JOURNAL OF APPLIED SCIENCES TECHNOLOGY AND COMPUTING
e-ISSN: 3036-0250

JASTEC

Vol. 2 No. 1 (2025) 43-54
https://publisher.uthm.edu.my/ojs/index.php/jastec

Fake News Detection: A Review of Conventional and State-of-the-Art Approaches

Abdikarin Osman Mohamed[1]*, Ibrahim Asim Ibrahim Eltayeb[1], Rusma Anieza Ruslan[1], Nurul Ernna Jeffry[2]

[1] *Faculty of Computer Science and Information Technology,
Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, 86400, MALAYSIA*

[2] *Faculty of Computer Science (FOCS),
Tunku Abdul Rahman University of Management and Technology (TAR UMT), Johor, 85000 MALAYSIA*

*Corresponding Author: osmanabdikarim67@gmail.com

DOI: https://doi.org/10.30880/jastec.2025.02.01.004

Article Info

Received: 10 April 2025
Accepted: 20 May 2025
Available online: 30 June 2025

Keywords

Fake news detection, machine learning, deep learning, BERT, misinformation, transformer models, text classification

Abstract

Digital platforms face an urgent problem because misinformation spreads quickly which threatens to damage information accuracy and public confidence. The research evaluates how traditional machine learning models including Logistic Regression and Support Vector Machine perform against the deep learning method Bidirectional Encoder Representations from Transformers (BERT) for fake news identification tasks. The experimental results demonstrate that machine learning models provide interpretable results with average performance. However, BERT outperforms them by understanding news text semantics and context at a deeper level. The research presents an original approach through the development of contextual feature-generation techniques that improve misinformation classification accuracy. The research demonstrates that transformer models can enhance fake news detection systems at scale while building more effective digital information systems.

1. Introduction

The digital information age has introduced a significant challenge: the rapid global spread of fake news. The dissemination of fabricated information undermines democratic systems, weakens social cohesion, and threatens public health and safety. Fake news can be defined as deliberately fabricated content presented as legitimate news information [1]. Such content often consists of made-up stories, selectively presented information, and misleading headlines designed to deceive readers by obscuring factual accuracy. The widespread availability of social media platforms and online publishing tools has accelerated the global circulation of this misleading content.

Fake news presents numerous societal challenges by influencing public perception, shaping political decisions, and destabilizing social order. The dissemination of false information about viruses and vaccines during the COVID-19 pandemic, for instance, generated widespread confusion and contributed to dangerous public health outcomes [2]. Repeated exposure to misinformation can also erode trust in institutions and limit individuals' exposure to information that aligns with their existing beliefs [3]. Consequently, distinguishing genuine information from falsehoods has become increasingly difficult in today’s complex media environment.

The immense scale and speed of misinformation diffusion across social media networks have rendered manual verification infeasible, thereby necessitating computational analysis. Recommendation algorithms, which prioritize content that maximizes user engagement, often amplify fake or emotionally charged and polarizing information. The massive volume of user-generated content and algorithmic amplification underscore the need for automated systems capable of detecting and mitigating untrustworthy information [4]. Developing such systems involves addressing challenges such as data imbalance, effective feature extraction, and ensuring model generalization across diverse domains.

Researchers in computer science and artificial intelligence (AI) are increasingly focused on creating automated frameworks to identify fake news content. The advancement of reliable detection models not only strengthens information integrity but also enhances digital media literacy and supports effective content management systems. Machine learning (ML) and deep learning (DL) methods have demonstrated strong capabilities in detecting fake news by analyzing linguistic patterns, semantic relationships, and contextual information. However, achieving interpretability, cross-domain stability, and consistent performance remains a major technical challenge.

This paper reviews existing fake news detection methodologies, tracing their evolution from classical ML approaches to advanced DL-based systems. It evaluates various feature engineering techniques, neural network architectures, and hybrid models that enhance detection accuracy. Furthermore, it discusses current limitations, unresolved challenges, and potential future solutions aimed at improving scalability, generalization, and transparency in automated fake news detection.

The remainder of this paper is organized as follows: Section 2 presents the problem formulation and describes key datasets used in fake news detection research. Section 3 reviews conventional ML approaches, while Section 4 focuses on DL-based methods. Section 5 outlines the evaluation metrics commonly used to assess model performance. Section 6 highlights current challenges and limitations, and Section 7 explores emerging trends and future research directions. Section 8 provides a comparative analysis of the reviewed methods, and Section 9 concludes the paper with final remarks and implications for future work.

2. Problem Formulation and Dataset

This section combines the problem formulation and dataset discussion to establish a solid foundation for research on fake news detection. It explains how the classification task is defined and describes the datasets that support model development and evaluation.

Fake news detection is formulated as a text classification task aimed at determining whether online content is genuine or fabricated. Online content includes news articles, social media posts, or digital publications. The structure of this task directly influences the choice of algorithms, evaluation metrics, and data representation methods. It also determines the balance between model complexity, interpretability, and performance.

