts or fact-check articles and then determine whether the evidence supports or contradicts the claim. This moves the problem closer to verification than pattern classification. The resulting reliability depends on both retrieval quality and reasoning quality. Provenance should therefore be exposed to users, and retrieval errors should be measured separately from final classification errors.

5. Traditional Machine-Learning Approaches

Traditional approaches represent posts or articles with manually selected features and train classifiers such as logistic regression, support-vector machines (SVM), Naïve Bayes, random forests or gradient boosting. TF-IDF remains a strong baseline because misinformation datasets often contain distinctive vocabulary. For term t in document d, TF-IDF can be written as $TF(t,d) \times \log(N/DF(t))$, where N is the corpus size and DF(t) is the number of documents containing t.

The strengths of classical models are low computational cost, modest data requirements and interpretability. Linear models make it relatively easy to inspect feature coefficients, while tree ensembles support feature-importance analyses. However, performance depends heavily on representation quality. Bag-of-words features do not naturally model negation, long-range dependencies or word sense, and handcrafted features can encode dataset-specific artifacts. If fake and real samples originate from different publishers, a random split may distribute publisher vocabulary across training and test sets and create an unrealistically easy problem. Traditional models remain scientifically essential as baselines. A proposed Transformer or hybrid architecture should outperform a well-tuned TF-IDF + linear classifier under the same leakage-controlled split; otherwise the added complexity may have limited practical justification. Reporting only weak baselines can exaggerate novelty.

6. Deep Learning: CNNs, RNNs, BiLSTMs and Attention

CNNs apply learnable filters over local windows of token embeddings. In text classification, a one-dimensional convolution can detect local phrase patterns analogous to learned n-grams, while max pooling selects high-activation features. Multiple kernel sizes capture patterns at different local scales. CNNs are parallelizable and computationally efficient but may miss long-range dependencies unless stacked deeply or combined with other encoders.

RNNs process tokens sequentially and maintain a hidden state that summarizes preceding information. LSTM and GRU architectures introduce gates that mitigate vanishing gradients, while bidirectional LSTMs (BiLSTMs) process text in forward and backward directions so each position can incorporate both left and right context. Hybrid systems

such as CSI combined text, temporal user response and source behavior, demonstrating the value of integrating content and social signals rather than treating them independently [13].

Attention mechanisms address the fact that not every token or hidden state contributes equally. Given states h_i and attention scores e_i, normalized weights are α_i = exp(e_i)/Σ_j exp(e_j), and a context vector is c = Σ_i α_i h_i. The mechanism can improve predictive performance and produces intuitive visualizations. However, a high attention weight is not automatically a causal explanation; faithfulness must be tested through perturbation or counterfactual analysis.

Hybrid CNN-LSTM architectures remain competitive and can complement Transformer representations. Hashmi et al. combined FastText with machine/deep learning models and explainability, while modern hybrid studies increasingly add pretrained Transformers and attention [22]. The methodological lesson is that hybridization should be justified by ablation: each branch must demonstrate incremental value under the same data split and optimization budget.

7. Transformer-Based Misinformation Detection

The Transformer replaced recurrence with self-attention, allowing direct interactions among all positions in a sequence [9]. The core scaled dot-product attention operation is:

$$\text{Attention}(Q,K,V) = \text{softmax}(QK^T / \sqrt{d_k})\, V$$

Multi-head attention computes several attention functions in parallel and concatenates their outputs, enabling different heads to learn distinct relational patterns. BERT introduced deep bidirectional pretraining with masked-language modeling and showed that a general encoder can be adapted to classification by fine-tuning [10]. RoBERTa subsequently demonstrated how strongly performance depends on pretraining corpus size, masking strategy and optimization choices [11].

For misinformation detection, pretrained Transformers offer two major advantages: contextual word representations and transfer learning. A token such as 'positive' is represented differently in a medical test, a financial outlook and everyday sentiment. Fine-tuning can therefore capture semantic subtleties that static embeddings miss. FakeBERT combined BERT with parallel CNN blocks and reported strong performance, illustrating the continuing appeal of Transformer-CNN hybrids [12]. Reviews focused specifically on Transformer rumor detection document a wide range of microblogging datasets, fine-tuning strategies and classification heads [17].

