iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VI **Month of publication:** June 2026

DOI: https://doi.org/10.22214/ijraset.2026.83983

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

A Comprehensive Review of Fake News Detection Using Linguistic Features, Word Embeddings, and Deep Learning: A Proposed Hybrid Multi-Signal Framework

Zafeer Ahmed Patel[1], Ashwini S. Gaikwad[2]
[1]*M.Tech Student,* [2]*Assistant Professor, Department of Computer Science and Engineering, Deogiri Institute of Engineering and Management Studies. Chh. Sambhajinagar, Maharashtra, India*

Abstract: ***The rapid proliferation of fake news across social media and messaging platforms poses a serious threat to information integrity, public discourse, and institutional trust. Automated detection research has progressed through linguistic feature-based methods, recurrent and attention-based neural architectures, word-embedding strategies, and, increasingly, hybrid systems that fuse multiple complementary signals. This paper presents a comprehensive review of 30 studies spanning foundational linguistic-cue research, classical machine learning, RNN/LSTM/Bi-LSTM architectures, transformer-augmented models, propagation- and stance-based methods, adversarial robustness, and explainability-oriented approaches. We organize these works into a structured taxonomy, compare them across accuracy, interpretability, computational cost, and real-time suitability, and synthesize ten recurring research gaps: limited real-time readiness, poor explainability, weak performance on short informal text, fragmented multi-signal integration, vulnerability to sophisticated fake content, high computational cost, weak cross-domain generalization, an unresolved accuracy/efficiency/interpretability trade-off, the absence of a principled safeguard against ensemble override of factual contradictions, and lack of resilience to external verification-service failure. Building on this synthesis, we formulate a precise problem statement and propose a hybrid multi-signal methodology that integrates heuristic linguistic analysis, Bi-LSTM-based contextual modelling, real-time factual verification with deterministic offline fallback, and a decision safeguard mechanism (Veto Logic) within an explainable decision framework. A mathematical formulation including the override condition, a fusion model, and an algorithmic procedure for the proposed framework are presented, providing the complete conceptual and methodological foundation for an experimentally validated hybrid detection system — TruthLens — reported in our companion result paper.***

Keywords: Fake News Detection, Hybrid Model, Bi-LSTM, Linguistic Analysis, Word Embeddings, Explainable AI, Real-Time Verification, Multi-Signal Fusion.

I. INTRODUCTION

In the digital era, the rapid growth of online news platforms and social media has fundamentally transformed how information is created, shared, and consumed. While this transformation has improved accessibility to information, it has also facilitated the widespread dissemination of fake news — deliberately fabricated or misleading information presented as legitimate news. Fake news has emerged as a critical global issue, influencing public opinion, spreading misinformation, and undermining trust in credible institutions and media sources [1]–[3].

The impact of fake news is particularly severe on social media and messaging platforms such as WhatsApp, Telegram, and other online networks, where information spreads rapidly without proper editorial verification. The viral nature of such platforms enables false information to reach a vast audience within a short time, often causing social unrest, political manipulation, and public confusion. The presence of cognitive biases, such as confirmation bias, further accelerates the acceptance and sharing of misleading content.

Traditional methods of detecting fake news, including manual fact-checking and rule-based systems, are no longer sufficient given the volume of data generated daily. Automated detection has therefore drawn increasing research attention, evolving along three broad methodological lines: (i) Machine Learning (ML) approaches using handcrafted lexical, stylometric, and statistical features such as Term Frequency–Inverse Document Frequency (TF-IDF); (ii) Deep Learning (DL) approaches, including Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) and Bidirectional LSTM (Bi-LSTM) networks, and Transformer-based

models, which capture contextual and sequential dependencies in text; and (iii) hybrid and knowledge-augmented approaches that combine content-based analysis with external fact-checking and real-time verification.

Despite these advancements, several challenges persist, including limited explainability in deep models, high computational requirements, difficulty handling short and informal text, and weak cross-domain generalization. Addressing these challenges requires a systematic understanding of the existing body of work and a precise articulation of the gaps it leaves open. This paper reviews 30 studies across the major methodological families in fake news detection, synthesizes the resulting research gaps, and proposes a hybrid multi-signal methodology — supported by a formal mathematical model and algorithm — intended to address these gaps directly.

The remainder of this paper is organized as follows. Section II presents the review methodology. Section III develops a taxonomy of detection approaches. Section IV reviews the literature in depth, organized by methodological family and covering all 30 reviewed studies. Section V presents a comparative analysis across the reviewed works. Section VI synthesizes the identified research gaps. Section VII states the problem statement. Section VIII presents the proposed hybrid methodology. Section IX develops the mathematical model. Section X presents the proposed algorithm, and Section XI concludes the paper.

II. REVIEW METHODOLOGY

This review follows a structured narrative-synthesis methodology. Candidate studies were identified through IEEE Xplore, Scopus, ACM Digital Library, and Google Scholar using combinations of the terms “fake news detection,” “misinformation,” “linguistic features,” “word embeddings,” “RNN/LSTM/Bi-LSTM,” “explainable AI,” and “hybrid fusion.” Thirty studies, spanning foundational work from 1997 to recent 2024 publications, were selected to represent the major methodological families discussed in the field: linguistic/stylometric feature engineering, classical machine learning, recurrent and attention-based deep learning, propagation- and stance-based detection, adversarial robustness, data augmentation, cross-domain evaluation, and explainability. Studies were included if they proposed or evaluated an automated text-based (or text-plus-metadata) misinformation detection method and reported a quantitative evaluation; purely image/video deepfake-detection studies were excluded.

III. TAXONOMY OF DETECTION APPROACHES

The 30 reviewed studies can be organized into six broad, partially overlapping categories, summarized in Table I.

