**Interdisciplinary Journal of AI, Machine Learning & Data Science (IJAIMLDS)**

**ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/xqp24h93**

## Article - e012

# A Comprehensive Literature Review on Fake News Detection: From Traditional Machine Learning to Deep Learning and Transformer-Based Approaches

**Gourav Yadav**[1], **Dr. Piyush Moghe**[2]

[1,2]*Department of IAC (Specialization), Institute of Advance Computing*
*SAGE University, Indore, Madhya Pradesh, India*

**Received**: 20/04/2026 **Revision Received**: 12/05/2026 **Accepted**: 28/05/2026

**ABSTRACT**

The unprecedented proliferation of digital media and social networking platforms has amplified the spread of fake news, posing serious threats to social stability, public trust, democratic integrity, and public health. This paper presents a comprehensive literature review of fake news detection methodologies, tracing the evolution from early rule-based and traditional machine learning approaches to modern deep learning architectures and transformer-based pre-trained language models. We systematically examine foundational frameworks for defining and categorizing misinformation, the application of feature engineering techniques including Bag-of-Words and TF-IDF representations, and the progression through classical classifiers such as Naive Bayes, Support Vector Machines, and Random Forests. We then review deep learning advances including Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, Bidirectional LSTMs with attention mechanisms, and state-of-the-art transformer models such as BERT, RoBERTa, and ALBERT. The review further examines word embedding techniques including Word2Vec and GloVe, benchmark datasets including LIAR, ISOT, FakeNewsNet, and the Kaggle Fake News Dataset, and standard evaluation frameworks. Research gaps including interpretability, multilingual detection, adversarial robustness, and real-time deployment are synthesized and discussed. The review concludes with a comparative analysis of model performance across approaches and a structured agenda for future research.

**KEYWORDS**: Fake News Detection, Natural Language Processing, Deep Learning, LSTM, BERT, Misinformation, Text Classification, Word Embeddings

**© 2026 The Author(s) | Published by - Scholiora Research Publications**
**Creative Commons Attribution 4.0 International License (CC BY 4.0)**

Interdisciplinary Journal of AI, Machine Learning & Data Science (IJAIMLDS)
ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/xqp24h93

## 1. INTRODUCTION

The rapid advancement of digital media and social networking platforms has fundamentally transformed how information is created, consumed, and disseminated worldwide. While this transformation has democratized access to information, it has simultaneously enabled an unprecedented proliferation of fake news — intentionally fabricated, manipulated, or misleading content presented as legitimate reporting — with documented and severe consequences for public health, electoral integrity, financial stability, and social cohesion. [1,2]

The dangers of fake news are well documented. Research has established that false information spreads significantly faster and farther than truthful content across digital platforms, [4] with direct consequences observed in political elections, public health crises such as COVID-19, and episodes of communal violence. [5,6] Traditional manual fact-checking mechanisms — relying on human editorial oversight and journalistic standards — are wholly insufficient to address the scale, velocity, and sophistication of modern misinformation. [1,2]

Automated fake news detection, grounded in Natural Language Processing (NLP) and Machine Learning (ML), has emerged as a critical research area in response to this challenge. [2] Over the past decade, the field has evolved dramatically — from simple rule-based keyword matching systems through statistical feature engineering and classical machine learning classifiers, to sophisticated deep learning architectures and large pre-trained transformer models capable of capturing nuanced semantic and contextual relationships in text. [2,17]

This paper presents a comprehensive and structured literature review of fake news detection methodologies. The review is organized thematically: beginning with foundational definitions and taxonomies, progressing through traditional machine learning approaches, examining deep learning and attention-based architectures, evaluating transformer-based pre-trained language models, reviewing benchmark datasets and evaluation frameworks, and concluding with a synthesis of research gaps and a structured agenda for future investigation.

### *1.1 Review Methodology*

This review was conducted following a structured literature search across multiple academic databases, including IEEE Xplore, ACM Digital Library, Google Scholar, Semantic Scholar, and arXiv. Search terms included combinations of “fake news detection,” “misinformation detection,” “disinformation,” “natural language processing,” “deep learning,” “transformer,” “BERT,” “LSTM,” and “text classification.” The primary publication timeframe considered spans 2011 to 2026, with emphasis on high-impact works published from 2017 onward. Inclusion criteria required that papers address automated computational detection of fake news or misinformation, present empirical evaluations on benchmark datasets, and be published in peer-reviewed venues or

© 2026 The Author(s) | Published by - Scholiora Research Publications
Creative Commons Attribution 4.0 International License (CC BY 4.0)

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/xqp24h93

as widely cited preprints. Papers addressing only manual fact-checking, purely sociological analyses without computational methods, or non-English detection without cross-lingual methodology were excluded. In total, over 120 candidate papers were identified, from which 40 core references were selected based on citation impact, methodological significance, and direct relevance to the scope of this review.

The primary contributions of this review are: (1) a systematic and chronologically organized survey of fake news detection methodologies; (2) a comparative analysis of model performance across approaches and datasets; (3) a structured synthesis of open research challenges; and (4) a forward-looking research agenda addressing key gaps in the literature.Unlike prior surveys (e.g., Zhou & Zafarani, 2020 [2]; Shu et al., 2017 [1]), this review uniquely integrates coverage of recent developments in Large Language Models (LLMs), multimodal detection, explainable AI, and lightweight transformer architectures (2023–2026) within a single unified comparative framework. It further provides a structured discussion of deployment constraints — including scalability, inference latency, and real-time implementation — that are underrepresented in earlier reviews.

**Scope Alignment**: As AI-driven systems increasingly underpin IoT ecosystems and industrial automation environments, ensuring the integrity and reliability of information inputs to these systems becomes critical. Automated fake news detection represents a foundational challenge in AI-enabled information systems.

## 2. DEFINITIONS AND TAXONOMY OF FAKE NEWS

Before reviewing detection methodologies, it is essential to establish a precise conceptual understanding of what constitutes fake news. The term encompasses a broad and nuanced spectrum of misinformation types, each presenting distinct computational detection challenges. [1,3]

### *2.1 Foundational Definitions*

Wardle and Derakhshan (2017) proposed one of the most widely cited taxonomic frameworks for categorizing problematic information, [3] distinguishing between three primary types: (i) Misinformation — false information shared without deliberate intent to deceive; (ii) Disinformation — deliberately fabricated or manipulated content shared with explicit harmful intent; and (iii) Malinformation — truthful information shared with intent to cause harm, such as unlawful exposure of private data.

Within the computational research domain, Shu et al. (2017) defined fake news as news articles that are intentionally and verifiably false and could mislead readers. [1] Zhou and Zafarani (2020) provided the most comprehensive computational survey of fake news research, [2] identifying four

© 2026 The Author(s) | Published by - Scholiora Research Publications
Creative Commons Attribution 4.0 International License (CC BY 4.0)

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/xqp24h93

primary research categories: detection, intervention, characterization, and propagation — with detection representing the most active research domain.

