Journal of Computing Theories and Applications ISSN:3024-9104

Review Article

Beyond Reported Accuracy: A Verifiability-Focused, Multi-Axis Review of Content and Context Based Fake News Detection

Long Dinh Tuan [1], Le Ngoc An [1,*], and Duong Dinh Thai [2]

[1] Faculty of Information Technology, Hanoi Open University, B101 Nguyen Hien, Bach Mai, Hanoi 10000, Vietnam; e-mail: dinhtuanlong@hou.edu.vn; anln@hou.edu.vn
[2] School of Information and Communication Technology, VNU University of Engineering and Technology, Vietnam National University, Hanoi, 144 Xuan Thuy Street, Cau Giay, Hanoi 10000, Vietnam; e-mail: dinhthaiduong123456@gmail.com
* Corresponding Author : Le Ngoc An

Abstract: Fake news undermines information integrity, and the proliferation of large language models (LLMs) has intensified both its generation and its detection. Prior surveys catalogue methods but rarely verify the performance figures they synthesize, restrict comparison to same-benchmark settings, or assess risk of bias, so reported accuracies are often over-read. This verifiability-focused review analyzes an enumerated, non-exhaustive corpus of 45 content- and context-based detection studies (2018-2025; 42 quantitative-core) in which each reported headline metric is traced to its primary source and confirmed against it wherever the source is accessible (36 of 45; the nine paywalled studies are flagged as reported-but-unconfirmed), and released with the corpus. Studies are grouped into five method families, Transformer/PLM (31.0%), multimodal (26.2%), LLM-based (21.4%), graph/propagation (16.7%), and traditional machine learning (4.8%), and interpreted through a four-axis framework of evidence source, detection time, data dependence, and deployment objective. Without metric conversion or pooling, reported accuracies are strongly heterogeneous and not comparable across families (Transformer 74.8-99.36%, multimodal 82.4-99.97%, graph 84.4-94.36% or 0.75-0.93 AUC, LLM-based 48.6-89.0%); within Weibo, seven multimodal models span 82.4-91.8%, and GPT-4 attains 68.2% (not the ~95% often cited) on LIAR-binary. Three findings follow. First, reported performance appears shaped as much by data and evaluation design as by architecture, so architecture-only interpretations can be misleading; near-perfect results may reflect dataset-specific separability, leakage, or favorable evaluation designs rather than superior transferable capability. Second, the field has evolved cumulatively from static content cues toward multi-source evidence integration and reasoning, so the families are complementary evidence sources rather than exclusive choices. Third, no family is optimal across scenarios, and a central open problem is sustaining reliability under drift, new languages, attacks, and AI-generated content. We derive a dependency-ordered research agenda and, from a Vietnamese case study, transferable design principles for low-resource languages.

Keywords: Fake news detection; Graph neural networks; Large language models; Misinformation; Multimodal learning; Risk of bias; Structured review; Transformer.

Received: July, 14th 2026
Revised: August, 18th 2026
Accepted: August, 28th 2026
Published: September, 11th 2026

1. Introduction

In the digital age, fake news has become one of the most serious challenges to information integrity and social trust. The widespread adoption of social media has dramatically accelerated the dissemination of false information at an unprecedented scale. According to the Global Risks Report 2024, misinformation and disinformation rank among the most severe short-term global risks, particularly as large language models (LLMs) increasingly facilitate the automated generation of deceptive content [1]. The societal consequences are far-reaching. During the COVID-19 pandemic, health-related misinformation caused measurable public-health harm, while in the political domain fake

news has been weaponized to interfere in electoral processes. A landmark study by Vosoughi et al. [2] demonstrated that false news is retweeted roughly 70% more frequently than true news and reaches people about six times faster, underscoring the urgency of automated detection.

Fake news is defined as information intentionally created with false or misleading content and disseminated in the form of news, but not adhering to professional journalistic standards [3]. Research on automated detection has evolved rapidly: from hand-crafted linguistic features and classical machine learning [4], [5], to deep neural networks, and since 2018-2019 to pre-trained Transformer language models such as BERT [6] and RoBERTa [7]. In parallel, two complementary directions matured, multimodal methods that jointly analyze text and images [8]–[10] and context-based methods that model how news propagates through social graphs [11], [12]. Most recently, generative LLMs have emerged in a dual role, acting simultaneously as detectors and as prolific generators of fluent misinformation [13]–[15].

Reported accuracies are high but are obtained on heterogeneous datasets, tasks, and metrics, which makes direct comparison across studies difficult. Reported values also vary with the evaluation setting: on the LIAR benchmark, LLM-based results in a zero-shot setting fall well below the values reported by fine-tuned models on easier corpora (see Section 4.2.4 for the exact figure). This motivates a synthesis that records each reported metric from its primary source and compares methods only under within-benchmark (same-dataset, same-task) conditions.

This review focuses on methods that analyze the content of news items (text and images) and the context of their diffusion (news-propagation structure and social-graph signals). We adopt this content-and-context scope because the field has converged on hybrid models that combine both dimensions. We exclude methods that rely solely on non-textual signal processing or on unlearned, heuristic account-metadata filters (full inclusion/exclusion criteria in Section 3.2). The temporal emphasis is 2021-2025, with a small number of seminal 2018-2020 works retained for architectural lineage.

Rather than only cataloguing methods, this review asks what explains their differences and how the field is evolving. Six research questions guide the study. RQ1: What are the main content- and context-based detection methods in 2021-2025? RQ2: Which data, task, and evaluation-design factors explain the differences in reported performance? RQ3: How do the method families trade off in-domain performance, out-of-domain generalization, explainability, data needs, and computational cost? RQ4: How has the field evolved from static content analysis toward multi-source evidence integration and reasoning? RQ5: How are the open challenges related through shared causes, and which research directions should be prioritized? RQ6: Which lessons from the Vietnamese low-resource case transfer to other low-resource languages? The contributions fall into two groups. The first is methodological, concerning how the evidence is assembled and read:

- An enumerated, source-verified evidence base of 45 studies (42 quantitative-core) in which each reported headline metric is traced to its primary source and confirmed wherever accessible (verification protocol and counts in Section 3.2), released as machine-readable supplementary material.
- A conservative comparison protocol that avoids metric conversion and cross-study pooling and confines direct comparison to within-benchmark (same-dataset, same-task) settings, which revises several performance figures repeated in secondary sources.
- A qualitative appraisal of potential sources of bias in the reported results, connecting near-perfect accuracies to data-design and evaluation weaknesses rather than to model quality.

The second group is conceptual, concerning what the synthesis reveals about the field:

- A multi-axis analytical framework (evidence source, detection time, data dependence, deployment objective) that positions the five method families across dimensions rather than on a single taxonomy, and explains why their performance differs.
- A trade-off map across in-domain performance, generalization, explainability, data needs, cost, and early detection, turned into concrete method-selection guidance.
- A five-stage account of the field's evolution, from feature-centric detection to robust adaptive detection, and a dependency-ordered research agenda derived from it.

- A set of transferable design principles, abstracted from the Vietnamese case, for fake news detection in other low-resource languages.

The remainder of this paper is organized as follows. Section 2 presents background and related surveys. Section 3 describes the methodology, including the search and screening criteria, classification rule, and synthesis approach. Section 4 presents the results: study characteristics, the method families, benchmark datasets, Vietnamese research, and trends and challenges. Section 5 provides a comparative analysis under within-benchmark conditions. Section 6 concludes with findings, recommendations, and limitations.

2. Background and Related Work

2.1. Review Design and Reporting

This paper is a verifiability-focused review rather than an exhaustive systematic literature review. It follows the reporting principles of the PRISMA statement and its 2020 update [16], [17] for transparency and the planning, conducting, and reporting structure of Kitchenham [18], but it does not claim to enumerate the entire literature. Instead, it prioritizes a source-verified evidence base in which each included study's reported metric is confirmed against the primary source (Section 3.2). The quantitative component is a descriptive synthesis and not a statistical meta-analysis; the distinction and its rationale are detailed in Section 3.5. Section 6.4 states the resulting limitations, chiefly that the corpus is enumerated rather than exhaustive and that per-database result counts and a two-reviewer screening log are therefore not reported.

2.2. Conceptual Framework: Content and Context Dimensions

Throughout, we use the following working distinctions. Misinformation is false or misleading content regardless of intent; disinformation is the subset spread with intent to deceive; and fake news denotes fabricated or manipulated news-style content and serves as the umbrella term for the detection task reviewed here. Rumor detection targets unverified claims propagating on social platforms and is often framed around propagation structure, while fact (or claim) verification assesses a specific claim against evidence or a knowledge base. These tasks overlap in data and methods but differ in their unit of analysis (article, post, or claim) and in label semantics; a study is included when its detection target falls within this content- and context-based scope, and its specific task is recorded per study.

Fake news detection methods can be classified along four analytical dimensions [4], [5]. By feature source, content-based methods analyze textual and visual content, context-based methods model propagation and social signals, and hybrid approaches combine both. By model architecture, the field spans traditional machine learning (SVM, Random Forest), deep learning (CNN, RNN, LSTM), Transformer models (BERT [6], RoBERTa [7]), graph neural networks (GCN, GAT), and large language models. By modality, methods are unimodal (text-only or image-only) or multimodal. By learning paradigm, they are supervised, semi-supervised, unsupervised, or few-/zero-shot.

A central methodological observation follows: these dimensions are not mutually exclusive. A multimodal detector built on a BERT text encoder is simultaneously Transformer-based and multimodal; an LLM-based detector is, by construction, Transformer-based; and a graph-based model whose nodes carry textual features combines content and context. Any single-axis tabulation must therefore adopt an explicit assignment rule and disclose the overlap it conceals (Sections 3.5 and 4.2.6).

2.3. Challenges in Fake News Detection

The field faces four interlocking challenge classes. (1) Data: scarcity of high-quality labeled data, class imbalance, collection difficulties under platform privacy policies, and annotation bias. (2) Content: increasingly sophisticated, multi-topic, multilingual, temporally drifting, and increasingly multimodal fake news. (3) Model: difficulty of interpreting deep-learning decisions, overfitting, poor generalization, and high computational cost. (4) Context: the need to model social diffusion, the role of users, and cultural and linguistic differences. These challenges motivate both the content and the context dimensions of our scope.

2.4. Related Surveys and the Position of This Review

Several surveys have mapped parts of this landscape. Zhou and Zafarani [3] provide a foundational account of theories and detection methods; Gong et al. [19] survey graph-based neural approaches; and Comito et al. [20] review deep-learning techniques for multimodal detection. These works are valuable but each concentrate on a single dimension and largely predates the consolidation of LLMs as both detectors and generators. The present review spans content and context within one protocol, records each reported metric from its primary source, and incorporates the 2023-2025 LLM literature together with a Vietnamese low-resource case study.

To make this positioning concrete, Table 1 compares the present review with nine prior surveys across the sub-areas of the field. Earlier surveys are strong on scope: they map content-based methods [3], [4], [21] graph and social context [3], [4], [19], [21] multimodality [20], [22], explainability [23], and, most recently, LLMs and AI-generated misinformation [24], [25]. Among the surveys examined in Table 1, none was found to jointly provide the methodological elements of this review. No prior survey verifies each reported metric against its primary source, confines quantitative comparison to within-benchmark (same-dataset, same-task) settings, or reports a risk-of-bias appraisal of the results it synthesizes; and only one addresses low-resource languages [25]. Those three empty columns in Table 1 are the gap the present review fills. The comparison was coded by the authors from each survey's abstract, scope statements, and section structure: a property is marked as provided only when a survey explicitly addresses it, and ambiguous cases are treated conservatively as not provided. The nine surveys were selected to span the sub-areas of the field (general, multimodal, graph, explainable, LLM, and low-resource), and the coding is our reading of those sources rather than a claim about every survey in existence.

Table 1. Comparison with prior fake-news-detection surveys.

Survey	Period	Coverage	Low-resource	Source-metric verification	Matched-setting comparison	Risk-of-bias
Shu et al. [4]	to 2017	C, G	—	—	—	—
Zhou & Zafarani [3]	to 2019	C, G	—	—	—	—
Guo et al. [21]	to 2020	C, M, G	—	—	—	—
Alam et al. [22]	to 2021	C, M, G	—	—	—	—
Athira et al. [23]	to 2022	C	—	—	—	—
Comito et al. [20]	to 2022	C, M	—	—	—	—
Gong et al. [19]	to 2023	C, G	—	—	—	—
Chen & Shu [24]	to 2024	C, L	—	—	—	—
Wang et al. [25]	to 2024	C, M, L	✓	—	—	✓
This review	2018-2025	C, M, G, L	✓	✓	✓	✓

Coverage lists the areas each survey addresses among content (C), multimodal (M), graph/context (G), and LLM (L). A check mark means the property is provided; a dash means it is not

Source-metric verification: each reported metric checked against its primary source. Within-benchmark comparison: quantitative comparison confined to a common (dataset, task) setting; a shared dataset name and task do not guarantee identical samples, splits, or protocols. Risk-of-bias: an appraisal of the synthesized results. Merely adding the 2023-2025 LLM literature to an existing taxonomy would not change how the evidence is read. The verifiability-focused stance does: once every number is tied to its source and comparison is restricted to within-benchmark settings, several widely repeated figures no longer support a ranking of families (Section 5), and the field's progress is better described as a change in the kind of evidence used than as a race for higher accuracy (Section 4.5). This reinterpretation, not the additional references, is the review's contribution to how the area is understood.