Nevertheless, contextual language modeling is not fact verification. A Transformer classifier can learn correlations between linguistic style and dataset labels without accessing evidence. This creates a central distinction between detecting content that resembles historical misinformation and determining whether a new claim is factually correct. Transformers also remain vulnerable to domain shift, source artifacts and temporal drift. Cross-dataset, source-held-out and chronological evaluation should therefore accompany random-split results.

Computational cost is another limitation. Full fine-tuning of large encoders requires GPU memory and can be unstable on small datasets. Lightweight encoders, distillation, parameter-efficient fine-tuning and frozen-feature approaches can reduce cost, but their accuracy-efficiency trade-off should be reported. A deployment paper should include parameter count, training time, inference latency and memory consumption rather than accuracy alone.

8. Graph Neural Networks and Propagation-Aware Detection

Social-media misinformation is intrinsically relational. Users follow, reply to and repost one another; posts refer to sources; and communities form temporal patterns around claims. A graph G=(V,E) represents entities as nodes and relationships as edges. Node features may include contextual text embeddings, user metadata, temporal features or source credibility. GNN layers update node representations by aggregating messages from neighbors, allowing a classifier to use both content and structure.

Research includes graph convolutional networks, graph attention networks, GraphSAGE, graph autoencoders and heterogeneous GNNs. Surveys by Phan et al. and Lakzaei et al. show that graph-based models can integrate content, social context and propagation while also emphasizing challenges in graph construction, scalability and transfer [18,19]. GCAN, for example, combined graph-aware modeling and co-attention to identify suspicious retweeters and textual evidence, explicitly linking performance with explanation [14].

A critical deployment issue is timing. Complete propagation graphs are unavailable when early intervention would be most valuable. Evaluation should therefore report how performance changes after the first few minutes, hours or number of interactions. Dynamic and temporal GNNs can model evolving graphs, but they increase reproducibility demands and computational complexity. Privacy is also important because user-level graphs can expose sensitive behavioral patterns; models should avoid unnecessary identity features and test whether predictions depend on protected or demographic proxies.

9. Multimodal Misinformation Detection

Modern misinformation is frequently multimodal. A false caption can be attached to an authentic image, a legitimate video can be recycled out of context, or synthetic media can be embedded in a deceptive narrative. The model must therefore reason both within and across modalities. Text encoders are commonly paired with CNNs or Vision Transformers for images, while fusion is performed by concatenation, gating, bilinear pooling or cross-attention.

Simple concatenation assumes that each modality is equally reliable, whereas gated or attention-based fusion can assign input-dependent weights. Cross-attention is particularly attractive because it can model whether textual entities align with visual regions. Fakeddit provides more than one million Reddit-derived text-image samples with 2-way, 3-way and 6-way labels [8], while FakeNewsNet supports experiments that combine content with social and spatiotemporal information [7].

Multimodal benchmarks introduce unique leakage risks. Near-duplicate images can cross train/test partitions; publisher logos or watermarks may correlate with labels; and missing modalities can become shortcuts if missingness differs by class. Robust studies should apply perceptual duplicate detection, source-held-out splits and missing-modality stress tests. They should also distinguish manipulated-media detection from contextual misinformation: a genuine image paired with a false claim is misinformation even when forensic image manipulation detectors correctly identify the pixels as authentic.

A 2025 systematic review covering multimodal fake-news detection identified Transformers and recurrent architectures as prominent model families and highlighted multilingual systems, hybridization, XAI and real-time detection as priorities [20]. More recent multilingual-multimodal work has combined Transformer text representations, CNN-BiLSTM components, image encoders, GraphSAGE and multi-head attention, illustrating how formerly separate architectures are converging [24].

10. Explainability, Calibration and Trustworthy AI

In high-stakes settings a detector should communicate more than a class label. At minimum, it should provide a calibrated probability or uncertainty estimate, identify influential evidence and expose external sources when factual verification is claimed. Explainability methods can be intrinsic, such as sparse rationales or evidence selection, or post-hoc, such as LIME and SHAP.