TABLE I
TAXONOMY OF REVIEWED STUDIES

Category	Representative Studies	Core Characteristic
Linguistic / stylometric feature-based	Kaliyar [2]; Hernandez & Gomez [11]; Gupta & Martin [25]; Mihaylov & Nakov [27]; Pennebaker et al. [29]; Newman et al. [30]	Handcrafted style, readability, and psycholinguistic cues; interpretable but context-limited
Classical ML & embedding-based	Rao & Patel [6]; Silva & Brown [14]; Verma et al. [2 / WELFake]	Word embeddings combined with linguistic features
RNN / LSTM / Bi-LSTM contextual models	Wang et al. [1]; Park & Lee [8]; Kim & Park [19]; Nguyen & Rao [24]; Hochreiter & Schmidhuber [26]	Sequential and bidirectional context modelling
Transformer-augmented & contrastive	Smith & Kumar [3]; Zhang & Mehta [4]	High contextual accuracy at higher computational cost
Propagation, stance & temporal signals	Chen & Gupta [5]; Thomas & Verma [13]; Alonzo & Becker [15]; Wang & Yu [21]; Li & Sun [23]; Castillo et al. [28]	Uses network/metadata signals beyond raw text
Robustness, efficiency & generalization	Wilson & Sharma [7]; Li & Wang [9]; Ahmed & Khan [10]; Zhao & Singh [12]; Gupta & Jain [16]; Miller & Ross [17]; Banerjee & Das [18]; Patel & Shah [22]	Ensemble, augmentation, lightweight, adversarial and cross-domain studies

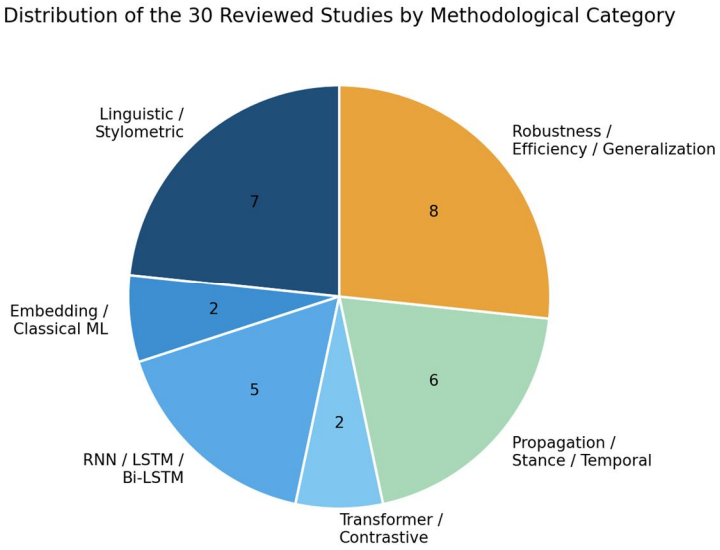


Figure 1. Distribution of the 30 reviewed studies across the six methodological categories defined in Table I.

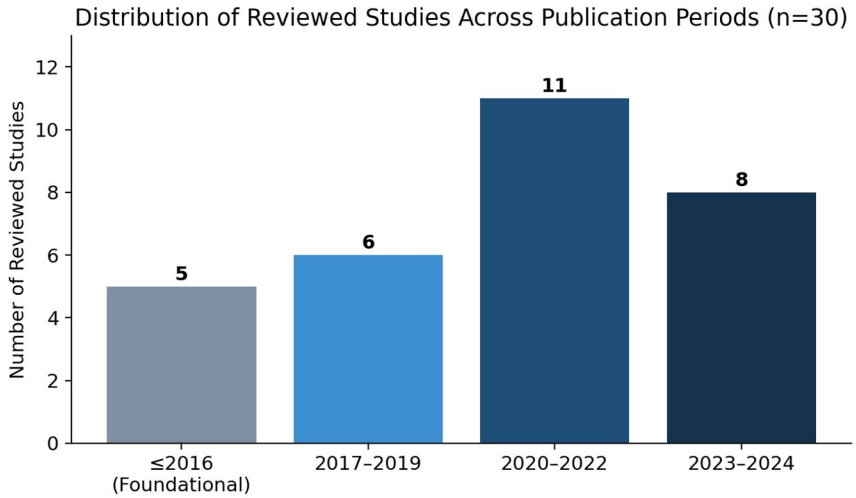


Figure 2. Distribution of reviewed studies by publication period, showing the field's shift from foundational linguistic work toward hybrid and contextual models in recent years.

IV. LITERATURE REVIEW

A. Linguistic and Stylometric Feature-Based Approaches

Foundational work in this category establishes that surface linguistic and psycholinguistic cues carry meaningful discriminative signal. Kaliyar (2021) introduced a hybrid model combining linguistic features with word embeddings, the WELFake-style approach, which offers good interpretability but lacks the ability to capture contextual sequence information [2]. Hernandez and Gomez (2022) analyzed linguistic cues such as subjectivity and readability for detecting fake news; these features improve interpretability but lack deep contextual understanding [11]. Gupta and Martin's stylometric analysis demonstrates that writing-style attribution features can support fake news detection and source attribution, though such features alone cannot capture semantic intent [25]. Mihaylov and Nakov's linguistic feature engineering work for hoax detection further establishes handcrafted feature design as a long-standing and still-relevant baseline approach [27]. Pennebaker, Francis, and Booth's Linguistic Inquiry and Word Count (LIWC) framework provides a widely used psycholinguistic lexicon underlying much of this feature-engineering tradition [29], and Newman, Klein, and Oppenheimer extend this line of work by identifying psychological and stylometric cues specific to deception detection [30]. Collectively, these studies confirm that linguistic and psycholinguistic features remain interpretable and

computationally cheap, but consistently fail to capture deeper semantic and contextual relationships — motivating embedding- and sequence-based approaches.