2.5. A Multi-Axis Analytical Framework

The five method families used for tabulation (Section 3.4) are convenient but conceptually uneven: Transformer/PLM and Traditional ML name an architecture,

Multimodal names a data type, Graph names a context representation, and LLM names both an architecture and a mode of use. A single-axis grouping therefore counts the literature well but explains it poorly. To interpret why methods differ and how the field is evolving, we organize the analysis along four complementary, though operationally interacting, dimensions (Figure 1).

- Evidence source: the kind of signal a method relies on, from intrinsic text and style, through cross-modal (text-image) consistency and social-context/propagation signals, to external knowledge or retrieval and, most recently, generated reasoning from LLMs.
- Detection time: when a decision is made, from pre-propagation (content only), through early propagation, to full propagation history.
- Data dependence: how the method obtains supervision, from supervised in-domain training, through transfer learning and few-/zero-shot use, to retrieval- or knowledge-grounded inference.
- Deployment objective: the property being optimized, among accuracy, generalization, explainability, early detection, computational efficiency, and adversarial robustness.

The five families map across these axes rather than onto any one of them. A multimodal detector adds cross-modal evidence but is otherwise a supervised in-domain content model; a graph model adds social-context evidence and shifts detection time later; an LLM method draws on generated-reasoning evidence and low data dependence but variable accuracy. Read this way, the field's trajectory (Section 4.5) is a movement toward richer evidence sources, later and more adaptive reasoning, and lower dependence on in-domain labels, rather than a succession of superior architectures.

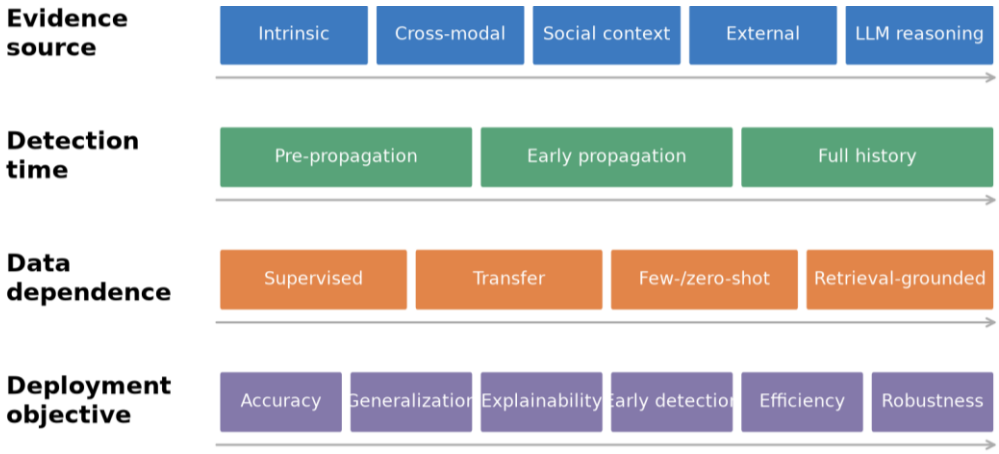

Figure 1. The four analytical dimensions and their ordered categories. The evidence axis is ordered by richness, not trustworthiness; Table 2 positions the five method families across the dimensions

Table 2. Positioning of the method families across the four framework dimensions (a '+' denotes an added evidence source relative to intrinsic content)

Family	Evidence source	Detection time	Data dependence	Deployment focus
Transformer/PLM	Intrinsic text	Pre-propagation	Supervised / transfer	Accuracy
Multimodal	+ Cross-modal	Pre-propagation	Supervised in-domain	Accuracy
Graph/context	+ Social context	Early-to-full propagation	Supervised in-domain	Context, propagation monitoring
LLM-based	+ Generated reasoning	Pre-propagation	Few-/zero-shot, retrieval	Generalization, explanation
Traditional ML	Intrinsic features	Pre-propagation	Supervised in-domain	Efficiency, explainability

Positions are the families' typical operating regions, not exclusive assignments; individual studies may combine dimensions (Section 4.2.6).

The evidence sources differ in reliability, and the axis orders them by richness, not by trustworthiness. Intrinsic, cross-modal, and social-context signals are observed from the item and its diffusion; external evidence is retrieved and may be outdated; and LLM-generated reasoning is model-produced and, unless grounded in verified sources, may be unfaithful or hallucinated. It is therefore useful to distinguish external factual evidence (retrieved documents, knowledge graphs, fact-checks), model-generated reasoning (an intermediate

rationale), and verified evidence-grounded reasoning (rationale tied to checkable sources), and to treat the last as the goal rather than the current norm. Table 2 positions the five families across the four dimensions; the framework structures the per-family analysis (Section 4.2), the trade-off summary (Section 4.2.7), and the research agenda (Section 6.3).

3. Research Methodology

3.1. Search Strategy

Relevant studies were identified in July-August 2026 using a two-pronged strategy. First, the titles, abstracts, and author keywords of records in the major academic databases and indexes for the field-IEEE Xplore, the ACM Digital Library, ScienceDirect, SpringerLink, arXiv, and Google Scholar-were searched with the concept query given below. Second, backward and forward snowballing was performed on the reference lists and citing papers of three recent surveys used as anchors [3], [19], [20], to recover method-family studies that keyword search alone can miss. Two publication windows were applied: a primary window of 2021-2025, which covers the Transformer-, multimodal-, graph-, and LLM-based methods that are the focus of this review, and a small set of seminal 2018-2020 works retained to establish architectural lineage. The concept query combined a topic block, a task block, and a method block:

("fake news" OR "misinformation" OR "disinformation" OR "rumor detection" OR "rumour detection" OR "fact verification") AND ("detection" OR "classification" OR "identification" OR "verification") AND ("deep learning" OR "Transformer" OR "BERT" OR "pre-trained language model" OR "graph neural network" OR "propagation" OR "multimodal" OR "large language model" OR "LLM").

Rumour- and rumor-detection terms are included deliberately, because several graph- and propagation-based studies in the corpus are framed as rumour detection rather than as fake-news detection, and would otherwise be missed.

Because identification deliberately combined database keyword searching with citation snowballing, a single per-database hit count would misrepresent how studies were actually located; we therefore report a transparent selection flow over the pooled candidate set (Figure 2) instead of a per-database PRISMA funnel. To make the process fully inspectable and reproducible, the complete enumerated set of screened candidates and included studies is released as supplementary material (corpus.csv). This is a structured, verifiability-focused review rather than a database-exhaustive systematic review; the resulting scoping limitations are stated in Section 6.4.

The 51 candidates therefore constitute a purposive sample rather than the output of an exhaustive harvest: they were assembled from the reference lists of the three anchor surveys, direct browsing of the leading venues within the review's scope, and backward and forward snowballing from studies already included. Accordingly, the family, country, and annual distributions reported in Section 4 characterize this corpus and should not be read as estimates of the wider field.

3.2. Inclusion and Exclusion Criteria

Screening was conducted in two stages. In the first (title-abstract) stage, records were screened against the topic, method, and evaluation criteria below. In the second (full-text) stage, the surviving records were read in full to confirm that the reported headline metric could be traced to an accessible primary source; records failing this verifiability check were excluded at this stage. A study was included only if it satisfied all four inclusion criteria:

- I1 - Topic: it addresses content- and/or context-based detection of fake news, misinformation, disinformation, or rumours.
- I2 - Method: it proposes or evaluates a machine-learning or deep-learning detection method.
- I3 - Evaluation: it reports an empirical evaluation on real-world data, with at least one quantitative headline metric (Accuracy, F1, macro-F1, or AUC) on a named dataset.
- I4 - Verifiability: the reported headline metric is confirmable against an accessible primary source-a peer-reviewed article, or, for well-cited preprints with complete evaluations, the arXiv record (flagged as a preprint).

- A study was excluded if any of the following applied:
- E1 - it concerns video or audio signal processing only, or image forensics without a misinformation-detection task.
- E2 - it relies on purely heuristic account- or metadata-based filters with no learned model.
- E3 - it reports no empirical evaluation or no quantitative detection metric (for example, a qualitative analysis only).
- E4 - its headline metric or citation could not be verified against an accessible source.

Applying these criteria, and after re-examining the initially unverifiable candidates (below), 51 screened candidates yield 45 included studies; the 6 exclusions are itemized below for transparency. Of the 45 included studies, 42 are classification studies that form the core quantitative set, and 3 are dataset or robustness papers retained as contextual references. Figure 2 shows the selection flow.

Records were deduplicated by DOI, arXiv identifier, and normalized title; where a preprint and a peer-reviewed version of the same work coexisted, the peer-reviewed version was retained. Screening and full-text verification were carried out by the authors, and borderline inclusion or verifiability decisions were resolved by discussion. The seven candidates subject to full-text verifiability review, with the reason each was retained or excluded, are listed below and are also released as excluded.csv in the supplementary material.

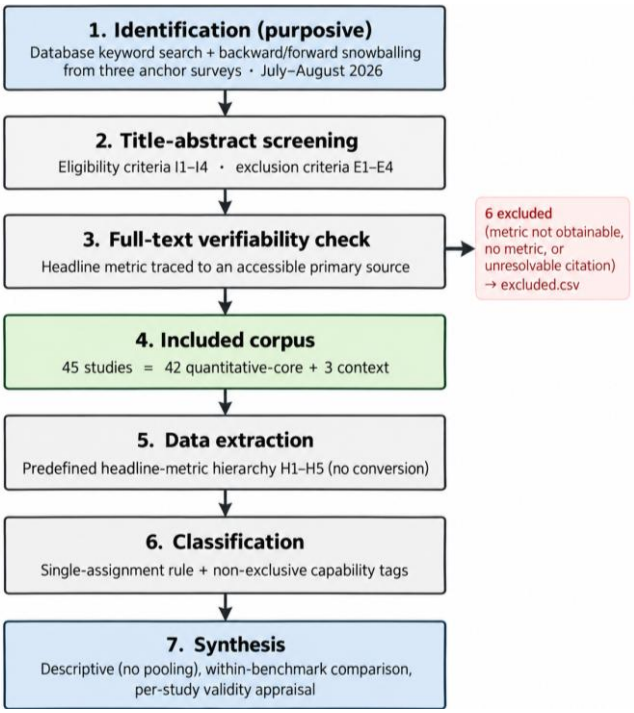

Figure 2. Review protocol: identification, title-abstract screening, full-text verifiability check, extraction, classification, and synthesis. Counts characterize this corpus; the full study list and the excluded studies (excluded.csv) are provided as supplementary material.

The corpus is enumerated rather than drawn from an exhaustive database funnel, and every included study is listed with its reported metric. The full list, with authors, year, venue, method family, modality, dataset, task, reported metric, and source, is provided as supplementary material (corpus.csv). The consequences of this scoping choice are discussed in Section 6.4.

At initial screening, seven candidates could not be verified and were set aside. On re-examination for this revision, one of them, Yang et al. [26] (SureFact, KDD 2022), was obtained in full from the camera-ready version and its headline confirmed (PolitiFact accuracy 94.36%); it is now included as a core study. The remaining six are excluded on the following grounds. For three genuine, in-scope papers the authors' headline metric could not be obtained from any accessible source, because the abstract reports no figure, the full text is

gated, and third-party tabulations are re-implementations that disagree (for example, one re-implementation lists AMFB at 0.75 accuracy, which is unsafe to record as the authors' figure): Kumari and Ekbal [27]; Xie et al. [28]; and Yuan et al. [29]. One purported graph-and-Transformer paper could not be identified because its citation locator resolves to a different article; one (Kula et al.) could not be uniquely identified among several same-author candidates; and one (Pavlyshenko) reports only qualitative analysis with no detection metric. The three genuine but currently unverifiable papers ([27]–[29]) are retained as cited related work but not used as quantitative data points. The full excluded list with reasons is released as excluded.csv.

3.3. Data Extraction

From each study we extracted: (1) bibliographic information (authors, year, venue, venue type, peer-reviewed vs preprint, first-author country); (2) method type, model architecture, and modality; (3) primary dataset, task (binary vs multi-class), and whether social context/propagation is used; (4) the reported headline metric with its exact name (Accuracy, F1, macro-F1, or AUC) and value, recorded as reported with no conversion; and (5) stated challenges and future directions. Each metric was recorded from its primary source.