### *2.2 Categories of Fake News*

Fake news manifests across several distinct categories that present varying degrees of computational detectability:

- Fabricated content: Entirely invented stories with no factual basis, designed to deceive
- Manipulated content: Genuine information altered through selective omission, decontextualization, or distortion
- Misleading headlines: Accurate body content paired with exaggerated or misrepresentative headlines
- Satire misrepresented as fact: Comedic or satirical content shared or interpreted without satirical context
- Propaganda: Ideologically motivated content designed to promote a particular viewpoint through selective presentation

Each category exhibits distinct linguistic patterns and requires different detection strategies. Fabricated content and manipulated content — the primary targets of computational detection systems — often mimic legitimate journalistic writing styles at the surface level, requiring deep semantic understanding for reliable identification. [1,8]

### *2.3 Social Dynamics of Fake News Spread*

Understanding why fake news spreads is as important as detecting it. Vosoughi, Roy, and Aral (2018) conducted a landmark study of 126,000 news stories shared on Twitter over a decade, conclusively demonstrating that false news spreads faster, more broadly, and more deeply than true news. [4] They attributed this largely to the novelty and emotional provocation of false content rather than to bot activity. Lazer et al. (2018) further documented the role of algorithmic amplification, confirmation bias, and echo chambers in accelerating fake news propagation. [5] Pennycook and Rand (2019) demonstrated that even brief interventions promoting analytical thinking can reduce susceptibility to fake news. [37]

## 3. TRADITIONAL MACHINE LEARNING APPROACHES

Early computational approaches to fake news detection relied on handcrafted feature engineering combined with classical supervised machine learning classifiers. While these methods established foundational benchmarks, they are fundamentally limited by their inability to capture semantic meaning, word order, and long-range contextual dependencies in natural language.

© 2026 The Author(s) | Published by - Scholiora Research Publications
Creative Commons Attribution 4.0 International License (CC BY 4.0)

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/xqp24h93

### *3.1 Feature Engineering*

#### *3.1.1 Bag-of-Words and TF-IDF*

The Bag-of-Words (BoW) model represents a document as an unordered collection of word frequencies, discarding all grammatical structure and word order. [10] Term Frequency-Inverse Document Frequency (TF-IDF) improves upon BoW by weighting word frequencies according to their informativeness across the corpus, reducing the influence of common but uninformative words. [10] While computationally efficient, both representations fundamentally fail to capture the semantic meaning of words or the contextual relationships between them — a critical limitation for fake news detection, where meaning is often conveyed through subtle contextual cues rather than individual word presence. [2]

#### *3.1.2 N-gram Features*

N-gram models capture sequences of N consecutive words, providing partial compensation for the word order blindness of BoW representations. Bigrams and trigrams proved useful for capturing short phrasal patterns characteristic of fake news, such as inflammatory compound expressions and sensationalist phrasing. However, N-gram feature spaces grow exponentially with vocabulary size, creating computational and sparsity challenges.

#### *3.1.3 Linguistic and Stylometric Features*

Pérez-Rosas et al. (2018) conducted a systematic study of linguistic features for fake news detection, [8] identifying that surface-level cues including readability scores, sentiment polarity, frequency of superlatives, pronoun usage, and part-of-speech distributions provide meaningful classification signals. Castillo et al. (2011) extended feature engineering to social signals on Twitter, [7] demonstrating that user credibility, retweet patterns, and content features could collectively support credibility assessment with promising accuracy.

### *3.2 Classical Classifiers*

#### *3.2.1 Naive Bayes*

The Multinomial Naive Bayes classifier, grounded in Bayes' theorem with the assumption of conditional feature independence, was among the earliest classifiers applied to fake news detection. Its computational efficiency and strong baseline performance on text data made it a widely used starting point. However, the fundamental violation of the feature independence assumption in natural language — where words co-occur in structured patterns — limits its ability to capture the complex linguistic dependencies characteristic of fake news.

#### *3.2.2 Support Vector Machines*

Support Vector Machines (SVM), which identify the optimal hyperplane maximally separating classes in a high-dimensional feature space, [11] emerged as the most effective traditional classifier

© 2026 The Author(s) | Published by - Scholiora Research Publications
Creative Commons Attribution 4.0 International License (CC BY 4.0)

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/xqp24h93

for fake news detection across multiple benchmark studies. [9] Ahmed et al. (2018) demonstrated that a Linear SVM combined with TF-IDF bigram features achieved over 92% accuracy on the ISOT Fake News Dataset, [9] establishing a strong traditional ML benchmark. Castillo et al. (2011) similarly reported strong SVM performance on Twitter credibility classification using handcrafted features. [7]

#### *3.2.3 Logistic Regression*

Logistic Regression, despite its relative simplicity, proved to be a competitive baseline for fake news classification. Its probabilistic output, interpretability, and efficient training made it a frequently used benchmark in comparative studies. Several research groups reported Logistic Regression with TF-IDF features achieving surprisingly competitive accuracy on well-curated datasets, though performance degraded significantly on more diverse or challenging datasets.

#### *3.2.4 Random Forest and Ensemble Methods*

Random Forest classifiers, which aggregate predictions of multiple decision trees trained on random feature subsets, [12] demonstrated improved robustness over single-model approaches through variance reduction. Ensemble methods including Gradient Boosting and AdaBoost showed promise in handling high-dimensional, noisy feature spaces typical of fake news datasets, though at the cost of reduced interpretability and increased computational overhead compared to linear methods.

### *3.3 Limitations of Traditional Approaches*

Despite establishing important baselines, traditional machine learning approaches share several fundamental limitations that motivate the transition to deep learning:

- Semantic blindness: BoW and TF-IDF representations capture word occurrence but not semantic meaning or contextual relationships
- Sequential structure loss: All traditional classifiers treat documents as unordered feature collections, losing grammatical structure and word order entirely
- Manual feature engineering: Handcrafted features are domain-specific, time-consuming, and difficult to generalize across news topics or writing styles
- Limited generalization: Models trained on specific datasets frequently fail to generalize to new domains, sources, or time periods
- Scalability challenges: Feature engineering pipelines are difficult to scale to the volume of content generated on modern digital platforms

| Approach | Best Accuracy | Dataset | Key Limitation |
|---|---|---|---|
| Naive Bayes | 84–87% | ISOT / Kaggle | Feature independence assumption violated in NL |
| Logistic Regression | 89–92% | ISOT / Kaggle | Linear boundary; no sequential understanding |

© 2026 The Author(s) | Published by - Scholiora Research Publications
Creative Commons Attribution 4.0 International License (CC BY 4.0)

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/xqp24h93

| Approach | Best Accuracy | Dataset | Key Limitation |
|---|---|---|---|
| SVM (Linear) [11] | 92–94% | ISOT [9] | No semantic understanding; bag-of-words only |
| Random Forest [12] | 88–91% | Multiple | High variance; domain-specific feature engineering |
| Ensemble Methods | 90–93% | Multiple | Computational cost; reduced interpretability |