LIME fits a simple local surrogate around an individual prediction after perturbing the input [15]. SHAP uses Shapley-value principles to assign feature contributions and unifies a family of additive attribution methods [16]. In text models, both can highlight influential words or phrases; in multimodal systems, attributions can be computed for tokens, image regions or modality-level features. However, explanation plausibility and explanation faithfulness are distinct. A rationale may look convincing to a human without reflecting the actual mechanism driving the prediction.

Faithfulness should be tested. Useful methods include deletion/insertion curves, comprehensiveness, sufficiency and counterfactual sensitivity. Human evaluation can measure whether explanations help fact-checkers reach better decisions, but evaluator expertise, sample selection and inter-rater agreement should be reported. Attention maps alone should not be presented as definitive explanations unless their causal relevance is validated.

Calibration is equally important. Softmax confidence is not automatically a probability of correctness. Expected calibration error, Brier score and reliability diagrams can reveal overconfidence. Selective prediction can allow a system to abstain on uncertain cases and return 'needs verification' rather than force an unjustifiably confident binary label. This is especially appropriate for emerging claims where ground truth or evidence is incomplete.

11. Benchmark Datasets and Their Implications

Benchmark choice defines what the model learns. LIAR contains 12.8K manually labeled short political statements with six truthfulness categories and associated metadata [6]. FakeNewsNet integrates article content with social and spatiotemporal context [7]. Fakeddit provides large-scale multimodal Reddit data [8]. PHEME-style rumor datasets focus on conversational threads and veracity rather than article-level fake news. These datasets are valuable but not interchangeable.

Table 2. Representative misinformation datasets and methodological cautions.

Dataset	Primary domain / platform	Modalities	Label structure	Key methodological issue
LIAR [6]	Political fact-check statements	Text + metadata	Six truthfulness levels	Short statements; political-domain transfer is uncertain
FakeNewsNet [7]	PolitiFact / GossipCop with social context	Text + user/social + temporal	Fake/real	Source/event overlap can create leakage
Fakeddit [8]	Reddit	Text + image + metadata	2-, 3-, 6-way	Distant supervision; subreddit and visual artifacts
PHEME / RumourEval	Twitter/X conversation threads	Text + reply tree + stance	Rumor/veracity classes	Thread structure differs from article classification
Weibo rumor sets	Chinese microblogging	Text + social + propagation	Rumor/non-rumor	Platform and language specificity
Custom fact-check corpora	News/social posts linked to fact-checkers	Usually text + source	Variable	Reproducibility depends on archived content and label rules

No single dataset captures the full misinformation problem. A robust study should ideally include an in-domain benchmark and at least one external evaluation with a different platform, source distribution or time period. Dataset documentation should state collection dates, label provenance, duplicate handling, known biases, licensing, missing modalities and intended use. Performance claims should be scoped to the conditions represented by the data.

12. Representative Method Families

Table 3. Comparative characteristics of major model families.

Family	Core representation	Main strength	Main limitation	Best use case
TF-IDF + linear ML	Sparse lexical features	Fast, transparent, strong baseline	Weak contextual semantics	Resource-constrained text classification

Family	Core representation	Main strength	Main limitation	Best use case
CNN	Local convolutional patterns	Efficient local feature extraction	Limited long-range context	Phrase-level stylistic cues
BiLSTM / GRU	Sequential hidden states	Ordered contextual dependencies	Slower, less parallel than Transformer	Sequence-sensitive text and conversations
Transformer	Contextual self-attention embeddings	Strong transfer learning and semantics	Compute cost; may learn shortcuts	Content-rich text classification
GNN	User/post/source graph embeddings	Captures propagation and social relations	Requires graph data; early-detection tension	Propagation-aware rumor/fake-news detection
Multimodal fusion	Text + image/video/audio	Captures cross-modal inconsistency	Missing/noisy modalities; duplicate leakage	Media-rich misinformation
LLM + RAG	Claim + retrieved evidence	Evidence synthesis and zero/few-shot flexibility	Hallucination, retrieval dependence, cost	Evidence-grounded verification

13. Evaluation Metrics and Experimental Design

Accuracy is intuitive but can be misleading when classes are imbalanced. Precision measures the proportion of predicted positive instances that are correct; recall measures the proportion of actual positive instances detected; F1 is their harmonic mean. For binary classification:
Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

F1 = 2 × Precision × Recall / (Precision + Recall)