When a study reported results for several datasets, metrics, model variants, or evaluation settings, a single headline figure was extracted using a predefined hierarchy, applied in order, to avoid best-of-the-paper selection:

- H1 - Model: take the figure for the authors' primary proposed model as identified in the abstract or conclusion, not an ablation, a baseline, or an oracle/upper-bound variant.
- H2 - Dataset: if several datasets are reported, take the one the authors feature in the abstract as their headline result; if none is featured, take the most widely used public benchmark among them (for example Weibo, LIAR, PolitiFact, ISOT, or Fakeddit) to maximize comparability.
- H3 - Metric: record the metric the authors report as their main result on that dataset, exactly as stated and with no conversion; for class-imbalanced multi-class settings, macro-F1 is preferred over accuracy when both are reported.
- H4 - Setting: prefer the standard supervised in-distribution test result over few-shot, zero-shot, or cross-domain variants, unless the study is specifically a zero-shot or cross-domain study, in which case its primary setting is used and flagged as such.
- H5 - Provenance: record where in the source the figure appears (abstract, results table, or leaderboard) and whether it is the best among several reported configurations; this flag feeds the result-selection item of the validity appraisal (Section 3.6).

Each study's extracted figure, the dataset and metric it refers to, its selection under this hierarchy, and the location of the supporting evidence are recorded per study in the supplementary corpus, so that the extraction can be independently audited.

3.4. Classification Rule

To produce a single-assignment tabulation despite the overlapping dimensions of Section 2.2, each study was classified by its primary methodological contribution, the component the authors positioned as their central novelty and around which the evaluation was organized, using the following priority order when a study spanned categories: (1) joint text-image fusion -> Multimodal; (2) else propagation/social-graph modeling -> Graph/propagation-based; (3) else prompting/zero-shot/instruction-tuned use of a large generative model -> LLM-based; (4) else pre-trained Transformer/PLM classification -> Transformer/PLM; (5) otherwise, classical feature-based models -> Traditional ML. This is a reporting convention, not a claim of architectural disjointness; Section 4.2.6 additionally reports non-exclusive capability tags, and the full per-study classification is in the supplementary corpus.

3.5. Synthesis Approach

Data were synthesized qualitatively (taxonomy, trends, challenge-solution mapping) and through a descriptive quantitative synthesis (counts, shares, per-year and per-region distributions, dataset-usage frequencies, and per-family reported-accuracy ranges). We do

NOT perform a statistical meta-analysis and do NOT convert between metrics. A meta-analysis requires comparable effect sizes with standard errors, a defined pooling model, and heterogeneity diagnostics; the primary studies report different metrics on different datasets, tasks, and class distributions, and rarely provide the confusion matrices or variance estimates such pooling would need. Likewise, Accuracy, F1, and AUC are not interchangeable and cannot be derived from one another without the underlying class distribution and prediction scores. Family-level performance is therefore reported as descriptive ranges with explicit caveats (Section 5), and any direct numerical comparison is confined to a common (dataset, task) setting. One study reporting 100% accuracy on a 30-article sample was retained in the corpus but excluded from the reported ranges as methodologically unreliable.

3.6. Quality and Risk-of-Bias Considerations

Confirming that a reported metric matches its source establishes provenance, not internal validity. We therefore appraised the reviewed studies against a qualitative risk-of-bias rubric covering the factors that most affect the credibility of fake-news-detection results. We report the rubric and a synthesized discussion rather than a per-study score, because many studies do not disclose enough detail (splits, seeds, code) to grade every item reliably. The rubric covers:

- Data leakage: whether train and test are properly separated, including near-duplicate items or shared events across splits.
- Split protocol: whether the train/validation/test split and ratio are reported and standard.
- Class balance: whether class imbalance is reported and handled.
- Sample size: whether the evaluation set is large enough to support the claim.
- External validation: whether the model is tested on a dataset different from the one it was trained on.
- Baseline fairness: whether baselines are tuned comparably to the proposed model.
- Statistical treatment: whether variance across runs or any significance test is reported.
- Result selection: whether the headline number is the best of many configurations.
- Reproducibility: whether code, weights, and data are publicly available.

Source verification also has limits worth stating precisely. Of the 45 studies, the headline metric of 36 was confirmed directly against an accessible primary source; for the remaining nine, whose full text sits behind paywalls, the metric is recorded as reported but not independently confirmed, and one of these (flagged in the corpus) shows a discrepancy between the corpus entry and the source. These nine are retained for coverage but marked in corpus.csv, and none is used as the sole basis for a within-benchmark claim.

Read across the corpus, the appraisal surfaces several concerns. First, the near-perfect accuracies (for example 99.97% on TruthSeeker and 99.36% on a single ~20k-article balanced sample (the best of several configurations)) are consistent with, but not proof of, explanations such as leakage, duplicate contamination, or dataset-specific separability; because we did not re-run the original train/test splits, we treat these as possible rather than established explanations and read the figures with caution rather than as evidence of superior transferable capability. Second, external validation is rare: most studies report in-distribution test accuracy on a single dataset, and the few cross-dataset or cross-task results (Section 5) show large drops. Third, most studies report a single run without variance or significance testing, so small differences between models are not reliably distinguishable. Fourth, headline numbers are often the best configuration among several, which inflates apparent performance. Fifth, code and data availability are inconsistent, limiting independent reproduction, and the Vietnamese studies additionally rely on small, self-collected datasets. These considerations reinforce the review's central stance: reported numbers are treated as provenance-checked but not risk-free, comparison is confined to within-benchmark settings, and no family is ranked by a single figure.

Table 3 summarizes this appraisal: each of the 42 core studies was coded on the nine criteria above, and the full per-study codes with their supporting evidence are released as risk_of_bias.csv so that the appraisal can be independently checked. The distribution shows that missing external validation (high concern in 24 of 42 studies), single-run reporting without variance or significance testing (14 of 42), and best-of-many result selection (15 of

42) are the most common threats to validity, while reproducibility is comparatively good (21 of 42 at low concern, reflecting public code or data).

Table 3. Per-study risk-of-bias appraisal of the 42 core studies: distribution of concern levels across nine criteria. Per-study codes with supporting evidence are released as risk_of_bias.csv.

Criterion	Concern level - Low / Some / High / Unclear (n = 42 core)	Predominant issue in this corpus
Data leakage / near-duplicates	9 / 16 / 1 / 16	train/test split usually given; deduplication rarely reported
Split protocol reported	22 / 4 / 1 / 15	split and ratio usually reported; random splits common
Class balance handled	15 / 6 / 3 / 18	often reported; sometimes imbalanced or unstated
Adequate evaluation sample	13 / 11 / 4 / 14	several test sets under 10k; Vietnamese sets small
External (cross-dataset) validation	2 / 3 / 24 / 13	mostly single-dataset in-distribution testing
Baseline fairness	11 / 18 / 1 / 12	baselines not always tuned comparably
Variance / significance testing	4 / 11 / 14 / 13	mostly single-run; no variance or significance
Result selection (headline = best of many)	11 / 6 / 15 / 10	headline often the best of several configurations
Reproducibility (code / data)	22 / 7 / 5 / 8	public for many; inconsistent overall

Each of the 42 core studies was coded per criterion as low, some, or high concern, or unclear where reporting was insufficient to judge; unclear is not treated as evidence of poor conduct. Counts are studies out of 42. The highest-concern items (missing external validation, absent variance/significance testing, best-of-many result selection) are those most able to inflate the reported numbers in Sections 4 and 5.

4. Results and Discussion

4.1. Characteristics of the Corpus

The 45 verified studies span 2018-2025, with publications rising through the period and peaking in 2024 (Figure 3); the emphasis of the corpus is on 2021-2025, while a small number of 2018-2020 works are retained for architectural lineage. Figure 4 shows the first-author geographic distribution, which is led by the United States and China (nine studies each) followed by Vietnam (eight, reflecting the review's low-resource focus) and India (four). Figure 5 shows the distribution of the 42 core classification studies by primary method family: Transformer/PLM 13 (31.0%), multimodal 11 (26.2%), LLM-based 9 (21.4%), graph/propagation 7 (16.7%), and traditional machine learning 2 (4.8%). LLM-based methods thus already constitute nearly a quarter of the recent core corpus, ranking third after Transformer/PLM and multimodal methods. As Section 4.2.6 makes explicit, these shares reflect primary contribution, not architectural exclusivity, most multimodal and all LLM studies also use Transformer components.Appendix A lists all 45 analyzed studies with their reported metrics and sources.

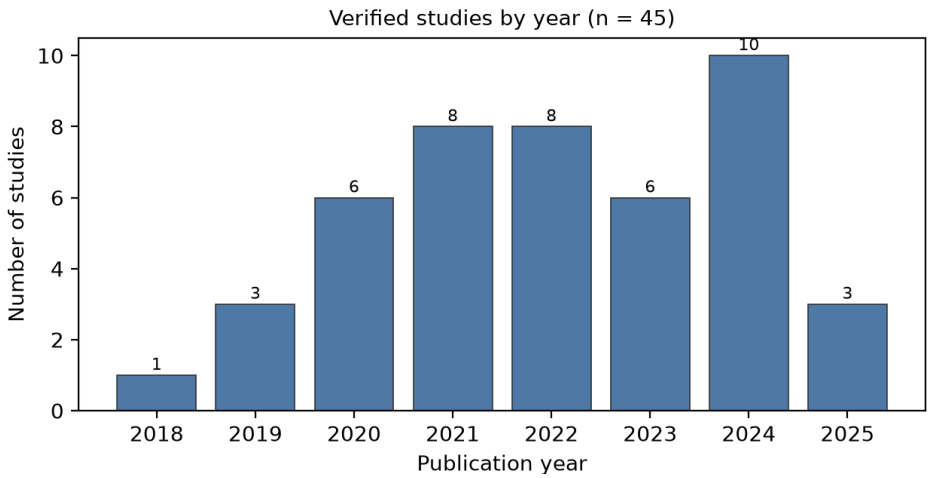

Figure 3. Verified studies by publication year (n = 45)

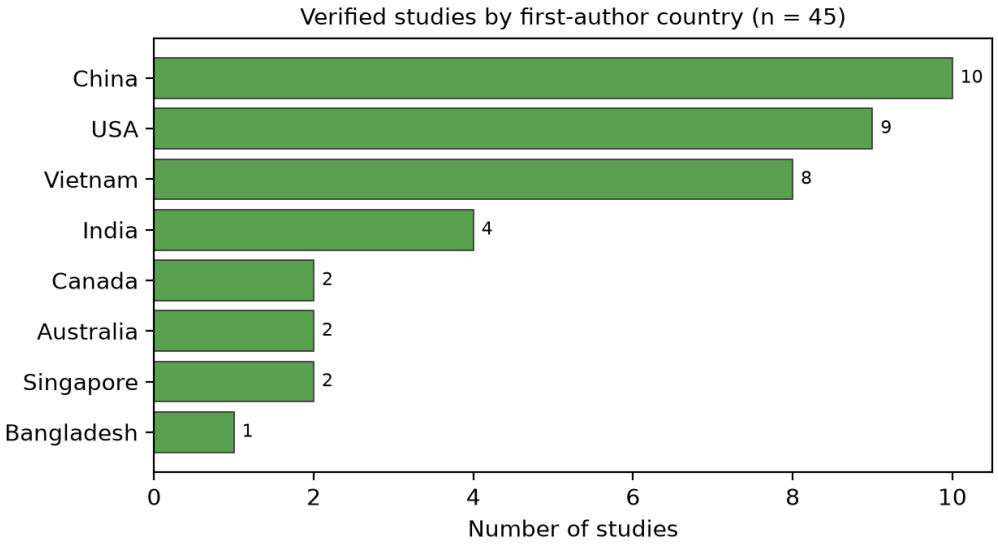

Figure 4. Verified studies by first-author country (top countries)

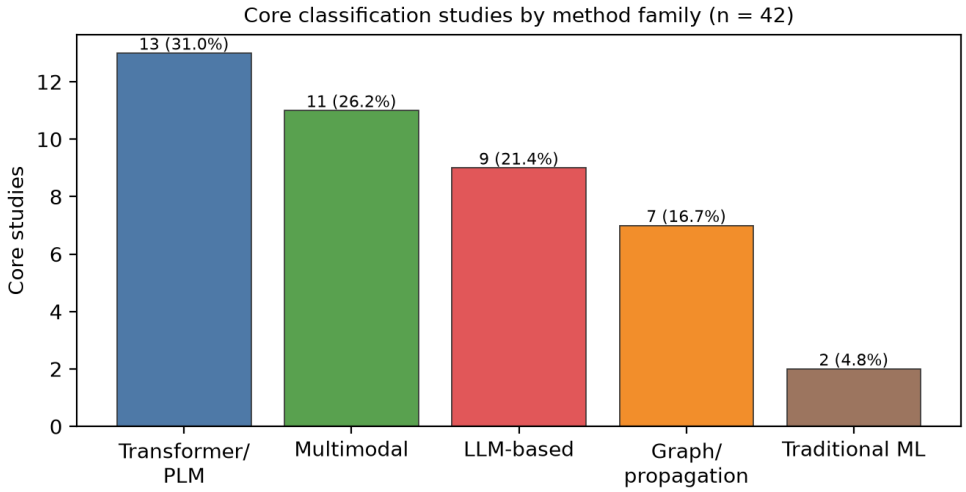

Figure 5. Core classification studies by primary method family (n = 42)

4.2. Classification and Analysis of Methods

The following per-family analyses combine the reported results with our interpretation of the reviewed evidence. Interpretive claims about why performance differs are flagged as such and cited to a reviewed study where one directly supports them; broader mechanisms (shortcut learning, missing-modality effects, explanation faithfulness, transfer bias) are stated as syntheses of the reviewed evidence to be tested, not as established facts.