*Table 1. Performance summary of traditional machine learning approaches for fake news detection*

### *3.4 Critical Assessment of Traditional Approaches.*

A comparative evaluation of the approaches summarised in Table 1 reveals important trade-offs. SVM with TF-IDF achieves the strongest performance among traditional classifiers (~92–94%), primarily because the kernel trick allows it to find non-linear decision boundaries in high-dimensional sparse feature spaces. However, it is brittle to domain shift: a model trained on ISOT (Reuters vs. unreliable web sources) transfers poorly to LIAR (short political statements), because the feature distributions differ substantially. Logistic Regression offers competitive accuracy at lower computational cost and greater interpretability, making it a practical choice when model explainability is required; however, its linear decision boundary cannot capture interaction effects between features. Naive Bayes, though fast and interpretable, is particularly susceptible to class-conditional feature dependencies — a fundamental property of natural language — and its accuracy degrades notably on datasets with longer, more complex articles. Random Forest addresses some of these weaknesses through ensemble averaging but introduces hyperparameter sensitivity and opacity. Crucially, all traditional approaches share the limitation that they cannot distinguish between a genuine article and a stylistically similar fabricated one that uses the same vocabulary, as they lack any mechanism for grounding claims against factual knowledge. This motivates the transition to representation-learning approaches reviewed in the following sections.

## 4. DEEP LEARNING APPROACHES

The introduction of deep learning to fake news detection represented a paradigm shift, enabling models to automatically learn complex hierarchical feature representations from large text corpora without manual feature engineering. [17,18] Deep learning models demonstrated substantial and consistent performance improvements over traditional classifiers, particularly on large and diverse

© 2026 The Author(s) | Published by - Scholiora Research Publications
Creative Commons Attribution 4.0 International License (CC BY 4.0)

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/xqp24h93

datasets, driven by their ability to capture sequential structure, semantic meaning, and long-range contextual dependencies in text. [2]

### *4.1 Word Embedding Foundations*

A critical prerequisite for deep learning NLP models is the transformation of discrete word tokens into dense, continuous vector representations that encode semantic relationships. Mikolov et al. (2013) introduced Word2Vec, [29,30] demonstrating that neural word embeddings trained on large corpora encode rich semantic and syntactic relationships — captured in the famous king − man + woman ≈ queen analogy. Pennington, Socher, and Manning (2014) introduced GloVe (Global Vectors for Word Representation), [28] which constructs embeddings by factorizing a global word co-occurrence matrix, capturing both local context and global corpus statistics. Bojanowski et al. (2017) further extended embeddings with subword information through FastText, [40] improving handling of rare and morphologically complex words.

These pre-trained embeddings provided a crucial head start for deep learning models by transferring semantic knowledge learned from billions of words to task-specific classification models, particularly beneficial when task-specific training data is limited.

### *4.2 Convolutional Neural Networks*

Convolutional Neural Networks (CNNs), originally developed for image recognition, were successfully adapted for text classification by Kim (2014) in a seminal paper demonstrating that CNNs with multiple convolutional filter sizes effectively capture local n-gram patterns across word embedding sequences. [16] Applied to fake news detection, CNNs extract locally significant textual features — capturing short phrases, idiomatic expressions, and stylistic patterns characteristic of fabricated content.

CNNs offer important practical advantages: they are computationally efficient, highly parallelizable, and require relatively modest training resources compared to recurrent architectures. However, their local receptive field is a fundamental architectural limitation — CNNs cannot inherently model dependencies spanning the full length of a news article, which may span hundreds of sentences.

### *4.3 Recurrent Neural Networks and LSTM*

#### *4.3.1 Standard RNNs and the Vanishing Gradient Problem*

Recurrent Neural Networks (RNNs) process text sequentially, maintaining a hidden state updated at each token to accumulate contextual information from preceding tokens. [17] While theoretically capable of modeling arbitrary-length dependencies, standard RNNs suffer from the vanishing gradient problem — during backpropagation through long sequences, gradient signals diminish exponentially, preventing the model from learning dependencies separated by more than a few tokens. [15]

© 2026 The Author(s) | Published by - Scholiora Research Publications
Creative Commons Attribution 4.0 International License (CC BY 4.0)

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/xqp24h93

#### *4.3.2 Long Short-Term Memory Networks*

Long Short-Term Memory (LSTM) networks, introduced by Hochreiter and Schmidhuber (1997), addressed the vanishing gradient problem through the introduction of a gating mechanism comprising input, forget, and output gates. [15] These gates selectively retain or discard information across arbitrarily long sequences, enabling LSTMs to capture long-range dependencies that are entirely inaccessible to standard RNNs or CNNs.

In the context of fake news detection, LSTM models demonstrated consistently strong performance by capturing long-range linguistic and rhetorical patterns across full article lengths. [15,14] Rashkin et al. (2017) demonstrated that LSTM models significantly outperformed traditional baselines on the LIAR benchmark, [14] with the model learning to associate specific linguistic patterns — including hedging language, superlative usage, and emotional tone — with different levels of statement truthfulness. [13]

#### *4.3.3 Bidirectional LSTM*

Bidirectional Long Short-Term Memory (BiLSTM) networks process text in both forward and backward temporal directions simultaneously, [19] maintaining two separate hidden state streams that are concatenated at each time step. This bidirectional processing enables the model to consider both preceding and following context when encoding each token — a significant advantage for understanding word meaning within complex sentence and paragraph structures.

Karimi and Tang (2019) employed a BiLSTM architecture with hierarchical attention mechanisms for fake news detection, [20] achieving state-of-the-art performance on multiple benchmark datasets. Their model demonstrated that attention-weighted bidirectional encoding could identify not only whether an article is fake but which specific segments of text are most suspicious — providing a foundation for interpretable detection.

### *4.4 Attention Mechanisms*

Yang et al. (2016) introduced the Hierarchical Attention Network (HAN), [22] applying attention mechanisms at both the word level (within each sentence) and the sentence level (across the full document). This hierarchical structure enables the model to dynamically weight the importance of individual words and entire sentences to the classification decision, rather than treating all content equally.

Applied to fake news detection, attention mechanisms provide two critical benefits: (1) performance improvement through dynamic, content-aware weighting of discriminative text regions; and (2) a degree of interpretability — attention weights can be visualized to identify which parts of an article the model found most suspicious, providing explainable outputs unavailable from standard LSTM or CNN classifiers.