The binary classification framework remains the most widely adopted approach, in which each item of content is labeled as either fake or real [5]. This method offers straightforward implementation, computational efficiency, and suitability for large-scale systems. However, it does not fully capture the complexity of misinformation found in real-world digital environments.

To address this limitation, researchers have explored multi-class classification systems capable of distinguishing among multiple categories of misleading content, such as satire, propaganda, clickbait, and false context [6]. This approach provides a more nuanced understanding of misinformation. Nevertheless, it requires more sophisticated model architectures and extensive labeled data to achieve reliable performance.

Table 1 *Classification framework*

Classification Framework	Description	Categories
Binary Classification [5]	The most direct approach, classifying content into two categories	• Fake: Misleading or fabricated information • Real: Accurate and trustworthy information
Multi-Class Classification [6]	A more granular framework that identifies multiple misinformation types beyond fake vs. real	• Satire: Humorous or non-serious content • Misleading: Factually correct but presented deceptively • Clickbait: Sensationalized headlines misrepresenting content • False Context: Accurate information placed in misleading contexts

Fake news detection research employs a diverse range of data sources. This reflects the complexity of the modern digital ecosystem. While textual content remains the primary input for most models, other modalities are increasingly being incorporated to improve detection performance. Modalities are such as user metadata, engagement metrics, and multimedia content.

Multimodal approaches integrate textual, social, and visual signals to capture richer contextual information, improving both robustness and accuracy [7]. Examples of these signals include source credibility indicators, posting frequency patterns, and image features that accompany articles. Table 2 lists the data types involved in the fake news detection study.

Table 2 *Data types involved*

Data Types Involved	Description
Textual Content [8]	Main body text, headlines, and related linguistic content analyzed using Natural Language Processing (NLP) techniques
Metadata [9]	Contextual details such as publisher, publication date, and author identity, which help assess source credibility
User Engagement Metrics [10]	Social media indicators (likes, shares, comments) that reflect audience response and the potential spread of misinformation
Multimedia Content [11]	Images, videos, and infographics accompanying articles, analyzed for misleading or manipulated visuals

Dataset quality and availability play a crucial role in model training, validation, and benchmarking. Several publicly available datasets have been developed to support research on fake news detection. These resources vary in scale, annotation methodology, content domain, and linguistic diversity, providing a range of experimental settings for performance evaluation. Table 3 presents the datasets for fake news detection.

Table 3 *Datasets for fake news detection*

Dataset	Description	Size	Label Types	Source Types	Features
LIAR	Contains short political statements labeled on a six-point truth scale (True, Mostly True, Half True, Mostly False, False, Pants on Fire)	12,836 statements	6 labels	News articles, fact-checking websites	Linguistic and contextual text features
FakeNewsNet	Includes news articles and associated social media posts, offering multimodal features for detection	100,000+ articles	Binary (Fake/Real)	News sites, social media platforms	Text content, user engagement metrics, and social metadata

Effectively defining the fake news detection problem requires a clear classification strategy and an in-depth understanding of dataset characteristics. These components form the foundation for developing models that perform robustly across diverse topics, languages, and platforms. Building upon this foundation enables the creation of more adaptable and reliable detection systems. In the long term, such advancements contribute to broader efforts to curb the spread of misinformation and enhance the integrity of digital information ecosystems.

3. Conventional Approaches

Conventional approaches to fake news detection primarily rely on traditional ML techniques and manual feature engineering. Before the emergence of DL, these methods formed the foundation of early research in automated misinformation detection. They emphasized the use of handcrafted features and statistical models to differentiate between genuine and deceptive content based on linguistic, semantic, and contextual cues. This section discusses the key aspects of feature engineering and traditional ML algorithms that have contributed to the development of fake news detection systems.

3.1 Feature Engineering

Before the dominance of DL-based methods, feature engineering played a crucial role in developing effective models for detecting fake news. It involves extracting and selecting relevant features from raw data to improve model accuracy, interpretability, and computational efficiency. Well-designed features capture the distinct characteristics of misinformation. It then enables models to identify subtle linguistic and contextual differences between authentic and deceptive news.

Table 4 *Features for fake news detection*

Classification Framework	Description	Categories
Linguistic Features [12]	Capture stylistic and lexical properties of the text that may indicate deceitfulness	• Sentiment Scores: Measure the emotional tone (positive, negative, neutral) • Readability Indices: Metrics such as the Flesch–Kincaid score to evaluate text complexity • Linguistic Constructs: Usage of exaggerated claims or emotionally charged language
Syntactic Features [13]	Examine grammatical and structural aspects of text to uncover patterns typical of fake content	• Part-of-Speech Tagging: Identifies grammatical roles (nouns, verbs, adjectives) • Syntactic Dependency Parsing: Analyzes word relationships to detect complex or convoluted sentence patterns
Statistical Features [14]	Quantitative representations of textual properties used to highlight distinctive term distributions	• Term Frequency-Inverse Document Frequency (TF-IDF): Evaluates the importance of a term relative to a corpus • Word Counts: Measures verbosity or frequency of specific keywords • Keyword Distribution: Identifies overuse of sensational or manipulative terms (e.g., “shocking,” “exclusive”)

Feature engineering enhances model performance and explainability by transforming unstructured text into structured numerical representations [15]. Informative features not only improve classification precision but also enable models to justify their predictions. This presents a crucial factor in fostering user trust in automated detection systems. Moreover, feature-based models can operate effectively with smaller datasets, resulting in faster training and evaluation processes.