4.2.1. Transformer- and PLM-Based Methods

Transformer-based models, particularly BERT [6] and its task-specific variants, are the largest family (31.0% of core studies). The canonical pipeline is: input text -> tokenization -> BERT encoder -> [CLS] representation -> dense layers -> softmax -> classification. Table 4 lists verified representative results. The family draws strength from transfer learning, bidirectional contextual understanding, and multilingual pre-trained models. Its reported accuracies, however, are dataset-dependent, spanning 74.8-99.36% across the corpus. The highest figure occurs on a merged five-dataset corpus evaluated in-distribution (99.36% [30]), while a transformer classifier on a more heterogeneous content-plus-context setting reaches 74.8% [31]. The reported accuracy of this family thus depends heavily on the evaluation dataset. Metrics are recorded exactly as reported by each primary source (no conversion). The near-perfect value (99.36%) versus 74.8% on a content-plus-context setting illustrates that a single accuracy figure is not a measure of transferable capability.

Table 4. Verified representative results for Transformer/PLM-based methods

Model	Dataset	Task	Reported metric	Ref.
FakeBERT (BERT+CNN)	Kaggle real/fake	binary	98.90% Acc	[32]
CT-BERT + BiGRU	Constraint COVID-19	binary	98.54% F1	[33]
ALBERT (best of 3 PLMs)	Indonesian corpus	binary	87.6% Acc	[34]
Doc-embedding + BiGRU	Merged 5-dataset corpus	binary	99.36% Acc	[30]
Content+context Transformer	NELA-GT-19 + Fakeddit	binary	74.8% Acc	[31]

In framework terms (Section 2.5), this family uses intrinsic textual evidence, supervised in-domain data, and pre-propagation timing, so its reported performance appears strongly influenced by how separable the training signal is, in addition to architecture. High reported accuracy typically arises when a dataset carries strong lexical, source, or editorial-style cues that correlate with the label, letting the model learn dataset shortcuts (topic, outlet name, writing style) rather than veracity; that such style cues can dominate is illustrated by style-transfer attacks that flip predictions without changing meaning [35]. The same fine-tuned model then loses much of that accuracy when the topic, time period, or platform changes. Transformer/PLM methods fit settings with ample labeled text, low-latency requirements, and no propagation data yet, and they serve as strong content-only baselines. They are weakest on grammatically fluent, neutrally worded, or LLM-generated fake news, where surface cues are absent.

4.2.2. Multimodal Methods

Multimodal methods (26.2% of core studies) combine text and image analysis, motivated by the observation that fake news often pairs text with mismatched, out-of-context, or manipulated images. The direction matured from event-adversarial and variational designs (EANN [8], MVAE [9]) through pre-trained dual-encoder frameworks (SpotFake [10]) to similarity- and ambiguity-aware fusion (SAFE [36], CAFE [37]), hierarchical contextual attention (HMCAN [38]), and bootstrapped multi-view representations (BMR [39]); related designs explore dual-stream fusion [40] and attention-based factorized bilinear pooling [27]. A key advantage of this family for synthesis is that many studies evaluate on the same Weibo benchmark, enabling the more comparable within-dataset summary used in Section 5. Table 5 lists verified results; on the shared Weibo benchmark the reported accuracies cluster in a comparatively narrow 82.4-91.8% band, while the highest value in Table 5 (99.97% on TruthSeeker [41]) reflects a dataset with strong dataset-specific separability.

Table 5. Verified representative results for multimodal methods (Weibo unless noted)

Method	Modality	Dataset	Task	Reported Acc.
EANN	text+image	Weibo	binary	82.7%
MVAE	text+image	Weibo	binary	82.4%
CAFE	text+image	Weibo	binary	84.0%
HMCAN	text+image	Weibo	binary	88.5%
SpotFake	text+image	Weibo	binary	89.2%
LLM-enhanced fusion	text+image	Weibo	binary	91.2%
BMR	text+image	Weibo	binary	91.8%
SAFE	text+image	PolitiFact	binary	87.4%
BERT+OCR fusion	text+image	TruthSeeker	binary	99.97%

Sources: EANN [8]; MVAE [9]; CAFE [37]; HMCAN [38]; SpotFake [10]; LLM-enhanced fusion [42]; BMR [39]; SAFE [36]; BERT+OCR fusion [41].

The seven Weibo entries form the within-benchmark comparison in Table 14; the TruthSeeker value reflects a dataset with strong dataset-specific separability. Multimodal methods add cross-modal evidence but otherwise share the supervised, in-domain, pre-propagation profile of content models. Fusion helps specifically when deception is signaled by a text-image inconsistency (out-of-context or manipulated images), which is the basis of similarity- and ambiguity-aware fusion [36], [37]; it adds little, and can hurt, when the image is merely decorative or missing, and it is sensitive to the quality of text-image alignment in the training pairs. The cost is higher: two encoders, paired multimodal data that is hard to collect

and label, and degraded behavior when a modality is absent. These methods therefore suit image-rich social platforms and are poorly suited when one modality is frequently missing.

4.2.3. Graph- and Propagation-Based (Context) Methods

Graph-based methods (16.7% of core studies) model diffusion as a graph in which nodes are users or posts and edges are interactions (retweet, comment, share) [19]. Because node features usually encode textual content, these are content-and-context hybrids rather than pure metadata methods, which places them within scope. Representative systems include bi-directional propagation (BiGCN [11]), geometric deep learning [12], user-preference-aware detection (UPFD [43]), continually updated social-context representations [44], domain-adversarial graph-attention networks [29], and the FANG social-graph model [45]. Table 6 shows verified results and highlights a methodological point: this family reports a mixture of accuracy and AUC, so its numbers cannot be placed on a single scale.

Graph/propagation methods move along the framework's evidence axis, adding social-context signals, and along its time axis, deciding later. They improve detection by learning from diffusion structure, community patterns, and source credibility that content-only models cannot see. That same dependence makes them unsuitable for early detection, since they need accumulated interactions; it also risks learning unstable platform- or community-specific structure and exposes them to API changes, privacy policies, and coordinated inauthentic behavior. They fit post-hoc monitoring once a story has begun to spread, rather than first-contact screening.

Table 6. Verified representative results for graph/propagation-based methods

Model	Dataset	Task	Metric	Value	Ref.
BiGCN	Twitter15	multiclass	Accuracy	88.6%	[11]
UPFD (GraphSAGE+BERT)	PolitiFact	binary	Accuracy	84.6%	[43]
GNN-CL (continual)	GossipCop	binary	Accuracy	84.4%	[46]
GCNFN (geometric DL)	Twitter	binary	AUC	0.927	[12]
Continual social-context GNN	FANG data	binary	AUC	0.835	[44]
SureFact	PolitiFact	binary	Accuracy	94.36%	[26]
FANG	FANG data	binary	AUC	0.752	[45]

Sources as cited. The coexistence of accuracy and AUC within one family illustrates the metric heterogeneity that precludes cross-study pooling (Section 3.5).

4.2.4. Large Language Model (LLM)-Based Methods

LLM-based methods (21.4% of core studies) apply models such as GPT-3.5/4, Llama, and Mistral to detection through: (1) zero-/few-shot prompting [13]; (2) fine-tuning of smaller models [47]; and (3) LLM-enhanced detection, where LLMs supply rationales or comments to smaller models [48], [49]. Table 7 lists verified results. LLM detection accuracy is the lowest and most variable of all families (48.6-89.0%)[15]. GPT-4 attains 68.2% on LIAR-binary in a zero-shot setting [13]. LLMs do offer natural-language explanations and few-shot flexibility, but at high inference cost and with strong sensitivity to prompt and setting. They also create new challenges: humans correctly identify only about 54.8% of ChatGPT-generated fake news [14], and whether LLM-generated misinformation is reliably machine-detectable remains open [15].

Table 7. Verified representative results for LLM-based methods

Model / method	Setting	Dataset	Metric	Value	Ref.
GPT-4	zero-shot	LIAR-binary	Accuracy	68.2%	[13]
CAPE-FND (GPT-3.5)	zero-shot	PolitiFact	Accuracy	78.1%	[50]
ARG (LLM-as-advisor)	trained	GossipCop	macro-F1	0.790	[48]
SheepDog	trained	PolitiFact	macro-F1	88.4%	[35]
DKFND	few-shot	PolitiFact	Accuracy	89.0%	[51]

Values span zero-shot to trained settings and are not comparable across rows. GPT-4's 68.2% on LIAR-binary corrects the frequently cited ~95%. One comparative study reporting

100% accuracy on a 30-article sample [47] is excluded from all ranges as methodologically unreliable.

LLM methods occupy the framework's generated-reasoning evidence and low-data-dependence region, but with unstable objectives. Zero-shot use is flexible yet poorly calibrated and highly prompt-sensitive, which is why the family's reported accuracy is both the lowest and the most dispersed. Their natural-language explanations are a genuine asset but are not guaranteed to be faithful to the actual decision process. Fine-tuning and retrieval grounding raise accuracy at higher cost. LLMs are most valuable where labeled data are scarce, coverage must be multilingual, or human analysts need assistance, and are not recommended as the sole basis for high-stakes decisions, given hallucination and output inconsistency.

4.2.5. Traditional Machine Learning

A small residual of studies (4.8%) rely on classical feature-based models (e.g., LightGBM, SVM) rather than deep architectures, typically in low-resource settings. In this corpus they appear mainly among Vietnamese studies (Section 4.4), reaching 85.0-88.2% accuracy on self-collected data. They remain relevant where labeled data and compute are limited and where interpretability is valued, but they are being displaced by pre-trained models as multilingual weights become available, as benchmark comparisons of classical and neural detectors confirm [52].

Traditional models should not be read merely as a family being displaced. In low-data, low-compute, or low-latency settings they remain competitive, provide transparent feature-level explanations, and make strong baselines; their generalization, however, is bounded by the quality of hand-crafted features. A practical role is as the first, inexpensive tier of a multi-stage pipeline that escalates only uncertain cases to heavier models.

4.2.6. Overlap Between Families and Non-Exclusive Capability Tags

Because families are defined by primary contribution rather than by disjoint architectures, Table 8 makes their overlap explicit by cross-tabulating each family against four capabilities. Transformer components are core to both the Transformer/PLM and LLM families and common in multimodal systems; the image modality is core only to multimodal methods; and propagation/graph structure is core only to graph-based methods. Thus 'Transformer/PLM (31.0%)' counts studies whose primary novelty is a text-only pre-trained classifier, whereas most multimodal and all LLM studies also use Transformers internally. Reporting these tags prevents the common misreading that the shares partition the field along a single axis.

Table 8. Non-exclusive capability tags across the four deep-learning families

Family (share)	Primary-contribution criterion	Transformer	Image	Graph	LLM
Transformer/PLM (31.0%)	Text-only PLM fine-tuning is the central novelty	Core	Rare	Rare	No
Multimodal (26.2%)	Joint text-image fusion	Common	Core	Occasional	Occasional
LLM-based (21.4%)	Prompted/instruction-tuned generative LLM	Core	Occasional	Occasional	Core
Graph/propagation (16.7%)	Propagation/social-graph modeling	Occasional	Rare	Core	Rare

Core = defining component; Common / Occasional / Rare = observed frequency among reviewed studies. Non-empty off-diagonal cells show the architectural and modality axes are not disjoint; the single-family assignment reflects each study's primary contribution (Section 3.4). Traditional ML (4.8%) uses none of these deep components by definition and is omitted.