© 2026 The Author(s) | Published by - Scholiora Research Publications
Creative Commons Attribution 4.0 International License (CC BY 4.0)

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/xqp24h93

### *4.5 Hybrid CNN-LSTM Architectures*

Recognizing that CNNs excel at capturing local n-gram patterns while LSTMs excel at modelling long-range sequential dependencies, several researchers proposed hybrid architectures combining both model types. In a typical CNN-LSTM pipeline, convolutional layers first extract local feature maps from the embedded input text, which are then processed by LSTM layers for sequential modelling. These hybrid models reported competitive performance by leveraging the complementary strengths of both architectures, achieving better results than either approach in isolation on several benchmark datasets.

### *4.6 Regularization Strategies for Deep Learning*

Overfitting is a significant challenge when training deep learning models on finite fake news datasets. Srivastava et al. (2014) introduced Dropout, [21] which randomly deactivates a proportion of neurons during training, preventing co-adaptation and forcing the model to learn more distributed and robust representations. Recurrent Dropout — applied specifically to the internal connections within LSTM layers — proved particularly effective for regularizing sequential architectures, reducing the tendency of LSTMs to memorize training sequence patterns rather than generalizing.

| Model Architecture | Accuracy Range | Dataset(s) | Key Advantage |
|---|---|---|---|
| CNN [16] | 87–93% | LIAR, ISOT | Efficient; captures local n-gram patterns |
| LSTM [15] | 91–96% | LIAR, ISOT, Kaggle | Long-range sequential dependencies |
| BiLSTM [19] | 93–97% | Multiple | Bidirectional context understanding |
| BiLSTM + Attention [20, 22] | 94–97% | Multiple | Improved accuracy + interpretability |
| Hybrid CNN-LSTM | 92–96% | Multiple | Combines local and sequential features |

*Table 2. Performance summary of deep learning approaches for fake news detection*

### *4.7 Critical Assessment of Deep Learning Approaches.*

The results in Table 2 reflect a clear progression in representational power. CNNs, while computationally efficient, are structurally limited to detecting patterns within a local window, making them less effective at capturing long-range rhetorical inconsistencies that span entire

© 2026 The Author(s) | Published by - Scholiora Research Publications
Creative Commons Attribution 4.0 International License (CC BY 4.0)

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/xqp24h93

articles. LSTMs overcome this limitation through gated memory but face practical constraints: training is sequential and does not parallelize across timesteps, limiting scalability to very long documents. BiLSTMs consistently outperform unidirectional LSTMs (+1–3%) owing to richer contextual encoding, but at the cost of doubled hidden state computation. The addition of hierarchical attention (Yang et al., 2016 [22]) provides further gains and introduces interpretability through attention weight visualization; however, attention weights have been shown to be poor proxies for feature importance in some configurations (Jain & Wallace, 2019), limiting their reliability as explanations. Hybrid CNN-LSTM models exhibit strong performance in practice but complicate hyperparameter tuning and increase training time. A practical limitation of all architectures in this section is their sensitivity to dataset-specific vocabulary and style: models trained on ISOT (formal news article text) frequently underperform on the LIAR dataset (short informal political statements), underscoring the need for domain-robust training protocols. Recent work by Nguyen et al. (2023) [41] has demonstrated that combining adversarial data augmentation with BiLSTM training can substantially improve cross-domain generalization.

## 5. TRANSFORMER-BASED PRE-TRAINED LANGUAGE MODELS

The introduction of the Transformer architecture by Vaswani et al. (2017) [23] and the subsequent development of large pre-trained language models represented the most significant advance in NLP in recent history, achieving state-of-the-art performance across virtually every language understanding benchmark and fundamentally changing the paradigm of NLP system development from feature engineering to fine-tuning.

### *5.1 The Transformer Architecture*

The Transformer architecture replaces sequential processing with a self-attention mechanism that allows every token in a sequence to directly attend to every other token simultaneously, [23] capturing global contextual relationships without the sequential bottleneck of RNNs. Multi-head attention enables the model to attend to multiple aspects of the input simultaneously — capturing syntactic, semantic, and pragmatic relationships in parallel. Positional encodings preserve token order information without requiring sequential processing.

### *5.2 BERT — Bidirectional Encoder Representations from Transformers*

Devlin et al. (2019) introduced BERT, [24] a landmark model pre-trained on massive text corpora (BooksCorpus and English Wikipedia — approximately 3.3 billion words) using two novel self-supervised objectives: (1) Masked Language Modelling, which predicts randomly masked tokens using bidirectional context; and (2) Next Sentence Prediction, which determines whether two sentences are semantically consecutive. BERT develops deep, bidirectional, contextual word

© 2026 The Author(s) | Published by - Scholiora Research Publications
Creative Commons Attribution 4.0 International License (CC BY 4.0)

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/xqp24h93

representations that encode rich semantic and syntactic knowledge — far superior to static word embeddings such as Word2Vec or GloVe.

When fine-tuned on fake news detection datasets, BERT-based models consistently achieved the highest reported accuracies in the literature. [24] Kula et al. (2021) demonstrated that fine-tuned BERT models achieved over 98% accuracy on the ISOT Fake News Dataset, [27] substantially outperforming all previous approaches and establishing the current state of the art.

### *5.3 BERT Variants and Extensions*

Following BERT's success, numerous variants have been developed to improve performance, efficiency, or domain specificity. RoBERTa (Liu et al., 2019) [25] removes the Next Sentence Prediction objective and uses larger batch sizes and more training data, achieving superior performance to the original BERT on multiple benchmarks. ALBERT (Lan et al., 2020) [26] introduces parameter-sharing and factorized embedding parameterization to reduce model size by up to 18× while maintaining competitive performance — addressing the computational cost barrier of transformer deployment. Each of these variants has demonstrated strong performance on fake news classification tasks.

### *5.4 Computational Cost Considerations*

Despite their impressive performance, transformer-based models carry significant practical limitations that constrain their applicability in resource-constrained settings:

- BERT-base contains 110 million parameters; BERT-large contains 340 million — compared to approximately 5 million for a typical LSTM model
- Training and fine-tuning require GPU clusters that are unavailable to many researchers, particularly in developing nations
- Inference latency may preclude real-time detection in high-throughput environments
- The black-box nature of transformer attention raises interpretability concerns in high-stakes deployment contexts

These constraints motivate continued research into computationally accessible alternatives — including optimized LSTM architectures, knowledge distillation, and model compression — that can achieve competitive performance within realistic resource constraints.

| Model | Accuracy | Dataset | Key Characteristic |
|---|---|---|---|
| BERT-base [24] | 97–98% | ISOT, Kaggle | 110M params; bidirectional; state-of-the-art |
| RoBERTa [25] | 97–99% | Multiple | Improved BERT pre-training; no NSP |
| ALBERT [26] | 96–98% | Multiple | 18× smaller; parameter sharing |
| Kula et al. BERT | 98%+ | ISOT | Fine-tuned; best reported on ISOT |

© 2026 The Author(s) | Published by - Scholiora Research Publications
Creative Commons Attribution 4.0 International License (CC BY 4.0)

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/xqp24h93

| Model | Accuracy | Dataset | Key Characteristic |
|---|---|---|---|
| [27] | | | |
| DistilBERT | 95–97% | Multiple | 40% smaller; 60% faster inference |

*Table 3. Performance summary of transformer-based approaches for fake news detection*

***5.5 Large Language Models (LLMs) in Fake News Detection.***

The emergence of generative Large Language Models (LLMs) — most notably GPT-4 (OpenAI, 2023) [42] and LLaMA-2/3 (Touvron et al., 2023) [43] — has introduced new paradigms for fake news detection that go beyond discriminative fine-tuning. Prompting-based approaches leverage LLMs as zero-shot or few-shot classifiers, circumventing the need for large labeled training datasets. Pelrine et al. (2023) demonstrated that GPT-4 with carefully crafted chain-of-thought prompts achieves competitive performance on LIAR-comparable benchmarks, correctly classifying over 70% of multi-class claims without any task-specific fine-tuning. However, LLMs introduce new risks: they can themselves generate highly convincing fake news (the "dual-use" problem), and their opaque reasoning chains may reflect biases present in pre-training corpora. Furthermore, the computational cost of deploying 70B+ parameter models in production classification pipelines remains prohibitive. Hybrid approaches — using LLMs for evidence retrieval and summarisation paired with lightweight classifiers for final prediction — are an emerging and promising direction. The tension between LLM capability and deployment feasibility represents one of the most important open questions in the current literature.

***5.6 Explainable AI and Lightweight Transformer Architectures.***

Explainability has emerged as a first-class requirement for fake news detection systems deployed in high-stakes contexts. Beyond attention visualization, recent work has integrated gradient-based attribution methods (Integrated Gradients, Sundararajan et al.) and SHAP values [36] directly into transformer-based detection pipelines, providing token-level importance scores that help human fact-checkers identify which claims within an article triggered the classification. Kotonya and Toni (2020) proposed an explainable automated fact-checking framework using BERT with structured explanations, demonstrating that explanations improve user trust and model audibility simultaneously. In parallel, the practical deployment barrier posed by large transformer models has motivated substantial research into lightweight architectures. MobileBERT (Sun et al., 2020) achieves 4.3× faster inference than BERT-base with only 0.6% accuracy loss. TinyBERT (Jiao et al., 2020) employs task-specific knowledge distillation to produce a 7.5× smaller model retaining 96.8% of BERT's performance. More recently, Rao et al. (2024) [45] demonstrated that a compressed 4-layer transformer fine-tuned with curriculum learning achieves 95.3% accuracy on

© 2026 The Author(s) | Published by - Scholiora Research Publications
Creative Commons Attribution 4.0 International License (CC BY 4.0)

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/xqp24h93

ISOT — within 2% of full BERT-base — at 6× lower inference latency, making it viable for real-time deployment on standard server hardware.

## 6. BENCHMARK DATASETS

The availability of high-quality labeled datasets has been fundamental to the progress of fake news detection research. The following datasets have been most widely used for training and evaluating computational detection systems.

### *6.1 LIAR Dataset*

Introduced by Wang (2017), [13] the LIAR dataset contains approximately 12,800 human-labeled short statements sourced from PolitiFact, annotated with six truthfulness labels ranging from pants-fire (completely false) to true. Its fine-grained, multi-class labeling makes it particularly challenging for automated classification — most models report accuracies in the 25–30% range on the six-class task — and it has become a standard benchmark for evaluating nuanced fake news detection systems. The dataset also includes metadata such as speaker, subject, context, and state, enabling multi-modal feature integration.

### *6.2 ISOT Fake News Dataset*

The ISOT dataset, developed by Ahmed et al. (2018) at the University of Victoria, [9] contains over 44,000 news articles collected from a diverse range of sources. Fake articles were sourced from websites flagged by PolitiFact as unreliable, while real articles were sourced from Reuters. The binary labeling, large size, and relative class balance (approximately equal proportions of fake and real) make it one of the most suitable datasets for training deep learning classifiers, and it is frequently used as the primary benchmark in comparative studies.

### *6.3 FakeNewsNet*

Shu et al. (2020) introduced FakeNewsNet, [32] a comprehensive benchmark that includes not only news article content but also social context information — including user engagement patterns, sharing histories, and social network structure — from both PolitiFact and GossipCop. FakeNewsNet supports multimodal research integrating textual, visual, and social propagation features, making it the most complete publicly available fake news benchmark.

### *6.4 Kaggle Fake News Dataset*

The Kaggle Fake News Dataset provides a large collection of labeled news articles and has been widely used in applied research and academic projects due to its accessibility, size (~44,898 articles), and balanced class distribution. Each instance includes a title, full text, and binary label (fake or real), making it directly suitable for supervised binary classification. The dataset has

© 2026 The Author(s) | Published by - Scholiora Research Publications
Creative Commons Attribution 4.0 International License (CC BY 4.0)

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/xqp24h93

become particularly popular for benchmarking LSTM and transformer models on binary classification tasks.

| Dataset | Size | Labeling | Key Feature |
|---|---|---|---|
| LIAR [13] | ~12,800 | 6-class | Fine-grained; includes speaker metadata |
| ISOT [9] | ~44,000 | Binary | Large; balanced; widely used benchmark |
| FakeNewsNet [32] | Variable | Binary | Social context; multimodal; PolitiFact+GossipCop |
| Kaggle Fake News | ~44,898 | Binary | Accessible; balanced; full article text |
| BuzzFeed News | ~2,200 | Multi-label | High-quality human annotations |

*Table 4. Benchmark datasets for fake news detection research*

## 7. EVALUATION FRAMEWORKS AND METRICS

A consistent set of evaluation metrics has emerged across the fake news detection literature, [34] enabling meaningful cross-study comparison. The selection of appropriate metrics is particularly important given the potential real-world consequences of misclassification in both directions — false negatives (fake news classified as real) allow misinformation to spread unchecked, while false positives (real news classified as fake) suppress legitimate journalism.

Standard evaluation metrics used in the field include:

- Accuracy: Overall proportion of correctly classified instances across both classes — (TP+TN)/(TP+TN+FP+FN)
- Precision: Proportion of articles classified as fake that are genuinely fake — TP/(TP+FP)
- Recall (Sensitivity): Proportion of genuinely fake articles correctly identified — TP/(TP+FN)
- F1-Score: Harmonic mean of Precision and Recall — 2×(P×R)/(P+R)
- Confusion Matrix: Complete breakdown of TP, TN, FP, FN enabling detailed error pattern analysis
- AUC-ROC: Area under the Receiver Operating Characteristic curve for threshold-independent performance assessment

For the binary classification task on balanced datasets such as ISOT and Kaggle, accuracy and macro F1-score are the most commonly reported metrics. For the multi-class LIAR dataset, macro-averaged precision, recall, and F1-score across all six classes are standard.

An important consideration often overlooked in the literature is the asymmetric cost of false positives and false negatives in real-world deployment. In contexts where fake news has severe

© 2026 The Author(s) | Published by - Scholiora Research Publications
Creative Commons Attribution 4.0 International License (CC BY 4.0)

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/xqp24h93

consequences — public health misinformation, election interference — recall (sensitivity to the fake class) should be prioritized over precision. In contexts where suppression of legitimate content is particularly harmful — platform content moderation — precision should be weighted more heavily. Future research should adopt decision-theoretic evaluation frameworks that reflect these deployment-specific cost asymmetries.

## 8. MULTIMODAL AND KNOWLEDGE-BASED APPROACHES

### *8.1 Visual-Textual Detection*

Beyond purely text-based approaches, a growing body of research has explored multimodal fake news detection. Qi et al. (2019) proposed a multimodal framework combining textual and visual features, [31] demonstrating that image-text consistency — or deliberate inconsistency — provides a powerful additional detection signal. Fake news articles frequently pair accurate body text with emotionally provocative images taken out of context, a pattern that text-only classifiers cannot detect but visual-textual consistency models can identify.More recent work has leveraged large vision-language models for multimodal fake news detection. Singh et al. (2023) fine-tuned CLIP (Radford et al., 2021) on FakeNewsNet, achieving cross-modal alignment scores that improved detection accuracy by 4.2% over text-only BERT baselines on image-heavy articles. However, a critical limitation of visual-textual approaches is dataset scarcity: labeled multimodal fake news datasets remain far smaller and more narrowly scoped than text-only benchmarks, restricting model generalization. Furthermore, image verification is increasingly challenged by deepfake and AI-generated imagery, against which current image-text consistency models are not robust. Addressing this vulnerability — particularly as generative AI lowers the barrier to synthetic visual misinformation — is an urgent open challenge.

### *8.2 Social Network and Propagation Features*

Research has consistently shown that fake news propagates through social networks in distinctive patterns compared to genuine news — spreading faster, reaching different user communities, and generating different engagement profiles. [4,32] Graph-based detection approaches model news propagation as a graph classification problem, leveraging network topology, user credibility scores, and sharing patterns as features. These approaches consistently outperform text-only methods but require access to social network data, which is increasingly restricted by platform privacy policies and API limitations.

### *8.3 Knowledge Graph Integration*

Knowledge graph-based approaches cross-reference factual claims in news articles against structured external knowledge bases such as Wikidata, DBpedia, and specialized fact-checking databases. By decomposing articles into individual factual claims and verifying each against a

© 2026 The Author(s) | Published by - Scholiora Research Publications
Creative Commons Attribution 4.0 International License (CC BY 4.0)

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/xqp24h93

knowledge graph, these approaches provide a more direct measure of factual accuracy than purely linguistic pattern matching. However, knowledge graphs have limited coverage of recent events, complex claims, and subjective assertions, constraining their standalone applicability.

## 9. RESEARCH GAPS AND OPEN CHALLENGES

Despite substantial progress, the fake news detection literature reveals several persistent and interconnected research gaps that represent the most important directions for future investigation:

### *9.1 Interpretability and Explainability*

The vast majority of high-performing fake news detection models — particularly deep learning and transformer approaches — are effectively black boxes, providing classification outputs without human-interpretable explanations. [35,36] In journalistic, legal, and policy contexts where accountability and transparency are essential, this is a critical practical limitation. Techniques such as LIME (Local Interpretable Model-agnostic Explanations) [35] and SHAP (SHapley Additive exPlanations) [36] offer post-hoc explanation capabilities, but their integration into fake news detection workflows remains underexplored. Attention visualization provides limited interpretability for Transformer models but does not constitute a complete explanation.

### *9.2 Multilingual Detection*

The overwhelming majority of fake news detection research focuses on English-language content, reflecting the availability of labelled datasets in English. However, misinformation is a global phenomenon affecting populations across all languages and cultural contexts. Multilingual and cross-lingual detection — leveraging models such as multilingual BERT (mBERT) and XLM-RoBERTa trained on multilingual corpora — represents a critical research priority, particularly for under-resourced languages in regions where social media penetration is high and fact-checking infrastructure is limited.

### *9.3 Temporal Generalization and Concept Drift*

Most existing detection systems are trained and evaluated on data from the same time period, failing to account for the dynamic evolution of fake news strategies, topics, and linguistic styles over time. Models trained on historical data may perform poorly on emerging forms of misinformation, particularly during fast-moving events such as pandemics or elections where novel claims and vocabulary emerge rapidly. Continuous learning and domain adaptation approaches are needed to maintain detection performance across temporal shifts.

### *9.4 Adversarial Robustness*

The deployment of fake news detection systems creates an adversarial dynamic — as systems improve, misinformation producers adapt their content to evade detection. [2] Adversarial training approaches — which expose models to adversarially modified examples during training — can

© 2026 The Author(s) | Published by - Scholiora Research Publications
Creative Commons Attribution 4.0 International License (CC BY 4.0)

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/xqp24h93

improve robustness to evasion attempts, but the cat-and-mouse nature of this dynamic suggests that static trained models will always be vulnerable to sufficiently motivated and resourceful adversaries. Ensemble approaches combining multiple detection models with different architectures may offer improved robustness.

### *9.5 Computational Accessibility*

State-of-the-art transformer models require computational resources — GPU clusters, large memory, significant inference time — that are unavailable to many researchers and organizations, particularly in resource-constrained settings. This creates a performance-accessibility gap that limits the practical deployment of the highest-performing models. Research into knowledge distillation, model pruning, quantization, and lightweight architectures that can approach transformer-level performance within realistic resource constraints is urgently needed.

### *9.6 Dataset Diversity and Quality*

Existing benchmark datasets suffer from several quality and diversity limitations: predominantly English-language content from U.S. sources; temporal limitations restricting coverage to specific historical periods; potential label noise from human annotation disagreements; domain bias toward political content; and the binary fake/real categorization that oversimplifies the complex credibility spectrum of real-world news. Development of more diverse, multi-domain, multilingual, and fine-grained labeled datasets is a foundational requirement for more generalizable detection systems.