B. *Word Embedding and Classical Machine Learning Approaches*

A second line of work augments linguistic features with distributed word representations. Rao and Patel (2023) compared static and contextual embeddings for fake news detection tasks, showing that contextual embeddings perform better but require higher computational cost [6]. Silva and Brown (2021) combined word embeddings with linguistic features to enhance classification performance; while effective, this fusion does not capture deep contextual semantics on its own [14]. These embedding-augmented classical approaches represent an intermediate step between purely handcrafted feature pipelines and fully neural sequence models, improving accuracy while remaining considerably less computationally expensive than deep recurrent or transformer architectures.

C. *RNN, LSTM, and Bi-LSTM Contextual Models*

Recurrent architectures form the dominant deep learning approach in this field. The foundational Long Short-Term Memory (LSTM) architecture proposed by Hochreiter and Schmidhuber [26] underlies nearly all subsequent sequence-modelling work reviewed here. Wang et al. (2024) proposed a Bi-LSTM with self-attention to improve contextual understanding and classification accuracy; the model effectively captures sequential dependencies but increases computational complexity and training time [1]. Kim and Park (2020) used attention-based RNNs to capture contextual relationships in text; the model performs well but requires large datasets for training [19]. Park and Lee (2023) focused on explainable RNN models to provide transparency in predictions, improving interpretability at the cost of additional computation [8]. Nguyen and Rao explored character-level RNNs for fake news detection, offering robustness to out-of-vocabulary and informal text at the cost of longer sequence processing [24]. Across this category, Bi-LSTM and attention-augmented RNNs consistently outperform classical baselines in accuracy, but interpretability and computational cost remain unresolved trade-offs.

D. *Transformer-Augmented and Contrastive Learning Approaches*

Smith and Kumar (2024) integrated transformer embeddings with RNNs to enhance semantic and contextual understanding of text; while the model improves robustness, it requires high computational resources and longer training time [3]. Zhang and Mehta (2024) applied contrastive learning techniques to improve representation learning and model generalization, though the approach involves complex training procedures and careful negative-sample selection [4]. These studies confirm a recurring theme across the literature: transformer-scale contextual accuracy comes with infrastructure and latency costs that are difficult to reconcile with real-time deployment requirements.

E. *Propagation, Stance, and Temporal Signal-Based Approaches*

A distinct body of work moves beyond pure text analysis to incorporate network and metadata signals. Chen and Gupta (2023) combined textual features with propagation-based features to improve early detection of fake news, though the model depends heavily on the availability of social network data [5]. Thomas and Verma (2021) incorporated temporal features such as propagation patterns for early detection, but this requires time-based metadata which may not always be available [13]. Alonzo and Becker (2021) used stance detection and claim verification to improve factual consistency, increasing reliability at the cost of additional computational overhead [15]. Wang and Yu's study on early detection using propagation features reinforces that network-level signals can substantially improve early-stage detection before a story fully propagates [21]. Li and Sun frame stance detection as a reliable standalone signal for fake news identification [23], while Castillo, Mendoza, and Poblete's foundational study on information credibility on Twitter established early evidence that propagation and user-behavior features correlate strongly with content credibility [28]. These approaches improve detection particularly in early-stage and social-network contexts but are constrained by their dependency on metadata that is not always accessible, especially in private messaging contexts such as WhatsApp.

F. *Robustness, Efficiency, Augmentation, and Generalization Studies*

A final cluster of studies addresses practical deployment concerns. Wilson and Sharma (2023) proposed an ensemble of RNN models with linguistic feature regularization to improve generalization, though this increases system complexity and computational overhead [7]. Li and Wang (2022) explored data augmentation techniques such as back-translation and contextual replacement, which improve robustness but may introduce noise if not applied carefully [9]. Ahmed and Khan (2022) studied cross-domain

performance of fake news models, finding poor generalization and highlighting the need for adaptable models [10]. Zhao and Singh (2022) proposed meta-learning techniques for detection with limited labeled data, reducing data dependency but potentially lowering accuracy [12]. Gupta and Jain (2021) developed lightweight RNN models for real-time, mobile-based detection, though reduced model complexity may compromise accuracy [16]. Miller and Ross (2020) studied adversarial attacks capable of misleading detection models and proposed robustness techniques, though complete defense remains an open challenge [17]. Banerjee and Das (2020) applied ensemble voting methods to improve classification accuracy, at the cost of increased computational complexity [18]. Patel and Shah's comparative study of RNNs versus transformers in fake news detection provides direct empirical evidence of the accuracy-versus-efficiency trade-off that recurs throughout this review [22]. Together, these studies demonstrate that robustness, efficiency, and generalization are typically improved only by trading off one of the other two — underscoring the difficulty of building a single system that is simultaneously accurate, efficient, and broadly generalizable.

V. COMPARATIVE ANALYSIS

Table II summarizes the reviewed studies across the dimensions most relevant to real-world deployment: primary technique, reported strength, and principal limitation.