4.2.7. Method-Family Trade-Offs and Application Scenarios

The framework and the per-family analyses can be summarized as a qualitative trade-off map (Table 9). The ratings are the authors' qualitative synthesis over the corpus, not measured scores; they use Low/Moderate/High/Conditional and are meant to guide method selection, not to rank families. No row dominates: each family trades strengths against weaknesses, so the right choice depends on the target task, the evidence available, the latency budget, and the deployment constraints. As a rough guide to best fit, traditional ML suits low-resource, low-latency settings; Transformer/PLM suits text-only, fast-response settings; multimodal suits image-rich platforms; graph/context suits propagation monitoring; and LLMs suit few-shot, human-in-the-loop use. Ratings were assigned from the dominant operating characteristics

observed across each family's studies. Because primary studies rarely report standardized cost or cross-domain measures, the ratings are structured qualitative judgments rather than empirical aggregate scores. This un-derpins the recommendations of Section 6.3. A rating is assigned as High when the property is consistently present across the representative studies of that family in the corpus, Moderate when it is mixed or setting-dependent, Low when it is typically absent, and Conditional when it depends on the availability of a second modality, propagation history, or external resources. Each rating is grounded in the per-family analyses of Sections 4.2.1-4.2.5 and in the per-study evidence released in the supplementary corpus and risk_of_bias.csv.

Table 9. Method-family trade-offs (qualitative synthesis over the corpus; Low/Moderate/High/Conditional, not measured scores)

Family	In-domain	Generali-zation	Explaina-bility	Labeled data	Compute	Early detection	Evidence avail
Traditional ML	Moderate	Low-Mod.	High	Low	Low	High	High
Transformer/PLM	High	Low-Mod.	Low	Moderate	Moderate	High	High
Multimodal	Conditional	Low-Mod.	Low	High	High	High	Conditional
Graph/context	Conditional	Low	Moderate	High	Mod.-High	Low	Low (early)
LLM-based	Conditional	Moderate	High*	Low	Very high	High	Conditional

Ratings are the authors' qualitative synthesis over the corpus, not measured values. Conditional = depends on the availability of a second modality, propagation history, or external resources. 'Labeled data' and 'Compute' are separated because LLMs need few labels but heavy compute. 'High*' for LLM explainability denotes fluent surface rationales not guaranteed faithful to the decision. 'Evidence avail.' is whether the evidence a family needs is typically available at decision time. No family dominates across columns, which is the basis for combining families in a pipeline (Section 6.3).

4.3. Benchmark Datasets

Figure 6 shows the datasets most frequently used by the 42 core studies. Weibo is the single most common benchmark (seven studies, all multimodal), followed by the PolitiFact/FakeNewsNet family [53] (six) and self-collected Vietnamese datasets (four); GossipCop appears in three studies and the Vietnamese ReINTEL shared task in four. The concentration on Weibo among multimodal methods is what enables the within-benchmark comparison in Section 5. Beyond the corpus, widely used resources include LIAR [54] (fine-grained six-way labels), Fakeddit [55] (over one million multimodal samples), ISOT, and health-focused or multilingual sets such as CoAID [56], WELFake [57], and the multilingual multimodal MuMiN [58]. Existing datasets share four limitations: (1) scale and balance, many contain fewer than 10,000 samples with class imbalance; (2) label quality, fact-checker or crowd labels carry bias; (3) temporal drift, narrow time windows reduce generalization; and (4) language coverage, most data are in English, and low-resource languages such as Vietnamese remain underrepresented. These limitations directly constrain how the performance numbers in Section 5 should be interpreted.

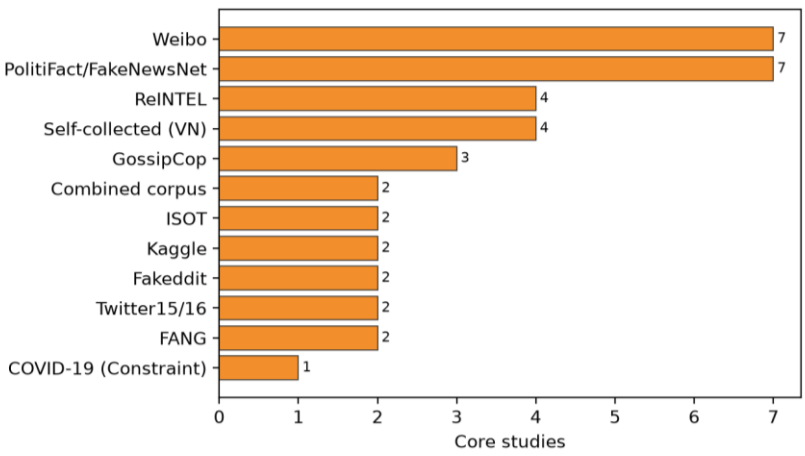

Figure 6. Most frequently used datasets among the 42 core studies

4.4. Vietnamese Fake News Detection: A Low-Resource Case Study

Vietnam contributes eight verified studies (RQ6), summarized in Table 10. A nuanced picture emerges that a single accuracy figure would obscure. On the ReINTEL 2020 shared task, Vietnamese systems built on PhoBERT [59] and vELECTRA achieve ROC-AUC around 0.95 (0.9462-0.9575); because this involves a different dataset and task, it is not directly comparable to English-benchmark results. On noisier, less curated self-collected tasks, however, the same ecosystem reaches only 85-88% accuracy, reflecting small datasets (roughly 4,000-10,000 items), noisy user-generated text, and limited pre-trained resources. Reporting only the lower accuracy figure, as is common, understates the shared-task result, while reporting only the AUC would overstate real-world robustness; both are given here. Metrics recorded as reported. The ReINTEL entries report ROC-AUC on a binary shared task; the self-collected entries report accuracy on noisier, less curated data. 'n/r' = exact values not openly reported.

Vietnamese detection faces four specific challenges. (1) Language: Vietnamese is isolating with no inflection, making word segmentation difficult; homophony and context-dependence add complexity. (2) Resources: high-quality pre-trained models are limited (mainly PhoBERT [41] and vELECTRA), and large labeled datasets and fact-checking infrastructure are scarce. (3) Social-media style: teencode, slang, code-switching, and emoticons complicate analysis. (4) Cultural context: fake news is tied to local topics, requiring deep contextual understanding. Emerging directions include fine-tuning multilingual models (mBERT, XLM-RoBERTa [60]); developing specialized pre-trained models; data augmentation via back-translation; combining PhoBERT features with metadata, including hybrid transfer-learning with TF-IDF [40]; multimodal models such as v3MFND [61]; and real-time detection [62].

Table 10. Verified Vietnamese fake news detection studies (2020-2025)

Study	Year	Method	Dataset	Reported metric
Hieu/Pham et al. [63]	2020	PhoBERT + ensemble	ReINTEL (10K)	AUC 0.9521 (1st place)
Trinh et al. [64]	2021	Hybrid deep NN	ReINTEL (10K)	AUC 0.9462
Pham et al. [65]	2021	Transfer learning + TF-IDF	ReINTEL (10K)	AUC 0.9538
Pham Dang et al. [66]	2023	vELECTRA + handcrafted	ReINTEL (10K)	AUC 0.9575
Vo et al. [67]	2022	CNN + RNN	Self-collected (~4K)	~85% Acc
Nguyen Thi et al. [61]	2022	v3MFND (multimodal)	Self-collected (10 domains)	per-domain AUC (+~6%)
Le et al. [68]	2024	LightGBM + NLP	Self-collected FB (~5K)	88.21% Acc
Huynh & Tran [62]	2025	PhoBERT + Word2Vec	Self-collected	best Acc/F1 (values n/r)

4.4.1. Transferable Lessons for Low-Resource Languages

The Vietnamese case suggests design principles that may transfer to other low-resource languages. Because they are drawn from a single language, we state them as transferable hypotheses to be tested, not as established generalizations (RQ6).

- Shared tasks accelerate progress, but a high score on one benchmark does not imply robustness in the wild; low-resource work should pair shared-task results with out-of-distribution evaluation.
- Language-specific pre-trained models (here PhoBERT and vELECTRA) help substantially, yet still require multi-domain and multi-time benchmarks to demonstrate durability.
- Code-switching, slang, and spelling variation are not Vietnamese-specific; they recur across low-resource social-media languages and should be treated as first-class modeling problems.
- Self-collected datasets are typically small and noisy, so event-based and temporal splits are essential to avoid over-optimistic estimates.
- Multilingual transfer (mBERT, XLM-R) is useful but can import biases from high-resource languages, so its outputs need language- and culture-specific validation.
- Small models with hand-crafted features retain value under compute constraints and as first-tier filters, echoing Section 4.2.7.

- The scarcity of local fact-checking infrastructure and knowledge resources is a bottleneck that larger models alone cannot remove; it is a data- and systems-design problem as much as a modeling one.

Two clarifications bound these principles. First, 'low-resource' here refers specifically to labeled fake-news datasets, domain-specific fact-checking resources, and benchmark diversity, not to Vietnamese NLP resources in general, which are relatively mature (for example PhoBERT [59]). More broadly, the resource-constrained setting also includes limited annotation budgets, restricted compute, limited fact-checking infrastructure, and weak access to platform-level propagation data. Second, the principles are supported by, but not established beyond, the Vietnamese case: comparable data-scarcity and language-variation issues appear in other settings, such as Indonesian fake-news detection [34], but confirming the transfer to Arabic, Hindi, or code-switched corpora requires evidence outside this review's scope.

4.5. Trends and Open Challenges

4.5.1. The Evolution of the Field

Read through the framework of Section 2.5, the corpus shows the field evolving through five overlapping stages rather than a sequence of superior architectures. The stages are analytical and overlapping rather than a strict chronological periodization: graph and multimodal work predates mature pre-trained language models, and robustness research predates AI-generated content. The 2024 publication peak (Figure 3) coincides with increased attention to LLM- and reasoning-oriented methods.

1. Stage 1, Feature-centric detection: hand-crafted linguistic and stylistic features with classical machine learning [4], [5]; now a small residual (4.8% of core studies) but still useful in low-resource settings (Section 4.2.5).
2. Stage 2, Representation-centric detection: deep representation learning and pre-trained language models [6], [7], which dominate the corpus (Transformer/PLM, 31.0%) and shift the emphasis from feature engineering to transfer learning.
3. Stage 3, Evidence-integration detection: joint use of text, images, users, and propagation (multimodal 26.2% and graph/context 16.7%), moving along the evidence and time axes of the framework and building on a lineage from EANN [8] to ambiguity-aware fusion [37] and social-graph hybrids [26], [28], [69].
4. Stage 4, Knowledge- and reasoning-centric detection: LLMs, retrieval, fact verification, and generated rationales (LLM-based, 21.4%, since 2023), adding external-knowledge and generated-reasoning evidence at low data dependence. Within this corpus, retrieval and fact verification appear mainly through the LLM studies rather than as a large independent body, and we distinguish LLM-generated reasoning from evidence-grounded verification (Section 2.5); the reasoning-centric direction is emerging rather than yet dominant.
5. Stage 5, Robust adaptive detection: an emerging stage concerned with domain and temporal adaptation, adversarial robustness, and AI-generated content, driven by the dual role of LLMs as both generators and detectors [15], [35].

Table 11. The five stages of the field's evolution and how each persists in later stages

Stage	Dominant evidence	Enabling technology	Unresolved limitation	Integration with later stages
Feature-centric	Intrinsic lexical/stylistic	Classical ML	Weak generalization	Features feed later models; first-tier filter
Representation-centric	Intrinsic semantic	Pre-trained language models	Dataset shortcuts, low explainability	Backbone for multimodal and LLM methods
Evidence-integration	Cross-modal, social context	Multimodal fusion, GNNs	Cost, late detection, missing modalities	Adds evidence to content models
Knowledge/ reasoning	External knowledge, generated reasoning	LLMs, retrieval	Unstable accuracy, unverified faithfulness	Supplies rationales and low-data coverage
Robust adaptive	All of the above, over time	Adaptation, adversarial training	Drift, AI-generated content	Wraps earlier stages for reliability

These stages are cumulative rather than substitutive: newer stages add evidence sources and adaptivity without retiring earlier ones, which is why every family remains present in the corpus. The central observation is that the field is moving from classifying content that looks fake toward assessing veracity from multiple, complementary sources of evidence and under changing conditions (drift, new languages, attacks, and AI-generated content). This reframing, not the counts themselves, is the contribution of the trend analysis, and it motivates the dependency-based reading of open challenges in Section 4.5.2. The stages are analytical and overlapping (Section 4.5.1); the last column indicates how each stage's contribution persists in the ones that follow, consistent with their cumulative nature.

4.5.2. Unresolved Challenges

Six challenges remain inadequately addressed:

- Challenge 1: Explainability. Deep models are black boxes whose decisions are hard to justify to users or regulators. Partial solutions include attention visualization, LIME, SHAP [70], explainable architectures such as dEFEND [71], LLM-generated rationales [48], and feature-importance analysis [72]. A performance-interpretability trade-off persists.
- Challenge 2: Cross-domain and cross-lingual generalization. Models often degrade markedly when transferred across domains or languages, and, as Section 5 shows, high in-dataset accuracy does not survive a change of task or corpus. Multi-task learning, domain adaptation, and multilingual models (XLM-R [60]) help, but domain-invariant representation remains open.
- Challenge 3: Early detection. Most methods need mature articles and propagation history before reliable classification, introducing latency. Source-credibility assessment, stylometry, early-burst modeling, and reinforcement learning [26] are being explored against a fundamental speed-accuracy trade-off.
- Challenge 4: Adversarial robustness. Small perturbations (synonym substitution, paraphrasing, style transfer) and injected malicious comments [73] fool classifiers without changing meaning. Adversarial training and ensembles help, but the arms race intensifies with LLM-empowered style attacks [35].
- Challenge 5: Scalability and real-time processing. Large models require heavy infrastructure. Mitigations include compression and distillation (DistilBERT, ALBERT [34]), hierarchical pipelines, and distributed processing.
- Challenge 6: AI-generated fake news. LLMs generate fluent misinformation at scale; humans identify only about 54.8% of ChatGPT-generated fake news [14], and machine detectability is not guaranteed [15]. Countermeasures include training detectors on AI-generated samples, statistical fingerprinting, hallucination detection, and content watermarking.

These six challenges are not independent. Table 12 traces them to a small number of root causes in data and evaluation design and to their practical consequences, links each to a research priority (Section 6.3), and shows why they resist simultaneous solution.

Table 12. Root causes, challenges, consequences, and the corresponding research priority

Root cause	Direct challenge	Deployment consequence	Priority
Small, domain-skewed labeled data	Poor generalization	Fails on new topics/languages	P1, P2
Static datasets, random splits	Over-optimistic results	Numbers miss temporal drift	P1
Large, complex models	Low explainability, high cost	Hard to justify and deploy	P3, P5
Dependence on propagation	Late detection	Early spread not prevented	P3
LLM-generated, shifting content	Adversarial attack and drift	Detectors outdated quickly	P4
No unified benchmark protocol	Incomparable results	Progress hard to establish	P1

The challenges share root causes in data and evaluation design; addressing the root causes (left column) is likely more effective than treating each challenge in isolation. P1-P5 are the research priorities of Section 6.3. The challenges also stand in tension, so no single design optimizes all at once: early detection trades against evidence completeness; accuracy trades against explainability; multimodal and graph enrichment trade against cost and latency;

large models trade against low-resource deployability; in-domain specialization trades against generalization; and using an LLM to explain a decision trades against the faithfulness of that explanation. A realistic agenda must therefore choose an order of priorities rather than pursue all objectives at once; Section 6.3 proposes one.

5. Comparative Performance Analysis

5.1. Why Family-Level Accuracy Pooling Is Not Valid

Family-level accuracy is not pooled into a single average, for three reasons. First, metric heterogeneity: studies report Accuracy, F1, macro-F1, or AUC, and these are not interchangeable: none can be derived from another without the confusion matrix, class distribution, or prediction scores, which primary studies rarely provide. Our corpus makes this concrete: the graph family reports both accuracy (84-89%) and AUC (0.75-0.93), quantities that cannot share a scale. Second, dataset and task heterogeneity: even among accuracy-reporting studies, tasks differ (binary vs multi-class), as do dataset size, domain, class balance, and splits. Third, pooling validity: a mean over such studies has no clear estimand, and a between-study standard deviation is not a confidence interval. We therefore report descriptive ranges (Section 5.2) and confine direct comparison to within-benchmark (same-dataset, same-task) settings (Section 5.3).

5.2. Descriptive Summary by Family

Table 13 summarizes, per family, the number of core studies and the reported-accuracy range of the studies that report accuracy; Figure 7 plots the individual values.

Table 13. Descriptive summary of reported accuracy by family (42 core studies)

Family	#Studies (share)	#Reporting accuracy	Reported accuracy range	Other metrics reported
Transformer/PLM	13 (31.0%)	7	74.8-99.36%	F1
Multimodal	11 (26.2%)	10	82.4-99.97%	F1
LLM-based	9 (21.4%)	4	48.6-89.0%	macro-F1
Graph/propagation	7 (16.7%)	4	84.4-94.36%	AUC 0.75-0.93
Traditional ML	2 (4.8%)	2	85.0-88.21%	F1

DESCRIPTIVE ONLY; not a ranking of intrinsic quality. Studies use different datasets, tasks, class distributions, and metrics; no conversion or pooling was performed. The graph family also reports AUC on a different scale. One 30-article study reporting 100% is excluded from the LLM range as unreliable. Per-study values are in the supplementary corpus.

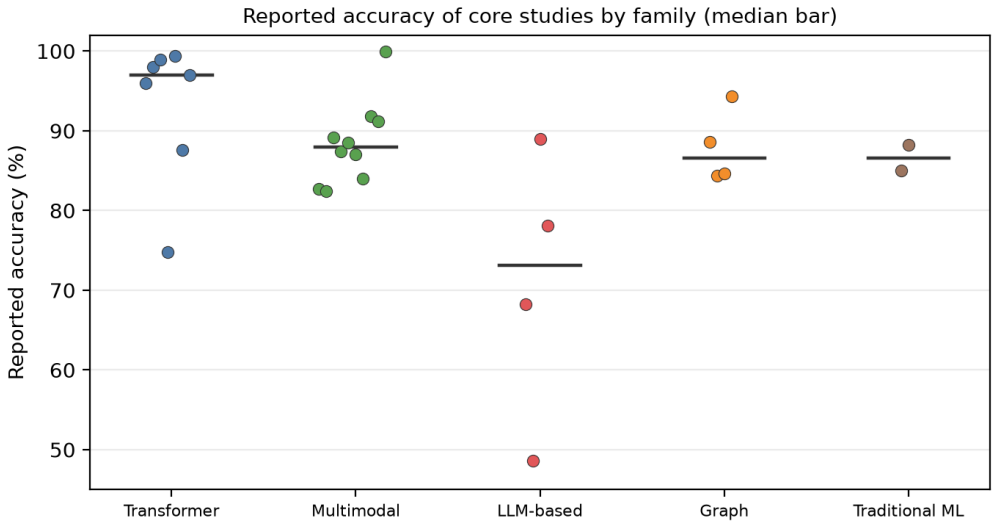

Figure 7. Reported accuracy of individual core studies by family; each point is one study and the short horizontal bar marks the median as a visual reference only, not a ranking. The wide, overlapping spreads, especially for LLM-based methods, show why a single pooled figure per family would be misleading

The reported accuracies overlap across families, and no family separates clearly once dataset heterogeneity is taken into account. The Transformer/PLM range is wide (74.8-99.36%), with its highest values coming from a few benchmark corpora with strong dataset-specific separability. The LLM-based family has the lowest and most dispersed accuracy (48.6-89.0%). Figure 7 shows these spreads; a single pooled figure per family would not represent them.

5.3. Within-Dataset and Within-Task Comparison

Comparison is meaningful only when dataset and task are fixed. Table 14 gathers the within-benchmark groups present in the corpus. Three observations follow. (1) On Weibo (binary), seven multimodal models span 82.4-91.8%, a more comparable within-dataset summary, although differences in data splits and evaluation protocols may remain; the spread across MVAE/EANN (2019) to BMR (2023) does not by itself establish architectural progress, because their splits, preprocessing, and protocols are not verified to be identical. (2) On PolitiFact (binary), trained detectors report higher scores (DKFND 89.0% accuracy; SheepDog 88.4% macro-F1) than a zero-shot GPT-3.5 pipeline (CAPE-FND, 78.1% accuracy); because these entries mix accuracy and macro-F1, only rows sharing the same metric and, ideally, the same split are strictly comparable. (3) On the Vietnamese ReINTEL task, four systems cluster at AUC 0.946-0.958. On LIAR, GPT-4 attains 68.2% on the binary task in a zero-shot setting [13]. Moreover, LIAR-binary and LIAR-6-class are different tasks on the same corpus, with random-chance baselines of 50% and 16.7%; reported drops from binary to six-class therefore reflect intrinsic task difficulty as much as any generalization loss, and must not be read as the latter alone.

Table 14. Within-dataset and within-task comparison (same-dataset, same-task groups in the corpus)

Dataset (task)	Method	Metric	Value	Ref.
Weibo (binary)	MVAE	Accuracy	82.4%	[9]
Weibo (binary)	EANN	Accuracy	82.7%	[8]
Weibo (binary)	CAFE	Accuracy	84.0%	[37]
Weibo (binary)	HMCAN	Accuracy	88.5%	[38]
Weibo (binary)	SpotFake	Accuracy	89.2%	[10]
Weibo (binary)	LLM-enhanced fusion	Accuracy	91.2%	[42]
Weibo (binary)	BMR	Accuracy	91.8%	[39]
PolitiFact (binary)	CAPE-FND (GPT-3.5)	Accuracy	78.1%	[50]
PolitiFact (binary)	UPFD	Accuracy	84.6%	[43]
PolitiFact (binary)	SheepDog	macro-F1	88.4%	[35]
PolitiFact (binary)	DKFND	Accuracy	89.0%	[51]
ReINTEL (binary)	PhoBERT ensemble	AUC	0.9521	[63]
ReINTEL (binary)	Hybrid deep NN	AUC	0.9462	[64]
ReINTEL (binary)	Transfer learning + TF-IDF	AUC	0.9538	[65]
ReINTEL (binary)	vELECTRA + features	AUC	0.9575	[66]

Comparisons are valid only within a single (dataset, task) group. Even within a group, metrics may differ (e.g., PolitiFact mixes accuracy and macro-F1), so rows sharing a metric are the strictly comparable ones. Random-chance baselines: binary 50%, LIAR-6-class 16.7%.

5.4. Mechanisms Behind Reported Performance Differences

In the reviewed corpus, differences in reported performance appear to track data and evaluation design at least as much as model architecture. The main sources of variation, largely orthogonal to family, are: task difficulty (binary versus multi-class, with different chance baselines); class balance; dataset size; event-based versus random splits; the time span of collection; in-domain versus cross-domain evaluation; clean edited text versus noisy social-media text; the presence of near-duplicates; reliance on source or topic cues; the availability of propagation history; and the pre-trained model and language used. Any of these can move a reported number by as much as the difference between two families.

These factors act as a chain: data design -> the kind of signal a model can exploit -> how the data are split -> the metric reported -> transferable performance. The data design determines which signals exist; a model exploits the most predictive of them, which may be source names, topics, or editorial style rather than veracity; the split determines whether the test set shares those signals with training (random, non-event splits usually do, inflating results); the reported metric determines what is even visible; and only the combination determines what transfers. This is why the same architecture can report 99.97% on one dataset and 74.8% on a more heterogeneous content-plus-context setting, the chain, not the model, differs. The most inflating links map directly onto the risk-of-bias rubric of Section 3.6: data leakage, split protocol, and best-of-many result selection (Figure 8).

It is therefore useful to separate three notions the literature often conflates: reported performance (the number a paper states), source-verified reported performance (that number confirmed against the primary source, as done here), and transferable performance (what holds under distribution shift). This review can establish the first two; the third requires the out-of-distribution evaluation prioritized in Section 6.3.

Figure 8. Hypothesized, evidence-consistent chain linking data and evaluation design to reported performance (Section 5.4). The ordering summarizes how the reviewed evidence is interpreted; it is not a demonstrated causal ranking, and architecture remains one factor within it

5.5. Interpretation

Descriptively (Table 13, Figure 7), no family dominates once heterogeneity is taken into account, and the Transformer and LLM extremes correspond to easy and hard evaluation settings respectively. Under within-benchmark conditions (Table 14), multimodal fusion shows measurable progress on Weibo, trained detectors exceed zero-shot LLMs on PolitiFact, and Vietnamese systems reach high AUC on ReINTEL. Model selection should thus be guided by the target task, data availability, and deployment constraints rather than by maximum reported accuracy.

6. Conclusions and Recommendations

6.1. Summary of Main Findings

This review compiled an enumerated, source-verified corpus of 45 studies (42 quantitative-core) on content- and context-based fake news detection, and read it through a multi-axis framework to answer six research questions.

- RQ1, Methods. The core corpus spans five method families, led by Transformer/PLM and multimodal approaches, with LLM-based methods now the third-largest family (full breakdown in Table 13 and Figure 5). Non-exclusive capability tags (Table 8) show most multimodal and all LLM studies also use Transformer components.
- RQ2, What explains performance differences. In this corpus, reported accuracy appears to be shaped mainly by data and evaluation design rather than by architecture alone: task difficulty, split protocol, dataset curation, near-duplicates, and reliance on source or topic cues each can move a number by as much as the gap between two families (Section 5.4), which is consistent with the same architecture reporting very different accuracies on different corpora, as detailed in Section 5.4.
- RQ3, Trade-offs. No family dominates (Table 9): Transformer/PLM offers strong in-domain accuracy but weak explainability and generalization; multimodal helps only when a second modality is informative; graph methods add context but detect late; LLMs are flexible and label-efficient but variable and expensive; traditional models remain valuable under tight resources.

- RQ4, Evolution. The field has moved through five cumulative stages, from feature-centric to robust adaptive detection (Section 4.5.1), shifting from classifying content that looks fake toward assessing veracity from multiple evidence sources under changing conditions.
- RQ5, Shared causes and priorities. The six open challenges share root causes in data and evaluation design (Table 12) and trade off against one another, so they require a prioritized agenda (Section 6.3) rather than simultaneous solution.
- RQ6, Low-resource transfer. Eight Vietnamese studies show a two-sided picture, strong on the shared-task benchmark but markedly weaker on noisier self-collected data (see Table 10 and Section 4.4), and yield transferable design principles for other low-resource languages (Section 4.4.1).

Beyond the individual answers, four synthesis insights stand out. First, in the reviewed evidence, model architecture does not appear to be the main driver of measured performance; data and evaluation design often appear to dominate. Second, the method families are better viewed as complementary evidence sources than as mutually exclusive choices. Third, a realistic deployed system is likely to be a time-adaptive pipeline: content-first for early screening, then adding knowledge, multimodal, and propagation evidence as it becomes available. Fourth, the defining problem of the next period is not maximizing in-domain accuracy but sustaining reliability under drift, new languages, adversarial attacks, and AI-generated content; the Vietnamese case shows that data, infrastructure, and cultural context are system-design constraints, not merely missing models.

6.2. Contributions

Synthesizing these results, the contributions are methodological, a source-verified evidence base, a within-benchmark-only comparison protocol, and a per-study risk-of-bias appraisal, and conceptual: the multi-axis framework, the family trade-off map, the five-stage account of the field's evolution, the dependency-based reading of open challenges, and a set of transferable low-resource design principles. The conceptual contributions, rather than the additional references, distinguish this review from prior surveys (Table 1): they change how the reported evidence is interpreted, not merely how much of it is cataloged.

6.3. A Prioritized Research Agenda and Recommendations

The trade-offs above imply an order of priorities for future research, ranked by their leverage over other challenges, their feasibility, and their importance for reliable deployment.

1. Priority 1, Evaluation that reflects generalization: temporal, event-disjoint, cross-domain, cross-platform, and cross-lingual splits, with uncertainty and multiple runs reported. Without this, apparent progress cannot be trusted.
2. Priority 2, Multilingual and low-resource datasets: high-quality labels, standardized metadata, drift control, and coverage of AI-generated content.
3. Priority 3, Evidence-integrating yet early-detecting models: combining content, source, and knowledge retrieval, updating as propagation accrues, with uncertainty-aware prediction.
4. Priority 4, Robustness and provenance: adversarial evaluation, detection of LLM-generated misinformation, source attribution, and content provenance.
5. Priority 5, Verifiable explanation: rationales grounded in evidence, citation-grounded justifications, and tests of explanation faithfulness.

The ordering is dependency-based, not arbitrary. P1 is placed first because reliable evaluation is a prerequisite for judging progress under all subsequent priorities. P2 follows because robust multilingual and low-resource datasets are needed to operationalize the evaluation protocols of P1. P3-P5 concern model-design and deployment properties whose benefits cannot be established reliably until the evaluation and data foundations of P1-P2 are in place.

These priorities in turn shape practical recommendations. For academic research: report interpretability alongside accuracy; evaluate cross-domain and cross-lingual generalization explicitly; release code, weights, and datasets; verify performance claims against primary sources; and treat adversarial robustness as a first-class criterion. For application

development: combine complementary methods (Transformer + multimodal + graph); keep humans in the loop for high-stakes decisions; monitor and retrain as patterns drift; provide explanations; and use tiered pipelines to balance throughput and accuracy. For policy-making: fund standardized multilingual datasets emphasizing low-resource languages; set accuracy, transparency, and fairness standards; ensure countermeasures preserve free expression; and support media-literacy programs. For education: integrate media literacy and critical thinking into curricula and train journalists and fact-checkers on the capabilities and limits of automated tools.

6.4. Limitations

First, the corpus is enumerated rather than exhaustive (45 studies), so some relevant studies are omitted, and the family shares describe this corpus rather than the entire field. Second, searches were mainly in English and Vietnamese, so Chinese-, Spanish-, and other-language works are underrepresented. Third, seven candidate studies were excluded because their reported metric or citation could not be verified; this improves reliability but may bias toward well-documented work. Fourth, the corpus admits well-cited preprints alongside peer-reviewed papers, trading some formal review for currency. Fifth, reported metrics reflect each study's own datasets and protocols; we therefore avoid pooling and confine comparison to within-benchmark settings, but residual incomparability remains. These limitations point to the future directions under RQ5.

6.5. Threats to Validity

Beyond these limitations, four threats to validity are worth stating explicitly. Selection bias: candidates were assembled purposively and screened by the authors, so relevant studies outside the anchor surveys and browsed venues may be missing, and the requirement that a headline metric be source-verifiable can favor better-documented work; the released corpus and excluded list let readers judge this directly. Extraction bias: reducing each study to a single headline figure, even under the predefined hierarchy of Section 3.3, discards within-study variation, and a different hierarchy could yield somewhat different ranges; we mitigate this by recording the chosen metric, its setting, and its evidence per study. Classification bias: the single-assignment rule of Section 3.4 imposes discrete families on methods that are genuinely hybrid, which the non-exclusive capability tags only partly offset. Publication and reporting bias: the literature over-represents positive, high-accuracy results and under-reports variance, external validation, and negative findings, so the corpus-level picture is likely optimistic. Finally, source verification establishes that a reported number matches its primary source; it does not re-run experiments or audit the underlying data, so it certifies provenance, not internal validity or independent reproducibility.

6.6. Concluding Remarks

Fake news detection is a complex, fast-moving challenge. Deep-learning methods achieve strong benchmark numbers, but these vary with the evaluation dataset and setting; across this corpus, performance is heterogeneous and overlapping, no family dominates, and LLM detectors are so far weak and highly variable. The same LLMs are potent misinformation generators, producing content humans identify only slightly above chance, which makes robust, interpretable, and adaptive detection essential. For Vietnam, sustained investment is needed in high-quality datasets, stronger Vietnamese-specific pre-trained models, and collaboration among academia, industry, and fact-checkers.

Author Contributions: Conceptualization: D.T.L. and L.N.A.; Methodology: L.N.A.; Formal analysis and quantitative synthesis: D.T.D.; Investigation and literature search: D.T.L.; Data curation: D.T.L. and D.T.D.; Writing, original draft preparation: D.T.L.; Writing, review and editing: L.N.A. and D.T.D.; Supervision: L.N.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by Hanoi Open University, Vietnam [grant number MHN2025-02.22], within the framework of the project "A Study on the Development of a

Social Media Post Collection and Classification System: An Application at Hanoi Open University.

Data Availability Statement: The analyzed studies with their reported metrics and sources, the per-study appraisal, the analysis script, and the computed results are provided as supplementary material openly at https://github.com/dtlong1979/fake-news-review-supplement, which provides corpus.csv, excluded.csv, risk_of_bias.csv, results.csv, and analyze.py; the repository README documents the schema and reproduction steps. No new primary datasets were generated. All reviewed studies are publicly accessible through the sources listed in Section 3.1 and corpus.csv.

Acknowledgments: During the preparation of this manuscript, the authors used Claude Opus 4.8 and GitHub Copilot as assistive tools for language review and limited coding support, including wording checks, clarity improvements, terminology consistency, code drafting, debugging, and refac-toring. These tools were not used to design the experiments, make analytical decisions, generate or alter data, fabricate numerical results, or determine the substantive interpretation of the findings. All AI-assisted code suggestions were reviewed, modified where necessary, tested, and executed by the authors. The authors independently verified all analyses, numerical results, and interpretations, and take full responsibility for the final content of the manuscript and the reproducibility of the analyses.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. List of Analyzed Studies

The 45 studies analyzed in this review, with their reported metrics, datasets, tasks, and sources, are provided as supplementary material, together with the analysis script (analyze.py) and the computed summary (results.csv), all openly available at https://github.com/dtlong1979/fake-news-review-supplement.

References

[1] World Economic Forum, "The Global Risks Report 2024," 2024. [Online]. Available: https://www.weforum.org/publications/global-risks-report-2024/

[2] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science (80-.).*, vol. 359, no. 6380, pp. 1146–1151, Mar. 2018, doi: 10.1126/science.aap9559.

[3] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Comput. Surv.*, vol. 53, no. 5, pp. 1–40, 2020, doi: 10.1145/3395046.

[4] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explor. Newsl.*, vol. 19, no. 1, pp. 22–36, 2017, doi: 10.1145/3137597.3137600.

[5] X. Zhang and A. A. Ghorbani, "An overview of online fake news: Characterization, detection, and discussion," *Inf. Process. Manag.*, vol. 57, no. 2, p. 102025, 2020, doi: 10.1016/j.ipm.2019.03.004.

[6] J. Devlin, M.-W. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North*, Stroudsburg, PA, USA: Association for Computational Linguistics, Oct. 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[7] Y. Liu *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv Prepr. arXiv1907.11692*, 2019, [Online]. Available: https://arxiv.org/abs/1907.11692

[8] Y. Wang *et al.*, "EANN: Event adversarial neural networks for multi-modal fake news detection," in *in Proc. 24th ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining (KDD 2018), 2018, pp. 849-8571145/3219819.3219903*, 2018. doi: 10.1145/3219819.3219903.

[9] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "MVAE: Multimodal variational autoencoder for fake news detection," in *in Proc. World Wide Web Conf. (WWW 2019), 2019, pp. 2915-29211145/3308558.3313552*, 2019. doi: 10.1145/3308558.3313552.

[10] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh, "SpotFake: A multi-modal framework for fake news detection," in *in Proc. IEEE 5th Int. Conf. Multimedia Big Data (BigMM 2019), 2019, pp. 39-471109/BigMM.2019.00-44*, 2019. doi: 10.1109/BigMM.2019.00-44.

[11] T. Bian *et al.*, "Rumor detection on social media with bi-directional graph convolutional networks," in *in Proc. AAAI Conf. Artificial Intelligence, vol. 34, no. 01, 2020, pp. 549-5561609/aaai.v34i01.5393*, 2020. doi: 10.1609/aaai.v34i01.5393.

[12] F. Monti, F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein, "Fake News Detection on Social Media Using Geometric Deep Learning," *arXiv Prepr. arXiv1902.06673*, 2019, doi: 10.48550/arXiv.1902.06673.

[13] K. Pelrine *et al.*, "Towards reliable misinformation mitigation: Generalization, uncertainty, and GPT-4," in *in Proc. 2023 Conf. Empirical Methods in Natural Language Processing (EMNLP 2023), 2023, pp. 6399-642918653/v1/2023.emnlp-main.395*, 2023. doi: 10.18653/v1/2023.emnlp-main.395.

[14] Y. Huang and L. Sun, "FakeGPT: Fake news generation, explanation and detection of large language models," 2023. doi: 10.48550/arXiv.2310.05046.