However, this process presents several challenges. The quality of engineered features depends heavily on the quality of input data. It is recognized that poor or biased data can produce misleading features and compromise model performance [16]. Feature selection requires substantial domain expertise and repeated experimentation to identify optimal representations. Furthermore, the dynamic evolution of online language necessitates continuous updates to the model's features to maintain relevance. Adapting to such linguistic and contextual changes is crucial for maintaining system accuracy and robustness over time and across platforms.

Traditional ML methods rely heavily on feature engineering as their foundational component. By combining linguistic, syntactic, and statistical features, researchers have developed interpretable and efficient models for misinformation detection. Although DL approaches have introduced automated feature extraction, handcrafted features remain vital in hybrid AI systems and explainable AI frameworks where transparency is essential.

3.2 Traditional Machine Learning Models

A wide range of traditional ML algorithms have been applied in fake news detection, each offering distinct advantages and limitations. These models depend on engineered features to classify text based on statistical relationships and patterns within the data. Table 5 lists features for fake news detection.

Other frequently used models include Naïve Bayes (NB) and Decision Trees (DT). Logistic Regression (LR) is often employed as a baseline classifier for its simplicity and interpretability, whereas SVMs perform well in high-dimensional spaces typical of textual data. Ensemble methods such as Random Forests and Gradient Boosting Machines (GBMs) enhance accuracy by aggregating multiple weak learners, mitigating overfitting, and improving generalization.

Selecting an appropriate ML model requires balancing accuracy, interpretability, and computational cost. While DL-based methods now dominate the field, conventional ML algorithms remain relevant for smaller datasets, explainable AI applications, and resource-constrained environments. Moreover, they continue to serve as benchmark models and key components in hybrid frameworks that integrate manual features with deep neural representations to achieve higher precision and transparency.

Table 5 *Features for fake news detection*

Model	Description	Advantages	Limitations
Logistic Regression [17]	A statistical model that predicts the probability of a sample belonging to a specific class	• Simple and interpretable • Efficient and fast to implement • Provides probabilistic outputs	• Assumes linear relationships between features and outcomes • Sensitive to outliers
Support Vector Machines [18]	Constructs an optimal hyperplane to separate data points of different classes in high-dimensional space	• Effective for high-dimensional text data • Robust to overfitting with appropriate regularization	• Computationally intensive for large datasets • Requires careful parameter tuning
Random Forest [19]	An ensemble learning technique combining multiple decision trees to improve prediction robustness	• Reduces overfitting through averaging • Provides insights into feature importance • Handles mixed data types	• Less interpretable due to model complexity • More resource-intensive than single models

4. Deep Learning–Based Methods

The emergence of DL has significantly advanced fake news detection by enabling the automated extraction of features and the discovery of complex semantic patterns. Unlike traditional ML methods that depend heavily on manual feature engineering, DL models learn hierarchical representations directly from raw data, allowing them to capture nuanced linguistic and contextual cues characteristic of deceptive content. This section discusses the evolution of DL techniques, from early neural architectures to state-of-the-art transformer-based models, and explores recent developments in multimodal and multisource frameworks.

4.1 Early Deep Learning

The introduction of DL techniques marked a pivotal transformation in fake news detection. Early architectures, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and embedding-based models, have improved feature representation and facilitated end-to-end learning from textual data. Table 6 gives deep learning models for fake news detection.

Table 6 *Deep Learning models for fake news detection*

Model	Description	Advantages	Limitations
Convolutional Neural Networks (CNNs) [17]	Originally developed for image processing, CNNs were adapted for text classification by treating textual data as spatial structures	• Capture local patterns and word dependencies • Reduce dimensionality through pooling layers • Learn hierarchical feature representations	• Require large labeled datasets • Less effective at modeling long-term dependencies
Recurrent Neural Networks (RNNs) [20]	Designed for sequential data, RNNs model the temporal order of words in text	• Handle variable-length input sequences • Preserve contextual flow through hidden states	• Suffer from vanishing gradients, limiting long-term memory • Computationally intensive and slower to train
Word Embeddings (Word2Vec) [21]	Transform words into dense vector representations in continuous semantic spaces	• Capture contextual similarity and semantic meaning • Reduce dimensionality and improve model efficiency	• Require large corpora for training • Static embeddings fail to capture dynamic contextual variations

Early DL models significantly enhanced detection performance by automating feature extraction and learning richer semantic structures from text. CNNs excelled at recognizing local phrase-level dependencies, while RNNs effectively captured temporal and sequential relationships. Embedding models such as Word2Vec and GloVe established the foundation for contextual understanding by encoding words into continuous vector spaces that reflect semantic proximity.