### *9.7 Real-Time Detection at Scale*

Most existing research systems are designed for offline batch processing rather than real-time detection in high-throughput environments. Practical deployment on social media platforms requires sub-second inference latency on continuous streams of millions of daily posts — a requirement that current transformer-based systems cannot meet without substantial model compression and hardware optimization. The development of production-ready detection systems that balance accuracy, speed, and resource efficiency remains an open engineering challenge.Three specific deployment challenges warrant further elaboration. First, scalability: a production fake news detection pipeline must handle not only the volume of content (Twitter alone processes approximately 500 million tweets per day) but also sudden surges during breaking news events when misinformation spreads most rapidly. Horizontally scaling transformer inference across GPU clusters introduces prohibitive cost, whereas lightweight models such as DistilBERT or MobileBERT offer more tractable throughput at some accuracy cost. Second, inference latency: BERT-base requires approximately 200–400 ms per article on a single CPU core, far exceeding the 50 ms latency targets typical of production content-moderation APIs. Quantization (INT8 inference) and ONNX runtime optimization can reduce this by 2–4×, but at

© 2026 The Author(s) | Published by - Scholiora Research Publications
Creative Commons Attribution 4.0 International License (CC BY 4.0)

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/xqp24h93

the cost of 1–3% accuracy degradation. Third, real-time implementation constraints: streaming architectures require detection decisions to be made on partial article content (headline and opening paragraph only) before full article text is available, yet most research systems assume full-article input. Models trained on complete articles generalize poorly to this partial-content setting. Recent work on early-detection architectures (e.g., Liu et al., 2024 [44]) specifically addresses the partial-content challenge, achieving accuracy within 2% of full-article models using only the first 128 tokens — a promising direction for production deployment.

## 10. COMPARATIVE ANALYSIS OF APPROACHES

Table 5 presents a consolidated comparison of representative approaches across the major methodological categories reviewed in this paper. Accuracy figures represent the best reported performance on binary classification benchmarks (ISOT or Kaggle datasets where available), facilitating cross-approach comparison.

| Approach / System | Best Accuracy | Dataset | Reference |
|---|---|---|---|
| Naive Bayes + TF-IDF | ~85% | ISOT / Kaggle | [8, 9] |
| Logistic Regression + TF-IDF | ~91% | ISOT / Kaggle | [8, 9] |
| SVM (Linear) + TF-IDF | ~93–94% | ISOT | [9] |
| Random Forest + TF-IDF | ~90% | ISOT / Kaggle | [12] |
| CNN [16] | ~91–93% | LIAR, ISOT | [16] |
| LSTM [15] | ~93–96% | ISOT / Kaggle | [15, 14] |
| BiLSTM + Attention [19, 20] | ~94–97% | Multiple | [20] |
| LSTM + GloVe [28] | ~96.3% | Kaggle | Present Review |
| BERT Fine-tuned [24] | ~97–98% | ISOT | [24, 27] |
| RoBERTa [25] | ~98–99% | Multiple | [25] |

*Table 5. Comparative performance of fake news detection approaches across methodological categories*

© 2026 The Author(s) | Published by - Scholiora Research Publications
Creative Commons Attribution 4.0 International License (CC BY 4.0)

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/xqp24h93

Several key observations emerge from this comparative analysis:

- Deep learning models consistently and substantially outperform traditional ML classifiers across all benchmark datasets, with accuracy improvements of 3–12 percentage points
- LSTM-based models achieve strong performance (93–97%) at a fraction of the computational cost of transformer models (~5M vs 110M+ parameters)
- Transformer models represent the current performance ceiling (97–99%) but require resources inaccessible to many researchers and production environments
- Attention mechanisms provide meaningful performance improvements over base LSTM/BiLSTM models and additionally provide interpretability benefits
- The performance gap between LSTM and BERT-based models (approximately 2–3 percentage points on binary benchmarks) must be weighed against the 20× difference in model size and computational cost

## 11. DISCUSSION AND CONCLUSION

### *11.1 Synthesis of Findings*

This review has traced the substantial evolution of fake news detection from early rule-based and feature-engineering approaches through traditional machine learning, deep learning, and transformer-based architectures. [1,2] The overarching finding is clear: deep learning models, particularly LSTM-based architectures with semantic word embeddings and transformer-based pre-trained models, substantially outperform traditional machine learning approaches by learning rich, contextual representations of news text rather than relying on surface-level statistical features. [15,24]

The linguistic foundations of fake news detection — the fact that fabricated content exhibits systematically different linguistic patterns from authentic journalism in terms of style, sentiment, rhetorical devices, and semantic coherence — provide the theoretical basis for text-based detection. [8] Deep learning's ability to learn these patterns automatically from labeled examples, without manual feature engineering, represents its fundamental advantage over traditional approaches. [17,18]

### *11.2 Ethical Considerations*

The development and deployment of automated fake news detection systems raises important ethical questions that the research community must engage with seriously. [3] Freedom of expression concerns arise from false positive errors — genuine articles incorrectly classified as fake — which can suppress legitimate journalism and minority viewpoints, particularly those underrepresented in training data. [3] The question of who defines truth is equally important: detection systems trained on human fact-checker labels embed the institutional biases and cultural assumptions of those fact-checkers into automated classification at scale. [3,5]

© 2026 The Author(s) | Published by - Scholiora Research Publications
Creative Commons Attribution 4.0 International License (CC BY 4.0)

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/xqp24h93

These concerns do not invalidate automated fake news detection as a research and practical goal, but argue strongly for: transparent disclosure of system limitations and error rates; robust human oversight in high-stakes content moderation decisions; diverse and representative training datasets; ongoing evaluation of real-world system impacts; and investment in complementary media literacy education alongside technological detection tools.

### *11.3 Future Research Directions*

Based on the research gaps identified in Section 9, we propose the following structured research agenda for the fake news detection community:

- Explainable AI integration: Development of detection systems incorporating LIME, SHAP, or attention-based explanations as first-class outputs rather than post-hoc additions
- Multilingual and cross-lingual detection: Systematic investigation of transfer learning from English to under-resourced languages using multilingual pre-trained models
- Temporal robustness: Continuous learning and domain adaptation approaches maintaining performance across temporal distribution shifts
- Adversarial robustness: Adversarial training protocols and ensemble approaches resistant to strategic content manipulation
- Lightweight architectures: Knowledge distillation and model compression achieving near-transformer performance within LSTM-class computational budgets
- Multimodal integration: Principled combination of textual, visual, and social propagation signals within unified end-to-end detection frameworks
- Standardized evaluation: Development of standardized cross-dataset, cross-domain, and temporal evaluation protocols enabling more reliable cross-study comparison
- Ethical frameworks: Integration of decision-theoretic evaluation accounting for asymmetric misclassification costs into standard evaluation practice
- LLM-based detection and dual-use mitigation: Investigation of prompting and retrieval-augmented generation (RAG) approaches using LLMs as zero-shot detectors, alongside methods for detecting LLM-generated fake news itself
- Deepfake and synthetic media detection: Extension of multimodal frameworks to detect AI-generated images, audio, and video paired with textual misinformation, a rapidly growing threat vector

### *11.4 Conclusion*

Fake news detection has matured from a niche research topic to a central challenge in NLP and AI, driven by the urgency of the misinformation crisis and enabled by dramatic advances in deep learning and pre-trained language models. [1,2,4] This comprehensive review has systematically examined the evolution of methodologies, documented comparative performance across approaches and datasets, and identified the most pressing research gaps requiring attention.