[15] C. Chen and K. Shu, "Can LLM-generated misinformation be detected?," in *in Proc. 12th Int. Conf. Learning Representations (ICLR 2024)*, 2024, p. 2024.
[16] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *PLoS Med.*, vol. 6, no. 7, p. e1000097, 2009, doi: 10.1371/journal.pmed.1000097.
[17] M. J. Page *et al.*, "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *BMJ*, p. n71, Mar. 2021, doi: 10.1136/bmj.n71.
[18] B. Kitchenham, "Procedures for performing systematic reviews," 2004.
[19] S. Gong, R. O. Sinnott, J. Qi, and C. Paris, "Fake news detection through graph-based neural networks: A survey," 2023. doi: 10.48550/arXiv.2307.12639.
[20] C. Comito, L. Caroprese, and E. Zumpano, "Multimodal fake news detection on social media: A survey of deep learning techniques," *Soc. Netw. Anal. Min.*, vol. 13, no. 1, p. 101, 2023, doi: 10.1007/s13278-023-01104-w.
[21] B. Guo, Y. Ding, L. Yao, Y. Liang, and Z. Yu, "The future of false information detection on social media: New perspectives and trends," *ACM Comput. Surv.*, vol. 53, no. 4, pp. 1–36, 2020, doi: 10.1145/3393880.
[22] F. Alam *et al.*, "A survey on multimodal disinformation detection," in *in Proc. 29th Int. Conf. Computational Linguistics (COLING), Gyeongju, Republic of Korea, 2022*, 2022, pp. 6625–6643.
[23] A. B. Athira, S. D. M. Kumar, and A. M. Chacko, "A systematic survey on explainable AI applied to fake news detection," 2023, doi: 10.1016/j.engappai.2023.106087.
[24] C. Chen and K. Shu, "Combating misinformation in the age of LLMs: Opportunities and challenges," *AI Mag.*, vol. 45, no. 3, pp. 354–368, 2024, doi: 10.1002/aaai.12188.
[25] X. Wang, W. Zhang, and S. Rajtmajer, "Monolingual and multilingual misinformation detection for low-resource languages: A comprehensive survey," 2024. doi: 10.48550/arXiv.2410.18390.
[26] R. Yang, X. Wang, Y. Jin, C. Li, J. Lian, and X. Xie, "Reinforcement subgraph reasoning for fake news detection," in *in Proc. 28th ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD 2022), 2022, pp. 2253-22621145/3534678.3539277*, 2022. doi: 10.1145/3534678.3539277.
[27] R. Kumari and A. Ekbal, "AMFB: Attention based multimodal factorized bilinear pooling for multimodal fake news detection," 2021, doi: 10.1016/j.eswa.2021.115412.
[28] B. Xie *et al.*, "Multi-knowledge and LLM-inspired heterogeneous graph neural network for fake news detection," *IEEE Trans. Comput. Soc. Syst.*, vol. 12, no. 2, pp. 682–694, 2025, doi: 10.1109/TCSS.2024.3488191.
[29] H. Yuan, J. Zheng, Q. Ye, Y. Qian, and Y. Zhang, "Improving fake news detection with domain-adversarial and graph-attention neural network," 2021, doi: 10.1016/j.dss.2021.113633.
[30] C.-O. Truică and E.-S. Apostol, "It's All in the Embedding! Fake News Detection Using Document Embeddings," *Mathematics*, vol. 11, no. 3, p. 508, 2023, doi: 10.3390/math11030508.
[31] S. Raza and C. Ding, "Fake news detection based on news content and social contexts: a transformer-based approach," *Int. J. Data Sci. Anal.*, vol. 13, no. 4, pp. 335–362, 2022, doi: 10.1007/s41060-021-00302-z.
[32] R. K. Kaliyar, A. Goswami, and P. Narang, "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach," *Multimed. Tools Appl.*, vol. 80, no. 8, pp. 11765–11788, Mar. 2021, doi: 10.1007/s11042-020-10183-2.
[33] J. Alghamdi, Y. Lin, and S. Luo, "Towards COVID-19 fake news detection using transformer-based models," 2023, doi: 10.1016/j.knosys.2023.110642.
[34] S. F. N. Azizah, H. D. Cahyono, S. W. Sihwi, and W. Widiarto, "Performance analysis of transformer based models (BERT, ALBERT and RoBERTa) in fake news detection," 2023. doi: 10.1109/ICOIACT59844.2023.10455849.
[35] J. Wu, J. Guo, and B. Hooi, "Fake news in sheep's clothing: Robust fake news detection against LLM-empowered style attacks," in *in Proc. 30th ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD 2024), 2024*, 2024, pp. 3367–3378. doi: 10.1145/3637528.3671977.
[36] X. Zhou, J. Wu, and R. Zafarani, "SAFE: Similarity-aware multi-modal fake news detection," 2020, doi: 10.1007/978 3-030-47436-2_27.
[37] Y. Chen *et al.*, "Cross-modal ambiguity learning for multimodal fake news detection," in *in Proc. ACM Web Conf. (WWW 2022), 2022, pp. 2897-29051145/3485447.3511968*, 2022. doi: 10.1145/3485447.3511968.
[38] S. Qian, J. Wang, J. Hu, Q. Fang, and C. Xu, "Hierarchical multi-modal contextual attention network for fake news detection," in *in Proc. 44th Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR 2021), 2021, pp. 153-1621145/3404835.3462871*, 2021. doi: 10.1145/3404835.3462871.
[39] Q. Ying, X. Hu, Y. Zhou, Z. Qian, D. Zeng, and S. Ge, "Bootstrapping multi-view representations for fake news detection," in *in Proc. AAAI Conf. Artificial Intelligence, vol. 37, no. 4, 2023, pp. 5384-53921609/aaai.v37i4.25670*, 2023. doi: 10.1609/aaai.v37i4.25670.
[40] I. Segura-Bedmar and S. Alonso-Bartolome, "Multimodal Fake News Detection," *Information*, vol. 13, no. 6, p. 284, Jun. 2022, doi: 10.3390/info13060284.
[41] M. Al-alshaqi, D. B. Rawat, and C. Liu, "A BERT-based multimodal framework for enhanced fake news detection using text and image data fusion," *Computers*, vol. 14, no. 6, p. 237, 2025, doi: 10.3390/computers14060237.
[42] J. Wang, Z. Zhu, C. Liu, R. Li, and X. Wu, "LLM-enhanced multimodal detection of fake news," *PLoS One*, vol. 19, no. 10, p. e0312240, 2024, doi: 10.1371/journal.pone.0312240.
[43] Y. Dou, K. Shu, C. Xia, P. S. Yu, and L. Sun, "User preference-aware fake news detection," in *in Proc. 44th Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR 2021), 2021, pp. 2051-20551145/3404835.3462990*, 2021. doi: 10.1145/3404835.3462990.
[44] N. Mehta, M. L. Pacheco, and D. Goldwasser, "Tackling fake news detection by continually improving social context representations using graph neural networks," in *in Proc. 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022), 2022, pp. 1363-138018653/v1/2022.acl-long.97*, 2022. doi: 10.18653/v1/2022.acl-long.97.

[45] V.-H. Nguyen, K. Sugiyama, P. Nakov, and M.-Y. Kan, "FANG: Leveraging social context for fake news detection using graph representation," in *in Proc. 29th ACM Int. Conf. Information and Knowledge Management (CIKM 2020), 2020, pp. 1165-11741145/3340531.3412046*, 2020. doi: 10.1145/3340531.3412046.

[46] Y. Han, S. Karunasekera, and C. Leckie, "Graph Neural Networks with Continual Learning for Fake News Detection from Social Media," *arXiv Prepr. arXiv2007.03316*, 2020, [Online]. Available: https://arxiv.org/abs/2007.03316

[47] S. Koka, A. Vuong, and A. Kataria, "Evaluating the efficacy of large language models in detecting fake news: A comparative analysis," 2024. doi: 10.48550/arXiv.2406.06584.

[48] B. Hu *et al.*, "Bad actor, good advisor: Exploring the role of large language models in fake news detection," in *in Proc. AAAI Conf. Artificial Intelligence, vol. 38, no. 20, 2024, pp. 22105-221131609/aaai.v38i20.30214*, 2024. doi: 10.1609/aaai.v38i20.30214.

[49] Q. Nan, Q. Sheng, J. Cao, B. Hu, D. Wang, and J. Li, "Let silence speak: Enhancing fake news detection with generated comments from large language models," 2024. doi: 10.1145/3627673.3679519.

[50] W. Jin *et al.*, "Veracity-oriented context-aware large language models-based prompting optimization for fake news detection," 2025, doi: 10.1155/int/5920142.

[51] Y. Liu *et al.*, "Detect, Investigate, Judge and Determine: A Knowledge-Guided Framework for Few-Shot Fake News Detection," in *Proceedings of the 25th IEEE International Conference on Data Mining (ICDM 2025)*, 2025, pp. 477–486. doi: 10.1109/ICDM65498.2025.00055.

[52] J. Y. Khan, M. T. I. Khondaker, S. Afroz, G. Uddin, and A. Iqbal, "A benchmark study of machine learning models for online fake news detection," *Mach. Learn. with Appl.*, vol. 4, p. 100032, Jun. 2021, doi: 10.1016/j.mlwa.2021.100032.

[53] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media," *Big Data*, vol. 8, no. 3, pp. 171–188, 2020, doi: 10.1089/big.2020.0062.

[54] W. Y. Wang, "Liar, liar pants on fire: A new benchmark dataset for fake news detection," in *in Proc. 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017), 2017, pp. 422-42618653/v1/P17-2067*, 2017. doi: 10.18653/v1/P17-2067.

[55] K. Nakamura, S. Levy, and W. Y. Wang, "Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection," in *in Proc. 12th Language Resources and Evaluation Conf. (LREC 2020), 2020*, 2020, pp. 6149–6157.

[56] L. Cui and D. Lee, "CoAID: COVID-19 Healthcare Misinformation Dataset," *arXiv Prepr. arXiv2006.00885*, 2020, [Online]. Available: https://arxiv.org/abs/2006.00885

[57] P. K. Verma, P. Agrawal, I. Amorim, and R. Prodan, "WELFake: Word Embedding Over Linguistic Features for Fake News Detection," *IEEE Trans. Comput. Soc. Syst.*, vol. 8, no. 4, pp. 881–893, Aug. 2021, doi: 10.1109/TCSS.2021.3068519.

[58] D. S. Nielsen and R. McConville, "MuMiN: A large-scale multilingual multimodal fact-checked misinformation social network dataset," in *in Proc. 45th Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR 2022), 2022, pp. 3141-31531145/3477495.3531744*, 2022. doi: 10.1145/3477495.3531744.

[59] D. Q. Nguyen and A. T. Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," 2020, doi: 10.18653/v1/2020.findings-emnlp.92.

[60] A. Conneau *et al.*, "Unsupervised Cross-lingual Representation Learning at Scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 8440–8451. doi: 10.18653/v1/2020.acl-main.747.

[61] C.-V. Nguyen Thi, T.-T. Vuong, D.-T. Le, and Q.-T. Ha, "v3MFND: A Deep Multi-Domain Multimodal Fake News Detection Model for Vietnamese," in *Asian Conference on Intelligent Information and Database Systems (ACIIDS 2022)*, in Lecture Notes in Artificial Intelligence, vol. 13757. 2022, pp. 608–620. doi: 10.1007/978-3-031-21743-2_49.

[62] A.-T. Huynh and P. Tran, "Utilizing transformer models to detect Vietnamese fake news on social media platforms," *KSII Trans. Internet Inf. Syst.*, vol. 19, no. 2, pp. 472–487, 2025, doi: 10.3837/tiis.2025.02.006.

[63] T. N. Hieu, H. C. N. Minh, H. T. Van, and B. V Quoc, "ReINTEL challenge 2020: Vietnamese fake news detection using ensemble model with PhoBERT embeddings," in *in Proc. 7th Workshop on Vietnamese Language and Speech Processing (VLSP 2020), 2020*, 2020, pp. 1–5.

[64] H. V. Trinh *et al.*, "ReINTEL Challenge 2020: A Comparative Study of Hybrid Deep Neural Network for Reliable Intelligence Identification on Vietnamese SNSs," in *Proceedings of the 7th International Workshop on Vietnamese Language and Speech Processing*, Hanoi, Vietnam: Association for Computational Linguistics, 2020, pp. 6–12. [Online]. Available: https://aclanthology.org/2020.vlsp-1.2/

[65] N.-D. Pham, T.-H. Le, T.-D. Do, T.-T. Vuong, T.-H. Vuong, and Q.-T. Ha, "Vietnamese fake news detection based on hybrid transfer learning model and TF-IDF," in *in Proc. 2021 13th Int. Conf. on Knowledge and Systems Engineering (KSE), 2021, pp. 1-61109/KSE53942.2021.9648676*, 2021. doi: 10.1109/KSE53942.2021.9648676.

[66] K. D. Pham, D. Van Thin, and N. L.-T. Nguyen, "Improving Vietnamese Fake News Detection Based on Contextual Language Model and Handcrafted Features," *Sci. Technol. Dev. J.*, vol. 26, no. 2, pp. 2705–2712, 2023, doi: 10.32508/stdj.v26i1.3927.

[67] T. H. Vo, T. L. T. Phan, and K. C. Ninh, "Development of a fake news detection tool for Vietnamese based on deep learning techniques," *Eastern-European J. Enterp. Technol.*, vol. 5, no. 2 (119), pp. 14–20, 2022, doi: 10.15587/1729-4061.2022.265317.

[68] T. N. Le, K. T. B. Nguyen, and N. T. V Khanh, "An application of machine learning models in detecting Vietnamese fake news," *Int. J. Softw. Innov.*, vol. 12, no. 1, pp. 1–15, 2024, doi: 10.4018/IJSI.362623.

[69] H. Rama Moorthy *et al.*, "Dual Stream Graph Augmented Transformer Model Integrating BERT and GNNs for Context Aware Fake News Detection," *Sci. Rep.*, vol. 15, p. 25436, 2025, doi: 10.1038/s41598-025-05586-w.

[70] M. Szczepanski, M. Pawlicki, R. Kozik, and M. Choras, "New explainability method for BERT-based model in fake news detection," 2021, doi: 10.1038/s41598-021-03100-6.

[71] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, "dEFEND: Explainable fake news detection," in *in Proc. 25th ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining (KDD 2019), 2019, pp. 395-4051145/3292500.3330935*, 2019. doi: 10.1145/3292500.3330935.

[72] E. Hashmi, S. Y. Yayilgan, M. M. Yamin, S. Ali, and M. Abomhara, "Advancing fake news detection: Hybrid deep learning with FastText and explainable AI," 2024, doi: 10.1109/ACCESS.2024.3381038.

[73] T. Le, S. Wang, and D. Lee, "MALCOM: Generating malicious comments to attack neural fake news detection models," in *in Proc. IEEE Int. Conf. Data Mining (ICDM 2020), 2020, pp. 282-2911109/ICDM50108.2020.00037*, 2020. doi: 10.1109/ICDM50108.2020.00037.