Macro-F1 is generally more informative than accuracy when class frequencies differ because each class receives equal weight. ROC-AUC summarizes threshold-independent ranking, but PR-AUC is often preferable when the positive class is rare. When probabilities are exposed to users, calibration measures such as Brier score and expected calibration error should be included. The split strategy can matter more than the architecture. Random splits can overestimate performance when duplicate posts, publishers, users or events appear in both training and test sets. Temporal splits evaluate concept drift; source-held-out splits test publisher generalization; event-held-out splits test topic transfer; and cross-dataset evaluation tests portability across collection procedures. These increasingly difficult settings form an evaluation ladder rather than mutually exclusive alternatives.

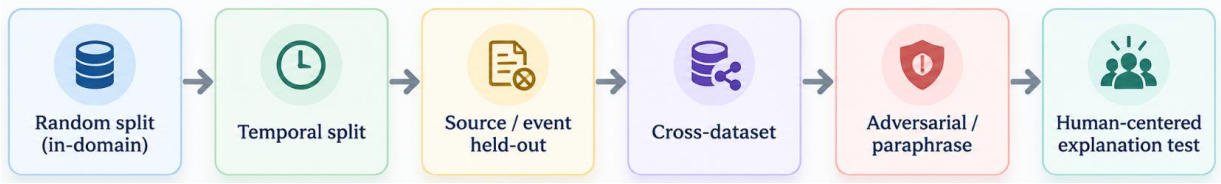

Figure 3. Deployment-oriented evaluation ladder for misinformation detection research.

Strong experimental reports should include multiple random seeds, confidence intervals or dispersion statistics, hyperparameter selection rules, model size, training cost, inference latency and ablation studies. Statistical

significance should be interpreted together with practical significance. A small in-domain accuracy gain may not justify a much larger model if cross-domain robustness, calibration and latency do not improve.

14. Literature-Derived Attention-Based Hybrid Framework

The literature suggests that different modules provide complementary inductive biases. A pretrained Transformer is strong at contextual semantics, CNNs capture local phrase patterns, BiLSTMs explicitly preserve sequential order, and attention-based fusion can learn input-dependent weighting. This motivates a hybrid research framework, illustrated in Figure 4. It is presented as a synthesis hypothesis derived from prior work—not as an experimentally validated model in this review.

Figure 4. Literature-derived Transformer-CNN-BiLSTM-attention fusion framework for future empirical validation.

Transformer encoder

A pretrained encoder such as BERT, RoBERTa or DeBERTa converts tokenized input into contextual embeddings H_T. The maximum sequence length should be selected from the empirical length distribution rather than chosen arbitrarily. Full fine-tuning, partial freezing and parameter-efficient fine-tuning should be compared because they differ in compute and overfitting risk.

CNN and BiLSTM branches

Parallel one-dimensional CNN filters with kernel sizes such as 3, 4 and 5 can transform H_T into local feature vectors F_CNN. A BiLSTM branch consumes the same contextual sequence and produces F_BiLSTM, which encodes forward and backward dependencies. Dropout and normalization should be reported explicitly.

Attention-based fusion

After projecting branch outputs to a common dimensionality, fusion can be learned rather than fixed. If F_i denotes a branch feature and e_i its learned score, normalized weights can be computed as $\alpha_i = \exp(e_i) / \Sigma_j \exp(e_j)$, followed by $F_{fused} = \Sigma_i \alpha_i F_i$. Multi-head fusion allows several distinct interactions among branches. Simple concatenation should be retained as a baseline to demonstrate that attention adds value.

Classification and explanation

A dense classifier with dropout maps F_fused to logits and class probabilities. Temperature scaling or another post-hoc calibration method can be fitted on a validation set. The user-facing output should include the predicted class, calibrated confidence, influential tokens and—if retrieval is used—the evidence documents. Ablations must remove each branch independently to determine whether the hybrid architecture is genuinely complementary or merely larger.

15. Large Language Models and Evidence-Grounded Verification

LLMs expand misinformation detection beyond supervised classification. They can act as zero-shot or few-shot classifiers, generate synthetic training examples, decompose complex claims and summarize retrieved evidence.

Recent reviews show rapidly increasing integration of LLMs with graph, multimodal and retrieval-based approaches [25]. This flexibility is valuable for emerging events where labeled data are scarce.