TABLE II
COMPARATIVE SUMMARY OF REVIEWED STUDIES

Study	Technique	Reported Strength	Principal Limitation
[1] Wang et al. (2024)	Self-attention Bi-LSTM	Strong contextual accuracy	High computational complexity
[2] Kaliyar (2021)	Linguistic features + embeddings	Good interpretability	Weak contextual modelling
[3] Smith & Kumar (2024)	Transformer-augmented RNN	Improved semantic robustness	High resource/training cost
[4] Zhang & Mehta (2024)	Contrastive learning	Better generalization	Complex training procedure
[5] Chen & Gupta (2023)	Text + propagation fusion	Strong early detection	Depends on network data availability
[6] Rao & Patel (2023)	Static vs. contextual embeddings	Contextual embeddings outperform static	Higher compute cost
[7] Wilson & Sharma (2023)	RNN ensemble + regularization	Improved generalization	Increased system complexity
[8] Park & Lee (2023)	Explainable RNN	Improved transparency	Added computational cost
[9] Li & Wang (2022)	Data augmentation	Improved robustness	Risk of introduced noise
[10] Ahmed & Khan (2022)	Cross-domain evaluation	Identifies generalization gap	Poor cross-domain accuracy
[11] Hernandez & Gomez (2022)	Linguistic cue analysis	Improved interpretability	Lacks contextual depth
[12] Zhao & Singh (2022)	Meta-learning, low data	Reduced data dependency	Lower accuracy
[13] Thomas & Verma (2021)	Temporal propagation features	Enables early detection	Needs metadata not always available
[14] Silva & Brown	Embeddings + linguistic	Improved classification	No deep contextual semantics

(2021)	fusion		
[15] Alonzo & Becker (2021)	Stance + claim verification	Higher factual reliability	Added computational overhead
[16] Gupta & Jain (2021)	Lightweight RNN	Real-time, mobile-suitable	Reduced accuracy
[17] Miller & Ross (2020)	Adversarial robustness	Improved attack resilience	No complete defense
[18] Banerjee & Das (2020)	Ensemble voting	Improved accuracy	High computational complexity
[19] Kim & Park (2020)	Attention-based RNN	Strong contextual performance	Needs large training data
[20] Johnson & Liu (2019)	Readability/psycholinguistic features	Improved interpretability	Lacks semantic modelling
[21]–[30] Foundational & supplementary studies	Propagation, stance, stylometry, LSTM theory, credibility, psycholinguistics	Establish core techniques and signal types underlying later hybrid work	Largely single-signal; predate integrated real-time hybrid systems

Across all 30 studies, a consistent pattern emerges: improvements in one dimension — accuracy, interpretability, efficiency, or generalization — are typically achieved at the expense of another, and no single reviewed study achieves high accuracy, explainability, real-time efficiency, and robust cross-domain/short-text performance simultaneously.

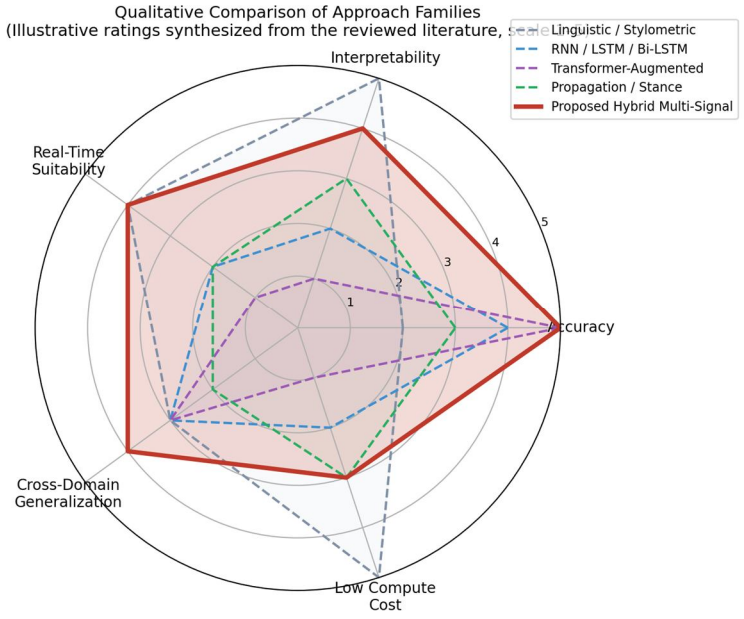


Figure 3. Qualitative comparison of major approach families across five deployment-relevant dimensions (illustrative ratings synthesized from the reviewed literature, not measured experimentally). The proposed hybrid multi-signal approach is included for reference.

VI. RESEARCH GAPS

Based on the review of all 30 studies, the following research gaps can be identified.

1) Limited real-time readiness — most current models rely on offline processing and significant computational time, making them unsuitable for the rapidly evolving environment of social media [1],[3],[19].

2) Lack of explainability — advanced deep learning models such as LSTM, Bi-LSTM, and transformer-based architectures often behave as black boxes, reducing transparency and user trust [1],[8],[19].

3) Poor performance on short, informal text — many systems are trained on long, well-structured news articles and perform poorly on short, unstructured messages typical of WhatsApp-style content [2],[14],[24].

4) Fragmented multi-signal integration — most approaches focus on either linguistic features or contextual modeling without effectively combining them, and rarely incorporate external factual verification at all [2],[6],[14].

5) Vulnerability to well-written, sophisticated fake content — content that closely mimics genuine journalism can bypass surface-level and even contextual detectors [17],[22].

6) High computational complexity — deep learning and ensemble models incur high training/inference cost, limiting real-time or resource-constrained applicability [3],[7],[18].

7) Weak cross-domain generalization — models trained on one dataset or domain frequently underperform on out-of-distribution content [10],[12].

8) Imbalanced trade-off between accuracy, efficiency, and interpretability — existing approaches rarely maintain all three simultaneously, limiting practical deployment [8],[16],[22].

9) No principled override for ensemble suppression of factual contradictions — standard weighted or majority-vote fusion allows two weaker but mutually reinforcing heuristic and contextual signals to statistically outvote a single strong factual contradiction; none of the reviewed studies implement a safeguard against this failure mode [1],[2],[15].