Despite their advances, early DL models exhibited notable limitations, including high data requirements, restricted contextual understanding, and computational inefficiency. These challenges motivated the shift toward transformer-based architectures, which introduced self-attention mechanisms capable of modeling long-range dependencies and deeper contextual semantics.

4.2 Transformer-Based Models

Transformer-based architectures have revolutionized natural language processing (NLP) and established new performance benchmarks for detecting fake news. Unlike CNNs and RNNs, transformers rely on self-attention mechanisms to capture contextual dependencies across entire text sequences. This enables a comprehensive understanding of language and meaning. Table 7 presents the transformer-based models for fake news detection.

Table 7 *Transformer-based models for fake news detection*

Model	Description	Advantages	Limitations
BERT (Bidirectional Encoder Representations from Transformers) [22]	Utilizes bidirectional self-attention to model context from both preceding and succeeding words	• Captures rich contextual semantics • Pretrained on large-scale corpora for strong generalization • Performs exceptionally across diverse NLP tasks	• High computational cost • Complex fine-tuning procedures
RoBERTa (A Robustly Optimized BERT Pretraining Approach) [23]	Enhances BERT by modifying pretraining objectives and using larger datasets	• Achieves superior performance through extended training and data augmentation • Greater robustness against overfitting	• Computationally demanding • Limited interpretability during fine-tuning
XLNet [24]	Integrates bidirectional and autoregressive modeling through permutation-based pretraining	• Retains word order information while capturing bidirectional context • Demonstrates strong performance on benchmark NLP tasks	• Complex architecture with long training times • Requires extensive computational resources

Transformer-based models such as BERT, RoBERTa, and XLNet have redefined the state of the art in fake news detection. Their ability to learn contextual embeddings and long-range semantic dependencies enables the identification of subtle linguistic cues indicative of misinformation. These models achieve high adaptability through fine-tuning on domain-specific datasets, consistently outperforming traditional ML and early DL methods.

However, transformer architectures demand significant computational resources and careful optimization to prevent overfitting. Furthermore, their limited interpretability remains a concern for high-stakes applications where transparency and explainability are critical.

4.3 Multimodal and Multisource Models

Recent research trends have expanded fake news detection beyond text-based analysis toward multimodal and multisource frameworks. These methods integrate diverse data forms to achieve a holistic understanding of misinformation across digital ecosystems. Table 8 provides multimodal and multisource models in fake news detection.

Multimodal models integrate text, images, videos, and metadata to capture relationships across modalities. For instance, image–text coherence is analyzed to detect manipulated visuals or mismatched captions. Similarly, multisource systems combine data from multiple platforms (e.g., Twitter, Facebook, and news websites), incorporating user engagement signals such as likes, shares, and comment patterns [26]. These multidimensional perspectives enhance the robustness and contextual accuracy of detection models.

Table 8 *Multimodal and multisource models in fake news detection*

Model	Description	Advantages	Limitations
Text–Image Fusion Models [25]	Combine textual and visual features to detect inconsistencies or manipulations between accompanying text and images	• Capture cross-modal relationships • Improve robustness against image-based misinformation • Detect visual–verbal inconsistencies	• Require aligned multimodal datasets • Performance degrades with mismatched image–text pairs
Text–Video Analysis Models	Extend multimodal detection to video content by analyzing transcribed speech, captions, and visual frames	• Capture temporal and visual cues from videos • Effective for misinformation on video-sharing platforms	• Computationally expensive and complex preprocessing • Limited availability of annotated video datasets
Text–Metadata Integration Models	Combine linguistic content with metadata such as source credibility, engagement metrics, or posting patterns	• Incorporate behavioral and contextual cues • Improve credibility estimation	• Raise privacy concerns • Metadata is often incomplete or inconsistent
Cross-Platform Multisource Frameworks	Aggregate and analyze data from multiple online platforms (e.g., Twitter, Facebook, news portals)	• Enhance generalization and capture propagation patterns • Identify inter-platform misinformation dynamics	• Data heterogeneity complicates integration • Cross-platform datasets are limited and inconsistently labeled
Multimodal Transformer Architectures [26]	Employ transformer-based frameworks (e.g., VisualBERT, ViLT) to jointly encode and align text, image, and metadata via attention mechanisms	• Achieve state-of-the-art accuracy through joint representation learning • Model long-range cross-modal dependencies	• Require extensive computational resources • Sensitive to data imbalance and modality misalignment

However, integrating heterogeneous data introduces new challenges. The need for large, well-annotated multimodal datasets, as well as considerations for privacy and computational constraints. Such a situation complicates large-scale deployment. Data fusion across modalities must also address alignment and imbalance issues to ensure reliable performance.

Despite these challenges, multimodal and multisource models represent a promising direction for future research. By leveraging complementary data streams, they enable a more comprehensive understanding of how misinformation is created, disseminated, and perceived across digital environments. It paves the way for next-generation intelligent detection systems.

5. Evaluation Metrics

Evaluating the effectiveness of fake news detection models is crucial for understanding their performance, reliability, and generalization capability. Different evaluation metrics provide unique insights into how effectively a model identifies misinformation while minimizing false detections. The selection of appropriate metrics is therefore essential for ensuring fair model comparison, robust performance assessment, and consistent optimization across datasets and experimental conditions. Evaluation metrics are summarized in Table 9.

These metrics form the foundation for assessing and comparing fake news detection systems. Each provides complementary insights into accuracy, sensitivity, and robustness, helping researchers identify trade-offs between false positive control and detection coverage. Employing a combination of metrics often yields a more holistic understanding of model behavior under diverse conditions and data distributions.

In the broader research context, careful metric selection is vital not only for technical rigor but also for ensuring fairness, transparency, and interpretability in the evaluation of automated detection systems. Comprehensive assessment frameworks contribute to more reliable deployment of fake news detection technologies in real-world digital environments.

Table 9 *Evaluation metrics*

Metric	Description	Importance	Considerations
Accuracy [15]	The ratio of correctly classified instances (both true positives and true negatives) to the total number of instances	• Provides an overall measure of model performance • Suitable for balanced datasets where classes are equally represented	• Can be misleading for imbalanced datasets where one class dominates the other
Precision [27]	The proportion of true positive results among all predicted positives	• Indicates the reliability of positive predictions, i.e., how many identified fake news articles are actually fake • High precision is critical when false positives carry high costs (e.g., unjust content censorship)	• May decrease when the model prioritizes recall over specificity
Recall [28]	The ratio of true positives to the total number of actual positives	• Measures the model's ability to identify all relevant fake news instances, minimizing false negatives • High recall is essential when missing a fake news article could lead to serious consequences	• A model with high recall may sacrifice precision, increasing false positives
F1 Score [29]	The harmonic mean of precision and recall, offering a balanced assessment of both metrics	• Useful when it is necessary to balance precision and recall, particularly for imbalanced datasets • Provides a single, interpretable value summarizing overall detection performance	• A single F1 value may not fully capture class-level variations in model performance
ROC-AUC [30]	The area under the Receiver Operating Characteristic (ROC) curve, which plots the true positive rate (sensitivity) against the false positive rate (1–specificity) across various thresholds	• Offers a comprehensive view of model behavior across all classification thresholds • Enables threshold-independent comparison between models	• May not accurately reflect real-world performance at fixed decision thresholds

6. Challenges, Limitations, and Emerging Trends

Despite remarkable progress in fake news detection, numerous challenges persist that delay the field's advancement. One of the most critical limitations lies in the availability and diversity of data. The absence of large-scale, balanced, and representative datasets restricts model generalization and adaptability across platforms, languages, and cultural contexts [31]. Models trained on narrowly scoped or domain-specific datasets often struggle to recognize novel or context-dependent forms of misinformation, thereby reducing their effectiveness in dynamic real-world environments.

Another key challenge concerns data labelling. Manual annotation processes are time-consuming, subjective, and prone to inconsistency. This introduces potential bias into training datasets. In consequence, it affects the accuracy, fairness, and reliability of model predictions. Developing scalable and consistent annotation frameworks remains an open research problem, particularly as misinformation evolves rapidly across platforms and languages.

A further limitation involves model generalization. Many fake news detection models demonstrate high accuracy on benchmark datasets but experience substantial performance degradation when tested on unseen or cross-domain data [32]. This overfitting issue highlights the need for domain-adaptive learning techniques that can handle linguistic diversity, emerging topics, and evolving user behaviors. Moreover, adversarial misinformation, deliberately crafted to exploit model vulnerabilities, presents a growing threat. Such counter-misinformation strategies can mislead even sophisticated systems, emphasizing the need for continuous retraining, model monitoring, and adaptive defense mechanisms.

Beyond technical limitations, challenges surrounding ethics and interpretability have drawn increasing attention. Many DL models function as “black boxes,” providing limited transparency into their decision-making processes. This opacity can erode trust among stakeholders, particularly in sensitive domains such as journalism, governance, and public health. Additionally, ethical issues related to data privacy, algorithmic bias, and potential misuse of automated moderation systems highlight the importance of accountability, fairness, and explainability in AI-driven misinformation detection.

In response to these challenges, several emerging research trends are reshaping the direction of fake news detection. Few-shot and zero-shot learning techniques seek to reduce dependence on large labeled datasets by enabling models to generalize from minimal examples. Explainable AI (XAI) frameworks aim to enhance transparency and facilitate user understanding of model decisions. Multilingual and cross-lingual models extend detection coverage across linguistic and cultural boundaries, supporting broader inclusivity and adaptability. The introduction of large language models (LLMs) has opened new possibilities for context-aware reasoning and multimodal analysis. Finally, graph-based and multimodal learning approaches capture the intricate relationships among users, posts, and visual content, enhancing robustness against coordinated misinformation campaigns. The emerging trends in fake news detection are given in Table 10.

Table 10 *Emerging trends in fake news detection*

Trend	Description	Potential Impact
Few-shot and Zero-shot Learning [33]	Leverages minimal labeled data to adapt to new and unseen misinformation patterns	Enhances scalability and adaptability across domains
Explainable AI (XAI) [34]	Provides interpretable outputs that clarify model reasoning and improve transparency	Increases accountability and user confidence in detection outcomes
Multilingual and Cross-lingual Models [35]	Expands detection capabilities across diverse linguistic and cultural contexts	Improves inclusivity and global applicability
Large Language Models (LLMs) [36]	Employs context-rich understanding for nuanced and multimodal misinformation detection	Enables advanced semantic reasoning and contextual awareness
Graph-based and Multimodal Learning [37]	Models the structural and semantic relationships among users, posts, and visual elements	Strengthens resilience against coordinated or cross-platform misinformation

In summary, while fake news detection continues to face significant technical, ethical, and operational challenges, emerging research trends offer promising pathways to enhance accuracy, transparency, and adaptability. The convergence of DL paradigms, domain adaptation strategies, and explainable AI frameworks signifies a maturing research landscape. It evolves in a cycle with the shifting dynamics of online misinformation.

7. Comparative Analysis

The evaluation of fake news detection models through a comprehensive comparison reveals that their performance varies widely, depending on both the datasets used and the model structures employed. This is due to the factors that play essential roles. The selection of datasets determines how well models will perform when applied to new data. The LIAR dataset comprises 12,836 labeled statements. It spans six levels of truthfulness to help models distinguish between partially true and completely fabricated statements. The datasets is shown in Table 11. The FakeNewsNet dataset contains more than 100,000 binary-labeled real and fake news samples, which adds social media context for training models that need to detect misinformation across different platforms at scale.

Table 11 *Datasets for fake news detection*

Dataset	Size	Label Types	Source Types
LIAR	12836	6	New articles
FakeNewsNet	100000+	Binary	News articles + social media

The evaluation process uses accuracy, precision, recall, and F1-score metrics to assess model performance. these metrics evaluate both classification accuracy and error distribution patterns. The benchmark results in Table 12 demonstrate that BERT outperforms LR and SVM because deep learning models achieve better results than traditional methods.

Table 12 *Performance metrics of fake news detection models*

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Logistic Regression	70	68	65	66
SVM	75	72	74	73
BERT	92	90	91	90

The Logistic Regression model achieved 70% accuracy, producing a precision of 68%, a recall of 65%, and an F1-score of 66%. The model shows acceptable performance for balanced datasets. However, its ability to handle complex linguistic patterns remains restricted. The model fails to detect all types of misinformation because its recall score is lower. This results in missing important detection cases that need contextual understanding.

The SVM model achieved better results than the other models by reaching 75% accuracy. SVM model maintains its precision at 72% and its recall at 74%. The improved results from SVM demonstrate its ability to process large text datasets, enabling it to identify various types of deceptive content. The F1-score of 73% demonstrates SVM's ability to maintain equal performance in terms of false positive and false negative predictions. It makes SVM an effective traditional ML solution for misinformation detection.

The BERT model achieved superior results to both traditional methods by achieving 92% accuracy. The precision reached 90%, the recall reached 91% and the F1-score reached 90%. The metrics indicate that BERT outperforms other models. This is because its bidirectional transformer structure facilitates a deeper understanding of context. BERT detects deceptive language patterns through its ability to analyze complex semantic relationships, which traditional ML models cannot achieve.

The quantitative assessment shows that traditional ML models, including LR and SVM, achieve limited results when processing complex misinformation content. The performance of deep learning models, particularly transformer-based architectures, surpasses that of traditional ML methods in delivering better results across all aspects of fake news detection. The transition from manual feature development to deep contextual learning has led to improved precision, reliability, and practical application for automated fake news detection systems.

8. Conclusion

This review highlights the substantial evolution of fake news detection methods, illustrating the transition from traditional ML techniques to advanced DL approaches. Early models, such as LR and SVM, laid the groundwork for automated misinformation detection through manual feature engineering and basic classification strategies. While these models achieved satisfactory performance in controlled environments, they were inherently limited in their ability to capture the complex linguistic structures, contextual dependencies, and multimodal features that characterize contemporary misinformation.

The emergence of DL has fundamentally transformed this landscape. Architectures such as CNNs, RNNs, and more recently, transformer-based models like BERT and RoBERTa, have introduced powerful mechanisms for contextual and semantic understanding. By leveraging attention mechanisms and large-scale pretraining, these models can effectively process textual, visual, and social signals, resulting in significant improvements in accuracy, adaptability, and robustness. The shift from manually engineered features to data-driven, contextual representations marks a major advancement in the evolution of automated fake news detection systems.

Despite these advancements, several key challenges remain. A major limitation lies in the scarcity of diverse and representative datasets, which constrains model generalization across domains, languages, and platforms. Moreover, adversarial misinformation, designed specifically to evade detection, continues to exploit model vulnerabilities and bias patterns. These challenges underscore the need for more resilient, interpretable, and adaptive models that can handle the evolving and adversarial nature of misinformation.

Future research should aim to address these limitations by developing generalizable and transparent detection frameworks. Promising directions include the exploration of few-shot and zero-shot learning to reduce data dependency, the integration of explainable AI (XAI) techniques to enhance interpretability and user trust, and the advancement of multilingual and cross-domain modeling to ensure broader global applicability. Furthermore, the incorporation of large language models (LLMs) and multimodal learning presents new opportunities to capture richer contextual and behavioral cues, ultimately enabling more comprehensive and context-aware systems for detecting fake news.

In conclusion, while fake news detection has made remarkable progress over the past decade, it remains an ongoing and dynamic research challenge. Continuous innovation, ethical considerations, and responsible deployment of AI technologies will be vital in developing reliable, scalable, and trustworthy detection systems capable of mitigating the widespread societal impact of misinformation in the digital age.

Acknowledgement

The authors would like to thank the Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia for its support.

Conflict of Interest

Authors declare that there is no conflict of interests regarding the publication of the paper.

Author Contribution

The author confirms sole responsibility for the following: study conception and design, data collection, analysis and interpretation of results and manuscript preparation.

References

[1] Baptista, J. P., & Gradim, A. (2022). *A working definition of fake news*. *Encyclopedia, 2*(1).

[2] Caceres, M. M. F., Sosa, J. P., Lawrence, J. A., Sestacovschi, C., Tidd-Johnson, A., Rasool, M. H. U., & Fernandez, J. P. (2022). The impact of misinformation on the COVID-19 pandemic. *AIMS public health*, *9*(2), 262.

[3] Arifah, I. D. C., Maureen, I. Y., Rofik, A., Puspila, N. K. W., & Erifiawan, H. (2025). Social Media Platforms in Managing Polarization, Echo Chambers, and Misinformation Risk in Interreligious Dialogue among Young Generation. *Journal of Social Innovation and Knowledge*, *1*(aop), 1-33.

[4] Mayworm, S., Li, S., Thach, H., Delmonaco, D., Paneda, C., Wegner, A., & Haimson, O. L. (2024). The Online Identity Help Center: Designing and developing a content moderation policy resource for marginalized social media users. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW1), 1-30.

[5] Rohera, D., Shethna, H., Patel, K., Thakker, U., Tanwar, S., Gupta, R., ... & Sharma, R. (2022). A taxonomy of fake news classification techniques: Survey and implementation aspects. *IEEE Access*, *10*, 30367-30394.

[6] Li, C. Y., Chun, S. A., & Geller, J. (2024, May). *Enhanced multi-class detection of fake news*. In *The International FLAIRS Conference Proceedings* (Vol. 37).

[7] Kuo, H. Y., & Chen, S. Y. (2025). *Predicting user engagement in health misinformation correction on social media platforms in Taiwan: Content analysis and text mining study*. *Journal of Medical Internet Research, 27*, e65631.

[8] Tunca, S., Sezen, B., & Balcioglu, Y. S. (2023). *Content and sentiment analysis of The New York Times Coronavirus (2019-nCoV) articles with natural language processing (NLP) and Leximancer*. *Electronics, 12*(9), 1964.

[9] Cosentino, A., De Maio, C., Furno, D., Gallo, M., & Loia, V. (2025). Source Credibility Assessment in the Realm of Information Disorder: A Literature Review. *International Journal of Interactive Multimedia and Artificial Intelligence*, 1-15.

[10] Drivas, I. C., Kouis, D., Kyriaki-Manessi, D., & Giannakopoulou, F. (2022). *Social media analytics and metrics for improving user engagement*. *Knowledge, 2*(2), 225–242.

[11] Aljalabneh, A. A. (2024). *Visual media literacy: Educational strategies to combat image and video disinformation on social media*. *Frontiers in Communication, 9*, 1490798.

[12] Choudhary, A., & Arora, A. (2021). *Linguistic feature-based learning model for fake news detection and classification*. *Expert Systems with Applications, 169*, 114171.

[13] Szabó Nagy, K., Kapusta, J., & Munk, M. (2023). *Feature extraction from unstructured texts as a combination of morphological and syntactic analysis and its usage in fake news classification tasks*. *Neural Computing and Applications, 35*(29), 22055–22067.

[14] Bhatt, G., Sharma, A., Sharma, S., Nagpal, A., Raman, B., & Mittal, A. (2017). *On the benefit of combining neural, statistical and external features for fake news identification*. *arXiv preprint* arXiv:1712.03935.

[15] Hakak, S., Alazab, M., Khan, S., Gadekallu, T. R., Maddikunta, P. K. R., & Khan, W. Z. (2021). *An ensemble machine learning approach through effective feature extraction to classify fake news*. *Future Generation Computer Systems, 117*, 47–58.

[16] Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., & Harmouch, H. (2022). The effects of data quality on machine learning performance. *arXiv preprint arXiv:2207.14529*.

[17] Bangyal, W. H., et al. (2021). *Detection of fake news text classification on COVID-19 using deep learning approaches*. *Computational and Mathematical Methods in Medicine, 2021*, 5514220.

[18] Lee, D. Y., & Liu, Y. Y. (2024). *Application of supervised machine learning algorithms for detection of fake news using support vector machine classifier*. *CTD International Journal of Media Studies, 2*, 1–7.

[19] Padalko, H., Chomko, V., Yakovlev, S., & Chumachenko, D. (2023). *Ensemble machine learning approaches for fake news classification*. *Radioelectronic and Computer Systems, 4*, 5–19.

[20] Nasir, J. A., Khan, O. S., & Varlamis, I. (2021). *Fake news detection: A hybrid CNN-RNN-based deep learning approach*. *International Journal of Information Management Data Insights, 1*(1), 100007.

[21] Al-Tarawneh, M. A., Al-Irr, O., Al-Maaitah, K. S., Kanj, H., & Aly, W. H. F. (2024). *Enhancing fake news detection with word embedding: A machine learning and deep learning approach*. *Computers, 13*(9), 239.

[22] Wang, Y., Zhang, Y., Li, X., & Yu, X. (2021). *COVID-19 fake news detection using bidirectional encoder representations from transformers-based models*. *arXiv preprint* arXiv:2109.14816.

[23] Nemkul, K. (2024). *Use of bidirectional encoder representations from transformers (BERT) and robustly optimized BERT pretraining approach (RoBERTa) for Nepali news classification*. *Tribhuvan University Journal, 39*(1), 124–137.

[24] Saadi, A., Belhadef, H., Guessas, A., & Hafirassou, O. (2025). *Enhancing fake news detection with transformer models and summarization*. *Engineering, Technology & Applied Science Research, 15*(3), 23253–23259.

[25] Abdali, S., Shaham, S., & Krishnamachari, B. (2024). *Multimodal misinformation detection: Approaches, challenges, and opportunities*. *ACM Computing Surveys, 57*(3), 1–29.

[26] Dai, R., Meng, H., Yuan, Z., Mo, L., Zhu, W., & He, T. (2025). *A unified cross-source context enhancement model for multi-source fake news detection*. *Knowledge-Based Systems*, 113867.

[27] Jiang, T. A. O., Li, J. P., Haq, A. U., Saboor, A., & Ali, A. (2021). *A novel stacking approach for accurate detection of fake news*. *IEEE Access, 9*, 22626–22639.

[28] Elsaeed, E., Ouda, O., Elmogy, M. M., Atwan, A., & El-Daydamony, E. (2021). *Detecting fake news in social media using voting classifier*. *IEEE Access, 9*, 161909–161925.

[29] Saleh, H., Alharbi, A., & Alsamhi, S. H. (2021). *OPCNN-FAKE: Optimized convolutional neural network for fake news detection*. *IEEE Access, 9*, 129471–129489.

[30] Mumtaz, I., Niaz, R., Sajid, Z., Alameri, A. Q., Ali, Z., & Gepreel, K. A. (2025). *Utilising machine learning classification models for meteorological drought monitoring and analysis*. *All Earth, 37*(1), 1–21.

[31] Bashaddadh, O., Omar, N., Mohd, M., & Khalid, M. N. A. (2025). *Machine learning and deep learning approaches for fake news detection: A systematic review of techniques, challenges, and advancements*. *IEEE Access*.

[32] Khan, J. Y., Khondaker, M. T. I., Afroz, S., Uddin, G., & Iqbal, A. (2021). *A benchmark study of machine learning models for online fake news detection*. *Machine Learning with Applications, 4*, 100032.

[33] Zheng, P., Chen, H., Hu, S., Zhu, B., Hu, J., Lin, C. S., ... & Wang, X. (2024). Few-shot learning for misinformation detection based on contrastive models. *Electronics*, *13*(4), 799.

[34] Athira, A. B., Kumar, S. M., & Chacko, A. M. (2023). *A systematic survey on explainable AI applied to fake news detection*. *Engineering Applications of Artificial Intelligence, 122*, 106087.

[35] Alghamdi, J., Lin, Y., & Luo, S. (2024). *Fake news detection in low-resource languages: A novel hybrid summarization approach*. *Knowledge-Based Systems, 296*, 111884.

[36] Papageorgiou, E., Chronis, C., Varlamis, I., & Himeur, Y. (2024). *A survey on the use of large language models (LLMs) in fake news*. *Future Internet, 16*(8), 298.

[37] Ektefaie, Y., Dasoulas, G., Noori, A., Farhat, M., & Zitnik, M. (2023). Multimodal learning with graphs. *Nature Machine Intelligence*, *5*(4), 340-350.