The central weakness is that an LLM's internal knowledge is neither complete nor guaranteed current. Asking a language model whether a claim is true can conflate linguistic plausibility with factual correctness. RAG addresses part of this problem by retrieving external documents and conditioning the model on evidence. A robust pipeline is: claim extraction → retrieval → evidence ranking → entailment/contradiction assessment → calibrated decision → cited explanation.

Evaluation should decompose this pipeline. Retrieval recall@k or nDCG measures whether relevant evidence is found; entailment or verification metrics assess reasoning; citation correctness checks whether the cited source actually supports the generated explanation. This makes failure analysis actionable: an error can arise because evidence was not retrieved, because correct evidence was misinterpreted, or because the claim was underspecified.

LLMs also change the threat landscape by making persuasive paraphrasing, style transfer and synthetic media easier to produce. Future detectors should therefore be stress-tested on AI-generated misinformation, multilingual rewrites, adversarial paraphrases and coordinated text-image manipulation instead of assuming that future misinformation resembles historical benchmark data.

16. Persistent Challenges and Future Research Agenda

Domain and temporal generalization

Most benchmark datasets represent a narrow set of events, sources and time periods. Models can therefore memorize publisher style, political topics or platform conventions. Future work should prioritize temporal splits, source-held-out evaluation, domain adaptation, continual learning and explicit concept-drift monitoring. Continual updating must be protected against noisy or adversarial labels.

Multilingual and code-mixed misinformation

English remains overrepresented in public benchmarks. Real social-media ecosystems are multilingual and frequently code-mixed. Translation into English can erase culturally specific cues and introduce systematic artifacts. Multilingual encoders, language-adaptive pretraining, cross-lingual retrieval and native-speaker annotation are required. Recent work on multilingual and multimodal fake-news detection underscores the importance of cross-lingual robustness and explanation quality [24].

Multimodal inconsistency and synthetic media

Future systems should distinguish three cases: manipulated media, genuine media used out of context and semantically inconsistent text-media pairs. A unified classifier can hide these mechanistically different errors. Modality-specific confidence and explicit cross-modal consistency scores would support more informative decisions.

Explainability and human factors

Explanations should be evaluated for both faithfulness and usefulness. Human-centered experiments should measure whether explanations improve fact-checking accuracy, reduce time or improve appropriate trust. Systems should support abstention and escalation rather than forcing automation in uncertain cases.

Reproducibility and computational reporting

Reproducibility remains uneven. Studies should report software versions, random seeds, exact splits, preprocessing, tokenizer version, maximum sequence length, training epochs, learning-rate schedules, early stopping, hardware and energy or compute proxies where feasible. Code and split files are often more important for reproducibility than architecture diagrams alone.

 ISSN: 2582-8150

Table 4. Research gaps and deployment-oriented evaluation priorities.

Research gap	Why it matters	Recommended experiment	Preferred outcome
Temporal drift	Narratives and vocabulary change	Train on earlier period; test on later period	Stable Macro-F1 and calibration over time
Source shortcut learning	Model may memorize publishers	Source-held-out split	Performance maintained on unseen sources
Cross-platform transfer	Platform norms differ	Train on platform A; test on platform B	Graceful degradation; effective adaptation
Multilingual / code-mixed data	English-only systems exclude major populations	Native multilingual test sets	Balanced performance across languages
Adversarial paraphrasing	Surface cues can be deliberately changed	LLM/human paraphrase stress test	Small robustness loss
Explainability	Users need auditable decisions	Faithfulness + human utility study	Explanations improve decisions without inducing over-trust
Evidence grounding	Classification is not verification	RAG with provenance and citation checks	Correct evidence retrieval and calibrated verification

17. Positioning Against Recent Reviews

Several recent reviews demonstrate both the maturity and fragmentation of the field. The 2024 IEEE Access systematic review by Alnabhan and Branco emphasizes deep-learning strategies, datasets, transfer learning and class imbalance [21]. Alghamdi et al. provide a broad machine-learning survey that spans theory, datasets and model families [23]. Anggrainingsih et al. focus specifically on Transformer-based rumor detection on microblogging platforms [17], while Phan et al. and Lakzaei et al. focus on GNNs [18,19]. Nasser et al. synthesize multimodal deep-learning studies through 2025 [20]. Bashaddadh et al. review machine- and deep-learning studies from 2020–2024 and highlight interpretability, generalizability and few-/zero-shot learning [26]. The 2026 quality-stratified review extends the scope toward LLM-GNN hybrids and explicitly critiques leakage and benchmark overfitting [25].

The present review differs in emphasis. Rather than treating model families as isolated tracks, it interprets them as complementary information-processing components and links them to the evidence available at deployment. It also foregrounds the boundary between classification and verification, recommends calibration and abstention, and proposes an evaluation ladder that increases distribution shift in stages. This perspective is intended to help researchers design experiments that demonstrate not only higher benchmark scores but also credible real-world value.

18. Recommended Reporting Checklist for New Studies

✓ Define misinformation, disinformation, rumor or fake news operationally for the dataset.

✓ Report dataset collection period, platform, label source, class balance and duplicate-control procedure.

✓ Provide exact train/validation/test split logic and test at least one leakage-resistant split.

✓ Use strong classical and Transformer-only baselines before claiming benefit from a hybrid architecture.

✓ Report precision, recall, Macro-F1 and PR-AUC where appropriate, not accuracy alone.

✓ Include calibration if prediction probabilities are presented to users.

✓ Perform ablation studies for CNN, BiLSTM, attention, graph, metadata and multimodal branches.

✓ Report computational cost: parameters, training hardware, training time and inference latency.

✓ Evaluate cross-domain, temporal, adversarial or multilingual robustness when deployment claims are made.

✓ Evaluate explanation faithfulness and, ideally, usefulness to human fact-checkers.

✓ For RAG/LLM systems, evaluate retrieval quality, citation correctness and hallucination separately.

✓ Release code, preprocessing details and split files whenever licensing permits.

19. Conclusion

Misinformation detection on social media has progressed from handcrafted lexical cues to contextual language models, graph reasoning, multimodal fusion and evidence-aware foundation-model pipelines. This progression has improved representational power, but it has not eliminated the central scientific challenge: distinguishing genuine generalization from exploitation of dataset shortcuts. High random-split accuracy does not demonstrate factual reasoning, cross-platform robustness or deployment readiness. Transformers are now a strong default for text representation, while CNNs, BiLSTMs and attention remain useful when their complementary inductive biases are demonstrated through ablation. GNNs add social and propagation structure, and multimodal architectures address text-image or text-video inconsistency. LLMs and RAG can extend classification toward evidence-grounded verification, but their reliability depends on retrieval quality, provenance, calibration and hallucination control. Future progress should therefore be measured by harder criteria: temporal and source generalization, multilingual and code-mixed performance, multimodal consistency reasoning, robustness to adversarial paraphrasing and AI-generated content, calibrated uncertainty, faithful explanations and human-centered utility. Misinformation detection should evolve from a benchmark classification problem into an auditable decision-support system that can say not only what it predicts, but why, with what evidence and with what degree of uncertainty.

References

[1] Lazer, D.M.J., Baum, M.A., Benkler, Y., Berinsky, A.J., Greenhill, K.M., Menczer, F., Metzger, M.J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S.A., Sunstein, C.R., Thorson, E.A., Watts, D.J., Zittrain, J.L. (2018). The science of fake news. Science, 359(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

[2] Vosoughi, S., Roy, D., Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

[3] Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22–36. https://doi.org/10.1145/3137597.3137600

[4] Zhou, X., Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Computing Surveys, 53(5), Article 109. https://doi.org/10.1145/3395046

[5] Castillo, C., Mendoza, M., Poblete, B. (2011). Information credibility on Twitter. Proceedings of the 20th International Conference on World Wide Web, 675–684. https://doi.org/10.1145/1963405.1963500

[6] Wang, W.Y. (2017). 'Liar, Liar Pants on Fire': A new benchmark dataset for fake news detection. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers), 422–426. https://doi.org/10.18653/v1/P17-2067

[7] Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H. (2020). FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big Data, 8(3), 171–188. https://doi.org/10.1089/big.2020.0062

[8] Nakamura, K., Levy, S., Wang, W.Y. (2020). Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. Proceedings of the 12th Language Resources and Evaluation Conference, 6149–6157. ACL Anthology: https://aclanthology.org/2020.lrec-1.755/

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

[10] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[11] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692.