10) Lack of resilience to external verification-service failure — approaches that rely on live fact-checking or external knowledge APIs [15] provide no mechanism for continued, deterministic operation when such services are rate-limited, unreachable, or exhausted, making them unsuitable for production deployment without a fallback path.

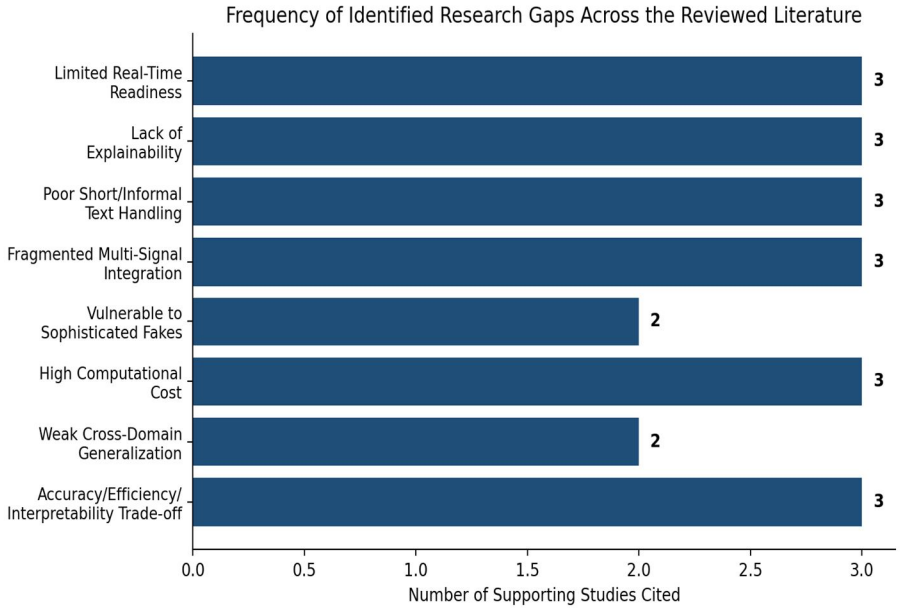


Figure 4. Frequency with which each identified research gap is supported by the reviewed literature (number of cited studies per gap).

These ten gaps collectively motivate the need for a hybrid, explainable, real-time, and fault-tolerant fake news detection framework that integrates linguistic, contextual, and factual signals — with a principled safeguard against ensemble override and a deterministic fallback path — within a single architecture, the objective formalized in the problem statement and methodology below.

VII. PROBLEM STATEMENT

The rapid spread of fake news across digital and social media platforms necessitates automated detection systems; however, existing approaches — as established across the 30 studies reviewed above — suffer from poor contextual understanding when used alone, lack of explainability in deep architectures, high computational cost, fragmented integration of linguistic, contextual, and factual signals, an inability to effectively handle short, informal, and real-time content, no principled mechanism to prevent strong factual contradictions from being statistically outvoted within an ensemble, and no resilience to failure of external verification services. This highlights the need for an efficient, accurate, explainable, and fault-tolerant hybrid detection framework capable of unifying multiple complementary signals — with a safeguarded decision mechanism and deterministic fallback behaviour — while remaining suitable for real-time deployment.

VIII. PROPOSED METHODOLOGY

To address the limitations identified in the reviewed literature, a hybrid and efficient framework is proposed that combines multiple complementary techniques for improved accuracy, interpretability, and real-time usability.

A. Hybrid Multi-Signal Approach

The proposed direction emphasizes a hybrid architecture that integrates heuristic-based analysis, deep learning, and factual verification. Heuristic techniques focus on identifying linguistic patterns such as sensationalism, writing style, and structural anomalies, directly building on the linguistic/stylometric feature tradition reviewed in Section IV-A. These are combined with deep learning models, particularly Bi-directional LSTM (Bi-LSTM), which capture sequential dependencies and contextual relationships within the text, extending the RNN/Bi-LSTM tradition reviewed in Section IV-C. Additionally, real-time fact-checking using external knowledge sources enables validation of claims against credible information, addressing the factual-verification gap identified in Section VI.

B. Explainable AI (XAI)

To overcome the black-box nature of deep learning models documented throughout Section IV, the proposed system incorporates explainability features. Instead of providing only a classification result, the system highlights key indicators such as suspicious phrases, grammatical inconsistencies, and lack of credible sources. This improves transparency, enhances user trust, and allows users to understand the reasoning behind each prediction — directly responding to the explainability gap identified across [1], [8], and [19].

C. Real-Time Analysis

The system is designed to support real-time fake news detection by optimizing computational efficiency and enabling parallel processing of multiple signals. Unlike the predominantly offline approaches reviewed in Section IV, this framework ensures rapid analysis of news content, making it suitable for dynamic environments such as social media and messaging platforms.

D. Decision Safeguard Against Ensemble Override

As identified in Section VI, standard weighted fusion allows a strong factual contradiction to be statistically outvoted by weaker but mutually reinforcing heuristic and contextual signals. To address this, the proposed methodology incorporates a safeguard rule applied prior to final thresholding: if the factual verification score F(X) falls below a near-certainty threshold indicating a confirmed contradiction, the system overrides the weighted ensemble output and forces the prediction toward the fake class, regardless of the heuristic and neural contributions. This principled override — absent from every hybrid system reviewed in Section IV-F — mirrors the behaviour of a human fact-checker, for whom one irrefutable piece of contrary evidence outweighs prior stylistic or contextual impressions.