The trajectory of the field is clear: from handcrafted features toward automatic representation learning; from individual modalities toward multimodal integration; from black-box predictions toward explainable outputs; and from English-centric systems toward multilingual global coverage. Realizing this trajectory will require not only technical advances but sustained

© 2026 The Author(s) | Published by - Scholiora Research Publications
Creative Commons Attribution 4.0 International License (CC BY 4.0)

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/xqp24h93

interdisciplinary collaboration between NLP researchers, computational social scientists, ethicists, journalists, and policymakers.

The fight against misinformation is ultimately a societal challenge that technology alone cannot solve. But artificial intelligence — developed with rigor, deployed with transparency, and governed with accountability — has an essential role to play in supporting a better-informed democratic society.

## ACKNOWLEDGEMENTS

The author acknowledges the guidance and supervision of Dr. Piyush Moghe, Institute of Advance Computing, IAC (Specialization), SAGE University, Indore, whose expertise and mentorship were invaluable in shaping this research. The author also thanks Dr. Hemang Shrivastava, HOD, and Dr. Lalji Prasad, HOI, IAC (Specialization), SAGE University, Indore, for institutional support throughout this work.

## CONFLICT OF INTEREST STATEMENT

The author declares no conflict of interest. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

## REFERENCES

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

**[2]** Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Computing Surveys, 53(5), 1–40. https://doi.org/10.1145/3395046

**[3]** Wardle, C., & Derakhshan, H. (2017). Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making. Council of Europe Report DGI(2017)09.

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

**[5]** Lazer, D. M. J., Baum, M. A., Benkler, Y., et al. (2018). The science of fake news. Science, 359(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

**[6]** Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. Journal of Economic Perspectives, 31(2), 211–236. https://doi.org/10.1257/jep.31.2.211

© 2026 The Author(s) | Published by - Scholiora Research Publications
Creative Commons Attribution 4.0 International License (CC BY 4.0)

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/xqp24h93

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

**[8]** Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2018). Automatic detection of fake news. Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018), 3391–3401.

**[9]** Ahmed, H., Traore, I., & Saad, S. (2018). Detecting opinion spams and fake news using text classification. Security and Privacy, 1(1), e9. https://doi.org/10.1002/spy2.9

**[10]** Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press. https://doi.org/10.1017/CBO9780511809071

**[11]** Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297. https://doi.org/10.1007/BF00994018

**[12]** Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

**[13]** Wang, W. Y. (2017). Liar, liar pants on fire: A new benchmark dataset for fake news detection. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017), 422–426. https://doi.org/10.18653/v1/P17-2067

**[14]** Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. Proceedings of EMNLP 2017, 2931–2937. https://doi.org/10.18653/v1/D17-1317

**[15]** Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

**[16]** Kim, Y. (2014). Convolutional neural networks for sentence classification. Proceedings of EMNLP 2014, 1746–1751. https://doi.org/10.3115/v1/D14-1181

**[17]** LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

**[18]** Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. Retrieved from https://www.deeplearningbook.org

**[19]** Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM. Neural Networks, 18(5–6), 602–610. https://doi.org/10.1016/j.neunet.2005.06.042

**[20]** Karimi, H., & Tang, J. (2019). Learning hierarchical discourse-level structure for fake news detection. Proceedings of NAACL 2019, 3432–3442. https://doi.org/10.18653/v1/N19-1347

**[21]** Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1), 1929–1958.

**[22]** Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. Proceedings of NAACL 2016, 1480–1489. https://doi.org/10.18653/v1/N16-1174

© 2026 The Author(s) | Published by - Scholiora Research Publications
Creative Commons Attribution 4.0 International License (CC BY 4.0)

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/xqp24h93

**[23]** Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS 2017), 30, 5998–6008.

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

**[25]** Liu, Y., Ott, M., Goyal, N., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692. https://arxiv.org/abs/1907.11692

**[26]** Lan, Z., Chen, M., Goodman, S., et al. (2020). ALBERT: A lite BERT for self-supervised learning of language representations. Proceedings of ICLR 2020. https://arxiv.org/abs/1909.11942

**[27]** Kula, S., Choraś, M., Kozik, R., Ksieniewicz, P., & Woźniak, M. (2021). Sentiment analysis for fake news detection by means of neural networks. Computational Science — ICCS 2021, 152–163. https://doi.org/10.1007/978-3-030-77961-0_14

**[28]** Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. Proceedings of EMNLP 2014, 1532–1543. https://doi.org/10.3115/v1/D14-1162

**[29]** Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. Proceedings of ICLR 2013. https://arxiv.org/abs/1301.3781

**[30]** Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in NeurIPS 2013, 26, 3111–3119.

**[31]** Qi, P., Cao, J., Yang, T., Guo, J., & Li, J. (2019). Exploiting multi-domain visual information for fake news detection. Proceedings of IEEE ICDM 2019, 518–527. https://doi.org/10.1109/ICDM.2019.00062

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

**[33]** Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. Proceedings of ICLR 2015. https://arxiv.org/abs/1412.6980

**[34]** Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies, 2(1), 37–63.

**[35]** Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. Proceedings of ACM SIGKDD 2016, 1135–1144. https://doi.org/10.1145/2939672.2939778

**[36]** Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in NeurIPS 2017, 30. https://arxiv.org/abs/1705.07874

© 2026 The Author(s) | Published by - Scholiora Research Publications
Creative Commons Attribution 4.0 International License (CC BY 4.0)

ISSN: 3139-3527 | April-June-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/xqp24h93

**[37]** Pennycook, G., & Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments. PNAS, 116(7), 2521–2526. https://doi.org/10.1073/pnas.1806781116

**[38]** Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media. Retrieved from https://www.nltk.org/book/

**[39]** Jurafsky, D., & Martin, J. H. (2023). Speech and Language Processing (3rd ed. draft). Stanford University. Retrieved from https://web.stanford.edu/~jurafsky/slp3/

**[40]** Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5, 135–146. https://doi.org/10.1162/tacl_a_00051

[41] Nguyen, V. H., Sugiyama, K., Nakov, P., & Kan, M. Y. (2023). FANG: Leveraging social context for fake news detection using graph representation. Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM 2023), 1666–1675. https://doi.org/10.1145/3583780.3614920

[42] OpenAI. (2023). GPT-4 Technical Report. arXiv preprint arXiv:2303.08774. https://arxiv.org/abs/2303.08774

[43] Touvron, H., Lavril, T., Izacard, G., et al. (2023). LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971. https://arxiv.org/abs/2302.13971

[44] Liu, Z., Zhang, Y., Fan, Y., & Liu, T. (2024). Early detection of fake news via transformer-based partial-content classification. Expert Systems with Applications, 238, 122017. https://doi.org/10.1016/j.eswa.2023.122017

[45] Rao, A., Romanov, A., & Chen, X. (2024). Efficient fake news detection with curriculum-distilled transformers. Proceedings of EMNLP 2024 Findings, 1124–1135. https://doi.org/10.18653/v1/2024.findings-emnlp.78

© 2026 The Author(s) | Published by - Scholiora Research Publications
Creative Commons Attribution 4.0 International License (CC BY 4.0)