[12] Kaliyar, R.K., Goswami, A., Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. Multimedia Tools and Applications, 80, 11765–11788. https://doi.org/10.1007/s11042-020-10183-2

[13] Ruchansky, N., Seo, S., Liu, Y. (2017). CSI: A hybrid deep model for fake news detection. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 797–806. https://doi.org/10.1145/3132847.3132877

[14] Lu, Y.-J., Li, C.-T. (2020). GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 505–514. https://doi.org/10.18653/v1/2020.acl-main.48

[15] Ribeiro, M.T., Singh, S., Guestrin, C. (2016). 'Why should I trust you?': Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778

[16] Lundberg, S.M., Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

[17] Anggrainingsih, R., Hassan, G.M., Datta, A. (2024). Transformer-based models for combating rumours on microblogging platforms: A review. Artificial Intelligence Review, 57, 212. https://doi.org/10.1007/s10462-024-10837-9

[18] Phan, H.T., Nguyen, N.T., Hwang, D. (2023). Fake news detection: A survey of graph neural network methods. Applied Soft Computing, 139, 110235. https://doi.org/10.1016/j.asoc.2023.110235

[19] Lakzaei, B., Haghir Chehreghani, M., Bagheri, A. (2024). Disinformation detection using graph neural networks: A survey. Artificial Intelligence Review, 57, 52. https://doi.org/10.1007/s10462-024-10702-9

[20] Nasser, M., Arshad, N., Ali, A., Alhussian, H., Saeed, F., 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. https://doi.org/10.1016/j.rineng.2025.104752

[21] Alnabhan, M.Q., Branco, P. (2024). Fake news detection using deep learning: A systematic literature review. IEEE Access, 12, 114435–114459. https://doi.org/10.1109/ACCESS.2024.3435497

[22] Hashmi, E., Yildirim Yayilgan, S., Yamin, M.M., Ali, S., Abomhara, M. (2024). Advancing fake news detection: Hybrid deep learning with FastText and explainable AI. IEEE Access, 12, 44462–44480. https://doi.org/10.1109/ACCESS.2024.3381038

[23] Alghamdi, J., Luo, S., Lin, Y. (2024). A comprehensive survey on machine learning approaches for fake news detection. Multimedia Tools and Applications, 83, 51009–51067. https://doi.org/10.1007/s11042-023-17470-8

[24] Jadhav, R., Meshram, V., Bhosle, A., Patil, K., Dash, S., Jadhav, S. (2025). Explainable multilingual and multimodal fake-news detection: Toward robust and trustworthy AI for combating misinformation. Frontiers in Artificial Intelligence, 8, 1690616. https://doi.org/10.3389/frai.2025.1690616

[25] Askarizade, M., Davoodijam, E., Tork Ladani, B. (2026). Advancing rumor detection in social media: A quality-stratified systematic review from CNNs to LLM-GNN hybrids. Array, 30, 100842. https://doi.org/10.1016/j.array.2026.100842

[26] Bashaddadh, O.M., Omar, N., Mohd, M., Khalid, M.N.A. (2025). Machine learning and deep learning approaches for fake news detection: A systematic review of techniques, challenges, and advancements. IEEE Access, 13, 90433–90466. https://doi.org/10.1109/ACCESS.2025.3572051

[27] Hu, L., Wei, S., Zhao, Z., Wu, B. (2022). Deep learning for fake news detection: A comprehensive survey. AI Open, 3, 133–155. https://doi.org/10.1016/j.aiopen.2022.09.001

[28] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

[29] Zhou, X., Zafarani, R., Shu, K., Liu, H. (2019). Fake news: Fundamental theories, detection strategies and challenges. Proceedings of the 12th ACM International Conference on Web Search and Data Mining, 836–837. https://doi.org/10.1145/3289600.3291382

[30] Shu, K., Wang, S., Liu, H. (2019). Beyond news contents: The role of social context for fake news detection. Proceedings of the 12th ACM International Conference on Web Search and Data Mining, 312–320. https://doi.org/10.1145/3289600.3290994