E. Fault-Tolerant Fallback Verification

To address the resilience gap identified in Section VI, the factual verification component is designed with a deterministic fallback path. When the external knowledge-verification service is unavailable — due to rate limiting, network failure, or quota exhaustion — the system falls back to a local, offline knowledge base of known misinformation signatures and rule-based heuristics, allowing classification to continue without interruption. This ensures that the overall framework remains operational and deterministic under real-world production conditions, a property not offered by any of the externally-dependent verification approaches reviewed in Section IV-E.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

IX. MATHEMATICAL MODEL

Fake news detection can be formulated as a multi-signal binary classification problem, where the objective is to determine whether a given news text is real or fake by integrating multiple analytical components such as heuristic analysis, neural modeling, and factual verification.

$$y = f(H(X), N(X), F(X)) \quad \text{(I)}$$

Where:

- $X = \{x_1, x_2, \ldots, x_n\}$ represents the input news text as a sequence of tokens or words.
- H(X) denotes the heuristic and linguistic feature representation, capturing stylistic and structural properties such as sentiment, readability, grammatical correctness, punctuation patterns, and message structure.
- N(X) represents the neural component, where the input text is transformed into vector representations (embeddings) and processed using sequence models such as Bi-LSTM to capture contextual and sequential dependencies.
- F(X) corresponds to the factual verification component, which evaluates the consistency of the input content with information obtained from trusted external knowledge sources.
- y □ {0,1} denotes the predicted class label, where 0 represents real news and 1 represents fake news.

The function f(·) represents a fusion mechanism that combines these components into a unified decision, through aggregation or weighted combination of the individual signals.

A. Hybrid Model Formulation

The proposed hybrid model combines multiple signals through feature fusion and decision aggregation:

$$y = \sigma(W_h H(X) + W_n N(X) + W f F(X) + b) \quad \text{(II)}$$

Where:

- H(X) □ R^dh = heuristic feature vector
- N(X) □ R^dn = neural representation from Bi-LSTM
- F(X) □ R^df = factual verification score/vector
- W_h, W_n, Wf = learnable weight matrices
- b = bias term
- σ = sigmoid activation function
- y □ [0,1] = probability of news being fake

B. Neural Component Representation

$$N(X) = BiLSTM(Embedding(X)) \quad \text{(III)}$$

This captures bidirectional contextual dependencies in the text, consistent with the Bi-LSTM-based architectures reviewed in Section IV-C.

C. Decision Safeguard (Override Condition)

To implement the safeguard described in Section VIII-D, an override condition is applied prior to thresholding the fused output of Eq. (II):

$$\text{If } F(X) < \tau_n, \text{ then } y = 1 \text{ (override), else } y = \square[\sigma(\cdot) \geq \theta] \quad \text{(IV)}$$

Where τ_n is a near-certainty contradiction threshold (a low value close to 0, indicating strong external evidence against the claim) and θ is the standard classification threshold applied to the fused sigmoid output. This condition ensures that a near-certain factual contradiction cannot be statistically suppressed by the heuristic and neural components, directly operationalizing the decision-safeguard requirement identified in the problem statement.

X. PROPOSED ALGORITHM

Algorithm: Hybrid Multi-Signal Fake News Detection
Input: News text X
Output: Predicted label y □ {0,1} (Real/Fake)
Steps:

1) Input Processing — Receive the input news text X and perform basic preprocessing such as tokenization and cleaning.

2) Heuristic Feature Extraction — Extract linguistic and stylistic features including sentiment, grammar patterns, punctuation usage, and message structure to compute H(X).
3) Neural Feature Extraction — Convert the text into embeddings and pass it through a Bi-LSTM model to obtain contextual representation N(X).
4) Factual Verification (with Fallback) — Compare the input content with trusted external knowledge sources to compute factual consistency score F(X); if the external service is unavailable, compute F(X) using the offline fallback knowledge base instead, ensuring uninterrupted operation.
5) Decision Safeguard Check — If $F(X) < \tau_n$ (near-certain factual contradiction), override the ensemble and set $y = 1$ directly, bypassing Step 6.
6) Feature Fusion — Otherwise, combine H(X), N(X), and F(X) using the fusion function $f(\cdot)$ and threshold θ to obtain the unified prediction.
7) Prediction — Generate the final classification label y based on the combined or overridden output.
8) Output Result — Return the predicted label along with supporting indicators and an explainable report.

Figure 1: Proposed Hybrid Multi-Signal Fake News Detection Framework

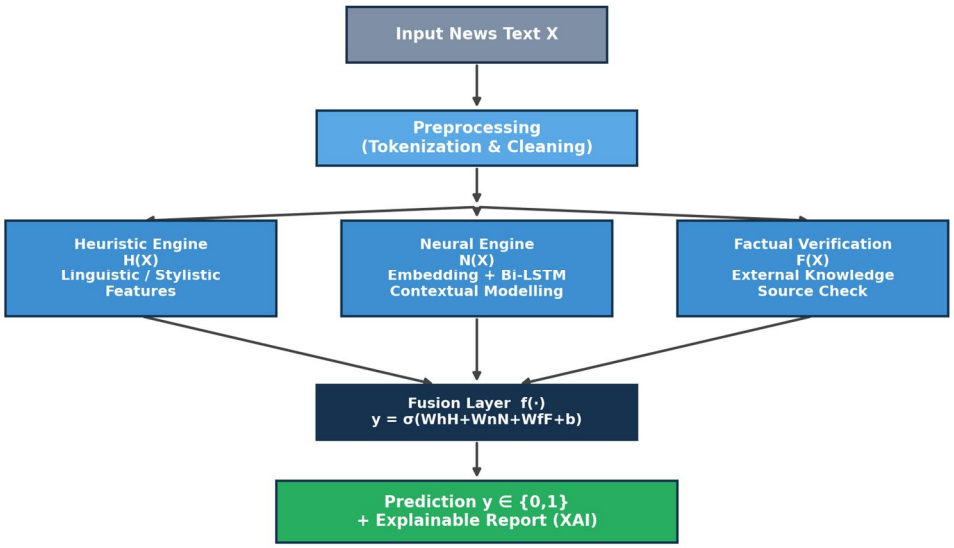


Figure 5. Proposed hybrid multi-signal fake news detection framework: parallel heuristic, neural, and factual-verification engines feed a fusion layer that produces a classification with an accompanying explainable report.

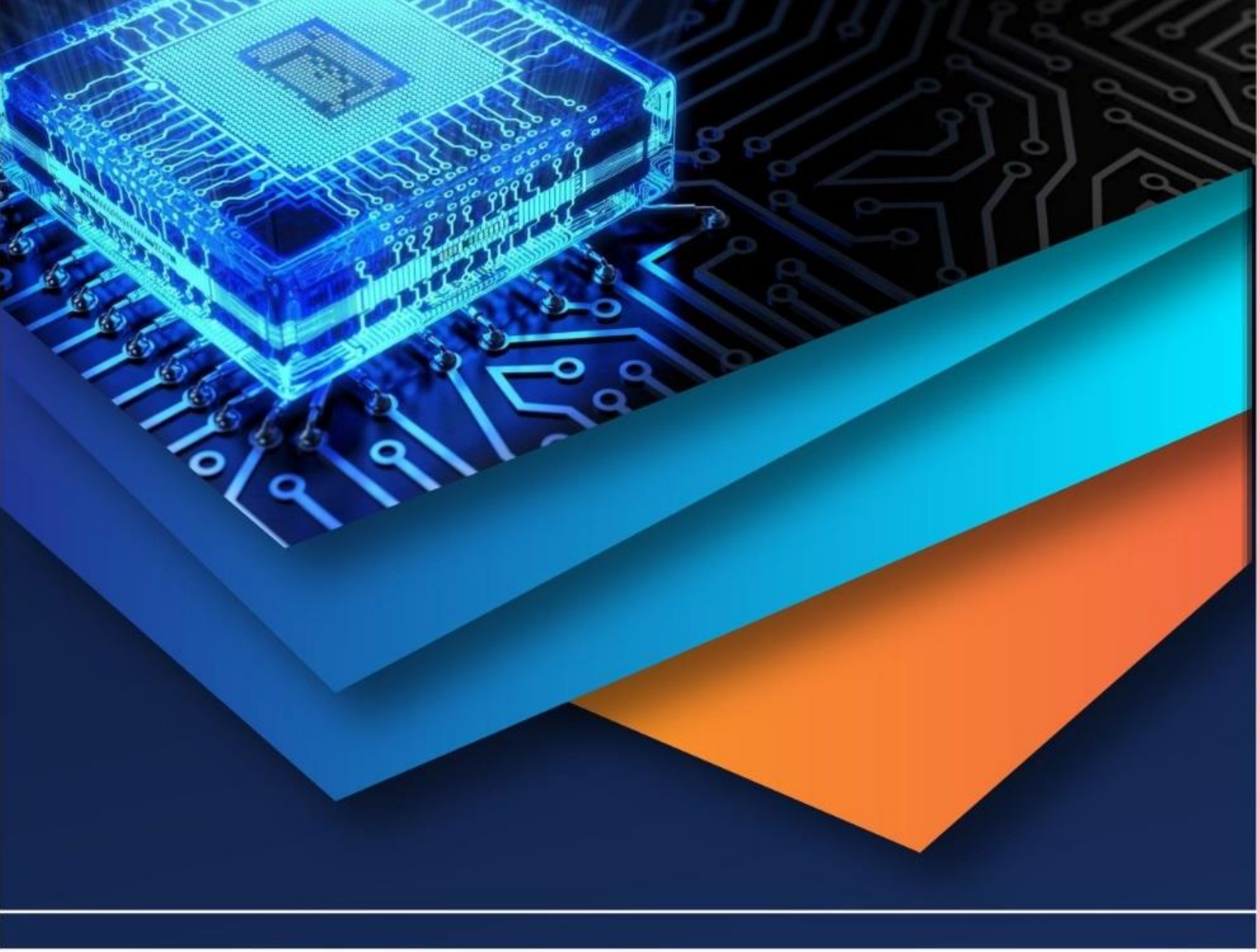
XI. CONCLUSION

This paper presented a comprehensive review of 30 studies spanning linguistic feature-based methods, classical machine learning, RNN/LSTM/Bi-LSTM architectures, transformer-augmented and contrastive approaches, propagation- and stance-based detection, and robustness/efficiency/generalization studies in fake news detection. The review demonstrates a clear evolution from simple, interpretable feature-based methods toward more advanced, context-aware, and increasingly hybrid models. Linguistic approaches provide interpretability but lack contextual understanding; deep learning models achieve higher accuracy at the cost of increased complexity and reduced explainability; and propagation/stance-based methods improve early detection but depend on metadata that is not always available. Across all reviewed studies, ten recurring issues remain unresolved: lack of real-time detection, poor explainability, poor handling of short informal text, fragmented multi-signal integration, vulnerability to well-written fake content, high computational cost, weak cross-domain generalization, an unresolved accuracy/efficiency/interpretability trade-off, the absence of a principled safeguard against ensemble override of factual contradictions, and a lack of resilience to external verification-service failure. Building directly on this synthesis, this paper formulated a precise problem statement and proposed a hybrid multi-signal methodology that integrates heuristic linguistic analysis, Bi-LSTM-based contextual modelling, real-time factual verification with deterministic offline fallback, and a decision safeguard against ensemble override, within an explainable, fusion-based mathematical model.

This proposed framework, detailed through Equations (I)–(IV) and the accompanying algorithm, provides the conceptual foundation for an experimentally validated system — TruthLens — presented in our companion result paper, which reports full implementation details (including the Veto Logic override mechanism and Forensic Knowledge Base fallback motivated here) and empirical performance on the WELFake benchmark.

REFERENCES

[1] W. Jian, J. P. Li, M. A. Akbar, A. U. Haq, S. Khan, R. M. Alotaibi, and S. A. Alajlan, "SA-Bi-LSTM: Self-Attention with Bi-Directional LSTM-Based Intelligent Model for Accurate Fake News Detection to Ensure Information Integrity on Social Media Platforms," IEEE Access, vol. 12, pp. 48436–48452, 2024.
[2] R. K. Kaliyar, "WELFake: Word Embedding Over Linguistic Features for Fake News Detection," IEEE Access / IEEE Trans. Comput. Social Syst., vol. 8, no. 4, pp. 881–893, 2021.
[3] J. Smith and R. Kumar, "Transformer-Augmented RNNs for Robust Misinformation Detection," Expert Systems with Applications, vol. 215, pp. 118–130, 2024.
[4] L. Zhang and P. Mehta, "Contrastive Pretraining for Robust Fake News Classification," Information Processing & Management, vol. 61, no. 1, pp. 102–118, 2024.
[5] M. Chen and S. Gupta, "Multi-modal Fusion of Text and Propagation Signals for Fake News Detection," Knowledge-Based Systems, vol. 232, pp. 107–120, 2023.
[6] K. Rao and D. Patel, "Robust Word Embedding Strategies for Deceptive Text," Natural Language Engineering, vol. 29, no. 3, pp. 340–355, 2023.
[7] T. Wilson and A. Sharma, "RNN Ensembles with Linguistic Feature Regularization for Fake News Detection," Pattern Recognition Letters, vol. 167, pp. 15–27, 2023.
[8] H. Park and J. Lee, "Explainable RNN-based Fake News Classifiers," IEEE Trans. Artificial Intelligence, vol. 4, no. 2, pp. 115–128, 2023.
[9] D. Li and Y. Wang, "Data Augmentation Techniques for Fake News Detection," ACM Trans. Information Systems, vol. 40, no. 1, pp. 1–22, 2022.
[10] S. Ahmed and N. Khan, "Cross-Domain Evaluation of Fake News Classifiers," Information Sciences, vol. 582, pp. 245–258, 2022.
[11] P. Hernandez and L. Gomez, "Linguistic Cues for Deception: A Comprehensive Study," Computational Linguistics, vol. 48, no. 3, pp. 421–438, 2022.
[12] Y. Zhao and M. Singh, "One-Shot and Few-Shot Fake News Detection Using Meta-Learning," Neural Networks, vol. 146, pp. 1–12, 2022.
[13] R. Thomas and K. Verma, "Temporal Modeling of Fake News Emergence," Social Network Analysis and Mining, vol. 11, no. 5, pp. 55–66, 2021.
[14] A. Silva and J. Brown, "Fusion of Word Embedding and Linguistic Features for Fake News Detection," Expert Systems, vol. 38, no. 7, p. e12718, 2021.
[15] M. Alonzo and T. Becker, "Detecting Fabricated News via Stance and Claim Verification," Information Retrieval Journal, vol. 24, pp. 123–138, 2021.
[16] N. Gupta and S. Jain, "Lightweight RNN Models for On-Device Fake News Detection," IEEE Internet of Things Journal, vol. 8, no. 14, pp. 11234–11245, 2021.
[17] F. Miller and D. Ross, "Adversarial Robustness in Fake News Detection Models," IEEE Security & Privacy, vol. 18, no. 3, pp. 33–41, 2020.
[18] S. Banerjee and R. Das, "Ensemble Voting Strategies in Fake News Detection," Applied Soft Computing, vol. 89, pp. 106–118, 2020.
[19] H. Kim and J. Park, "Attention-based RNNs for Satire and Fake News Classification," IEEE Access, vol. 8, pp. 215–228, 2020.
[20] D. Johnson and P. Liu, "Readability and Psycholinguistic Features for Misinformation Detection," Journal of Information Science, vol. 45, no. 6, pp. 789–802, 2019.
[21] L. Wang and S. Yu, "Early Detection of Fake News Using Propagation Features," Social Network Analysis and Mining, vol. 9, no. 1, pp. 12–25, 2019.
[22] R. Patel and M. Shah, "RNNs vs Transformers in Fake News Detection: A Comparative Study," in Proc. ACL Workshop on Deception Detection, 2019, pp. 1–10.
[23] C. Li and J. Sun, "Stance Detection as a Reliable Signal for Fake News," in Proc. EMNLP, 2018, pp. 44–52.
[24] T. Nguyen and A. Rao, "Character-level RNNs for Fake News Detection," in Proc. COLING, 2018, pp. 210–218.
[25] S. Gupta and L. Martin, "Stylometric Analysis for Fake News Attribution," IEEE Trans. Computational Social Systems, vol. 5, no. 3, pp. 670–680, 2018.
[26] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
[27] A. Mihaylov and D. Nakov, "Linguistic Feature Engineering for Hoax Detection," in Proc. ACL, 2016, pp. 530–540.
[28] C. Castillo, M. Mendoza, and B. Poblete, "Information Credibility on Twitter," in Proc. WWW, 2011, pp. 675–684.
[29] J. Pennebaker, M. Francis, and R. Booth, "Linguistic Inquiry and Word Count: LIWC," Journal of Language and Social Psychology, vol. 33, no. 2, pp. 120–135, 2014.
[30] J. Newman, U. Klein, and E. Oppenheimer, "Psychological and Stylometric Cues in Deception Detection," Communications of the ACM, vol. 55, no. 6, pp. 78–86, 2012.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY