MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 2460-2465

# A Comprehensive Review of Explainable AI Methods for Transformer-Based Fake News Detection

Kanchan shahi[1]
*Department of CSE*
*Krishna Institute of Engineering & Technology (KIET),*
*Ghaziabad, Delhi-NCR, Uttar Pradesh, India*
kanchan.mtechcse2604@kiet.edu

Seema Maitrey[2]
*Department of CSE*
*Krishna Institute of Engineering & Technology (KIET),*
*Ghaziabad, Delhi-NCR, Uttar Pradesh, India*
seema.maitrey@kiet.edu

**Abstract—**
*The rapid growth of online news dissemination through social media has significantly increased the spread of misinformation and fake news. Although recent advances in deep learning and transformer-based models have improved the accuracy of automated fake news detection systems, these models often function as black-box systems, limiting their transparency and trustworthiness. Explainable Artificial Intelligence (XAI) has emerged as a promising approach to address this limitation by providing interpretable insights into model predictions. This paper presents a comprehensive review of explainable AI techniques applied to transformer-based fake news detection systems. The study systematically analyzes widely used methods such as Local Interpretable Model-agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP), Integrated Gradients, and attention-based mechanisms. In addition, the paper examines commonly used datasets, model architectures, and evaluation strategies for both classification performance and explanation quality. Furthermore, a conceptual framework is proposed based on insights derived from the literature to integrate detection and explainability in a unified pipeline. The review identifies key challenges, including the lack of standardized evaluation metrics for explainability, computational complexity, and the gap between machine-generated explanations and human interpretability. Finally, future research directions are outlined to guide the development of more transparent, reliable, and trustworthy fake news detection systems.*

Keywords— Fake News Detection, Explainable Artificial Intelligence (XAI), Deep Learning, Transformers, Social Media, Misinformation Detection, Interpretability, LIME, SHAP.

## I. INTRODUCTION

The rapid growth of online communication platforms and social media has greatly accelerated the production and dissemination of information. While these platforms enable instant global communication, they also facilitate the widespread circulation of misinformation and fake news. The uncontrolled spread of misleading information has emerged as a serious challenge affecting public health, social stability, and democratic processes [1], [5]. During the COVID-19 pandemic, for example, false health-related claims spread rapidly across social networks and were found to propagate faster than verified information, significantly influencing public perception and behavior [8]. Similarly, misinformation related to elections has disrupted democratic processes in several countries [7], [15]. To address this problem, automated fake news detection

has become an important research area within natural language processing and machine learning. Early studies relied on traditional machine learning algorithms that used manually engineered linguistic and social features to classify news content [1], [3]. With the advancement of deep learning, models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) began to capture complex semantic patterns in textual data [17]. More recently, transformer-based models such as BERT and RoBERTa have demonstrated superior performance due to their ability to understand contextual relationships within text [10], [18]. Researchers have also explored multimodal approaches that combine textual, visual, and social context information to improve detection accuracy [11], [21]. Despite achieving high classification performance, many deep learning models operate as black-box systems, making it difficult to understand the reasoning behind their predictions. In critical domains such as journalism, policymaking, and public health, the lack of transparency can limit the trust and adoption of automated detection systems. To address this limitation, Explainable Artificial Intelligence (XAI) techniques have been introduced to provide interpretable insights into model predictions [26]. Methods such as Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) help identify the features that influence model decisions and improve transparency in machine learning systems [26], [20].

However, recent studies suggest that many explanation methods may produce unstable or incomplete interpretations when models or data distributions change, raising concerns about their reliability in real-world applications [20], [34]. Therefore, evaluating the effectiveness, robustness, and trustworthiness of explainability techniques remains an important research challenge in the development of reliable fake news detection systems.

## II. REVIEW METHODOLOGY

This study follows a structured literature review approach to systematically analyze existing research on explainable artificial intelligence (XAI) for transformer-based fake news detection.

### *A. Data Sources*

Relevant research articles were collected from major academic databases, including IEEE Xplore, SpringerLink, ScienceDirect, ACM Digital Library, and Google Scholar.

### *B. Search Strategy*

The literature search was performed using combinations of the following keywords:
"fake news detection", "misinformation detection", "explainable AI", "XAI", "transformers", "BERT", "RoBERTa", "interpretability", and "deep learning".

### *C. Inclusion Criteria*

The following criteria were used to select relevant studies:
• Papers published between 2017 and 2025
• Studies focusing on fake news or misinformation detection
• Research involving explainable AI techniques or interpretability
• Papers published in peer-reviewed journals or reputed conferences

### *D. Exclusion Criteria*

• Articles not related to fake news detection
• Papers without significant methodological contribution
• Duplicate or non-peer-reviewed sources

### *E. Study Selection Process*

An initial pool of research papers was identified through keyword-based search. After removing duplicates and irrelevant studies, the remaining papers were screened based on abstracts and full-text analysis. Finally, the most relevant studies were selected for detailed review and comparative analysis. This systematic approach ensures comprehensive coverage of the literature and reduces selection bias in the review process.

## III. LITERATURE REVIEW

Research on fake news detection has gained significant attention over the past decade due to the rapid spread of misinformation on social media platforms. Researchers have proposed various approaches ranging from traditional machine learning models to advanced deep learning architectures and explainable artificial intelligence (XAI) techniques. The following sections summarize the major developments in fake news detection research from 2017 to 2025, focusing on detection methods, datasets, interpretability approaches, and model performance. Early studies primarily focused on traditional machine learning approaches that relied on linguistic, social, and temporal features extracted from news content and user interactions. One of the earliest comprehensive analyses was

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 2460-2465

presented by Shu et al., who explored fake news detection from a data mining perspective and identified key feature categories for misinformation analysis [1]. Similarly, Wang introduced the widely used LIAR dataset, which became an important benchmark for evaluating fake news classification models [2]. During this period, conventional classifiers such as Naïve Bayes, Support Vector Machines, and Random Forests were commonly used, although they required significant manual feature engineering [3]. Subsequent research shifted toward deep learning techniques that could automatically learn complex textual representations. Neural network models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) demonstrated improved performance in detecting misleading content by capturing contextual patterns in textual data [17]. In parallel, studies analyzing information diffusion showed that false information spreads more rapidly and widely than factual information on social networks, highlighting the urgency of developing automated detection systems [5].

With the advancement of natural language processing, researchers increasingly adopted transformer-based architectures such as BERT and RoBERTa, which significantly improved contextual understanding in fake news detection tasks [10]. Additionally, several benchmark datasets, including FakeNewsNet, enabled researchers to integrate textual content with social context information for more comprehensive analysis [6]. Surveys conducted by Zhou and Zafarani and other researchers provided detailed taxonomies of detection approaches and identified key challenges such as dataset bias, domain dependency, and cross-platform generalization [7], [18].

More recently, research has focused on improving model transparency and interpretability through Explainable Artificial Intelligence (XAI) techniques. Methods such as Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) have been widely applied to explain predictions generated by complex deep learning models [26]. Several studies have also explored multimodal approaches, combining textual, visual, and social metadata to improve detection performance [15], [20]. Despite these advancements, challenges remain in evaluating the reliability and robustness of explanation methods, as many XAI techniques provide only approximate interpretations of model behavior [12]–[14].

Overall, the literature demonstrates a clear progression from traditional machine learning models to deep learning and explainable AI-based fake news detection systems. However,

the lack of standardized evaluation frameworks for explainability and the computational complexity of multimodal models indicate the need for further research to develop trustworthy and interpretable misinformation detection systems.

***Summary and Research Gap***

The literature indicates a clear evolution in fake news detection research, progressing from traditional machine learning approaches to deep learning models and more recently to explainable AI-based systems. While modern transformer models have significantly improved detection accuracy, the interpretability of these models remains a major challenge. Existing XAI techniques such as LIME, SHAP, and attention-based explanations provide valuable insights into model behavior; however, they often produce approximate explanations rather than fully accurate representations of model reasoning [12]–[14]. Additionally, current research lacks standardized methods for evaluating explanation reliability, robustness, and alignment with human understanding.

Therefore, there is a growing need for systematic evaluation of explainable AI methods in fake news detection systems. Investigating the effectiveness, robustness, and usability of existing XAI approaches is essential for developing trustworthy misinformation detection systems that can be applied in real-world decision-making environments. Literature Review of Fake News Detection flow can be depicted in the figure below- Fig.1 which illustrates the evolution of fake news detection approaches, highlighting the transition from traditional machine learning methods to deep learning, transformer-based architectures, and finally to explainable AI techniques. This progression reflects the increasing emphasis on both predictive accuracy and model interpretability.

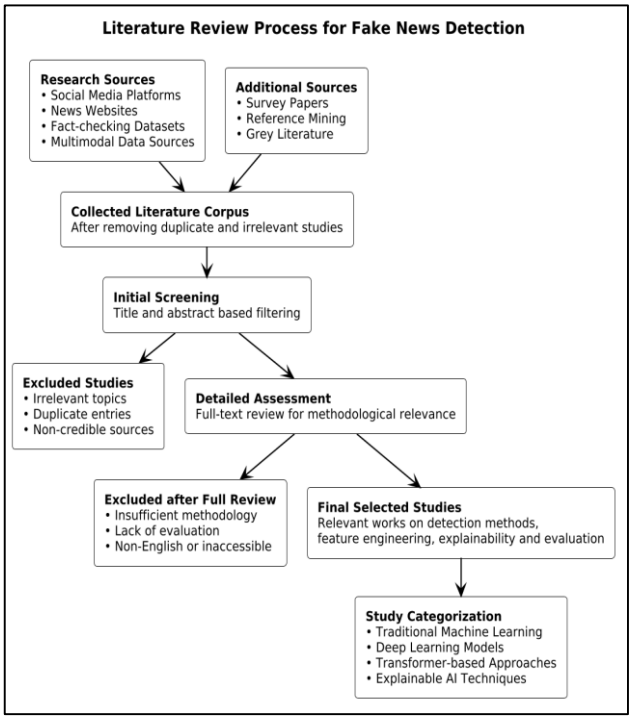


Fig. 1: Literature Evolution Diagram

The literature on fake news detection demonstrates a steady evolution from traditional machine learning approaches to advanced deep learning and transformer-based models. Recent studies have further incorporated explainable artificial intelligence techniques to enhance model transparency and trustworthiness. The following table summarizes key contributions in terms of detection methods, datasets, interpretability techniques, and performance evaluation.

**Table I: Explainable AI Techniques for Fake News Detection (2020–2025)**

| Year | Key Studies (References) | Dataset(s) Used | Model / Approach | XAI / Interpretability | Major Contributions | Key Limitations |
|---|---|---|---|---|---|---|
| **2017** | Shu et al. [1], Wang [2], Ruchansky et al. [3], Tacchini et al. [4], Rashkin et al. [17] | LIAR, Twitter, Facebook | Traditional ML (SVM, Naïve Bayes, Random Forest), early deep models | None | Established foundations of fake news detection; introduced benchmark datasets; early integration of textual and social features | Heavy feature engineering; limited interpretability |
| **2018** | Vosoughi et al. [5], Shu et al. [16] | Twitter datasets | Information diffusion analysis, social network analysis | None | Demonstrated that false information spreads faster than true | No automated explainable detection models |

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 2460-2465

| Year | Key Studies (References) | Dataset(s) Used | Model / Approach | XAI / Interpretability | Major Contributions | Key Limitations |
|---|---|---|---|---|---|---|
| | | | | | news; highlighted role of user behavior in misinformation spread | |
| **2020** | Zhou & Zafarani [7], Kumar & Shah [8], Shu et al. [6], de Beer & Matthee [18], Yang et al. [9] | FakeNewsNet, social media datasets | Deep learning and survey-based analysis | Conceptual interpretability | Provided comprehensive surveys and taxonomies of fake news detection; introduced datasets combining news content and social context | Lack of standardized datasets and evaluation metrics |
| **2021** | Hoy [30], Kim et al. | ISOT, Twitter, Weibo | CNN, CNN + BiLSTM | LIME, Attention mechanisms | Introduced explainability approaches in neural network models; improved interpretability of textual features | Attention mechanisms may not fully explain predictions |
| **2022** | Hu et al. [10], Mishra et al. [22], Castells [25] | FakeNewsNet, LIAR | Transformer models (BERT), Hybrid DL | SHAP, LIME, Layer-wise relevance propagation | Improved token-level interpretation of deep learning models; introduction of XAI methods for misinformation detection | High computational cost; scalability challenges |
| **2023** | Athira et al. [11], Ahmed [23], Bandi et al. [12], Gupta & Rao [26], Shu & Liu [29], Shen et al. [14] | Social media datasets, COVID-FN | Transformer models, deep neural networks | LIME, SHAP, Grad-CAM, Attention | Increased focus on explainable AI; analysis of model transparency and trustworthiness | Lack of evaluation frameworks for explanation quality |
| **2024** | Gongane [13], Jain et al. [15], Sultan et al. [24] | Multimodal datasets, social media | Multimodal DL (text + image + metadata) | SHAP, Attention | Integration of multimodal signals for improved detection accuracy; analysis of misinformation behavior | Computational complexity and limited real-time capability |
| **2025** | Moalla et al. [19], Nasser et al. [20], Alshuwaier [21], Geto et al. [28] | Multimodal and multilingual datasets | Hybrid deep learning models, CNN + Transformers | Feature visualization, explainability frameworks | Development of robust multimodal and multilingual fake news detection systems with explainability | Lack of standardized evaluation methods for XAI |

## IV. Research Objective

The primary objective of this review is to analyze and synthesize existing research on explainable artificial intelligence (XAI) methods applied to transformer-based fake news detection systems. The study focuses on understanding how different explainability techniques contribute to interpreting model predictions, evaluating their effectiveness, and identifying limitations in current approaches. Furthermore, the review aims to highlight key research gaps and outline future directions for developing more transparent, reliable, and human-interpretable misinformation detection systems.

## V. Contributions of This Work

This review presents a structured and interpretable approach to fake news detection by focusing not only on predictive performance but also on understanding how models arrive at their decisions. While recent research has largely emphasized improving accuracy using transformer-based architectures, the explainability aspect remains insufficiently explored. To address this gap, this study introduces a comprehensive framework that combines advanced detection models with explainable artificial intelligence (XAI) techniques.

The key contributions of this work are summarized as follows:

***(i) Comprehensive Literature Synthesis***

The study provides a detailed and structured review of fake news detection techniques, covering traditional machine learning, deep learning, transformer-based models, and explainable AI methods.

***(ii) Comparative Analysis of XAI Techniques***

A systematic comparison of widely used explainability methods, including LIME, SHAP, Integrated Gradients, and attention mechanisms, is presented in the context of fake news detection.

***(iii) Identification of Research Gaps***

The review highlights critical challenges such as the lack of standardized evaluation frameworks for explainability, instability of explanations, and limited alignment with human reasoning.

***(iv) Conceptual Framework for Explainable Detection***

A unified conceptual framework is proposed based on insights from existing literature, integrating transformer-based models with explainability techniques and evaluation strategies.

***(v) Future Research Directions***

The study outlines potential research directions to improve the reliability, robustness, and interpretability of fake news detection systems.

## Conceptual Framework Based on Literature Insights

The proposed methodology presents a framework for explainable fake news detection, which integrates transformer-based classification models with explainability techniques and evaluation mechanisms. The framework consists of several stages, including data processing, model prediction, explanation generation, and evaluation of explanations.

**1. Input Data**

The framework accepts news articles or social media posts as input. Each document $d$may include:

- textual content
- metadata (author, timestamp, source)
- associated images (optional)

These components provide contextual information useful for misinformation detection.

**2. Data Preprocessing**

Before model training, the textual content undergoes preprocessing steps to improve data quality. These steps include:

- removal of noise and irrelevant characters
- text normalization
- tokenization using transformer tokenizers
- conversion of text into numerical representations

This stage ensures that the input data is compatible with transformer-based models.

**3. Transformer-based Fake News Detection**

The processed textual data is then fed into transformer-based language models such as **BERT or RoBERTa**. These models capture contextual relationships between words and generate semantic embeddings that are used for classification.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 2460-2465

The model outputs a prediction label:

$$y \in \{0,1\}$$

where

- $y = 0$represents **real news**
- $y = 1$represents **fake news**

**4. Explainability Module**

To interpret the predictions generated by the model, the framework incorporates multiple explainability techniques, including:

- **LIME**, which explains predictions using locally interpretable models
- **SHAP**, which estimates feature contributions based on Shapley values
- **Integrated Gradients**, which measures feature importance using gradient-based attribution
- **Attention visualization**, which highlights important tokens in transformer models

These methods help identify the input features responsible for the classification decision.

**5. Evaluation of Explanations**

In addition to evaluating prediction performance, the framework assesses the quality of explanations using the following criteria:

- **Faithfulness:** alignment between explanations and actual model behavior
- **Consistency:** stability of explanations across similar inputs
- **Human interpretability:** ease with which explanations can be understood by users

This evaluation enables a more comprehensive understanding of model transparency.

**Conceptual Workflow of Explainable Fake News Detection-**

**Input:**

News document $d$

**Output:**

Prediction label $y$and explanation $E$

**Steps**

1. Collect dataset containing labeled news articles.
2. Perform preprocessing on textual data:
   - remove noise
   - tokenize text
   - convert text into transformer input format.
3. Load a transformer-based model (BERT / RoBERTa).
4. Train the model using labeled training data.
5. Provide document $d$as input to the trained model.
6. Generate classification prediction $y$.
7. Apply explainability techniques:
   - LIME
   - SHAP
   - Integrated Gradients
   - Attention analysis
8. Compute feature importance scores for input tokens.
9. Generate explanation $E$highlighting influential features.
10. Evaluate explanations based on:
    - faithfulness
    - consistency
    - interpretability.
11. Return prediction $y$and explanation $E$.

**The Framework Diagram can be depicted in the figure below as:**

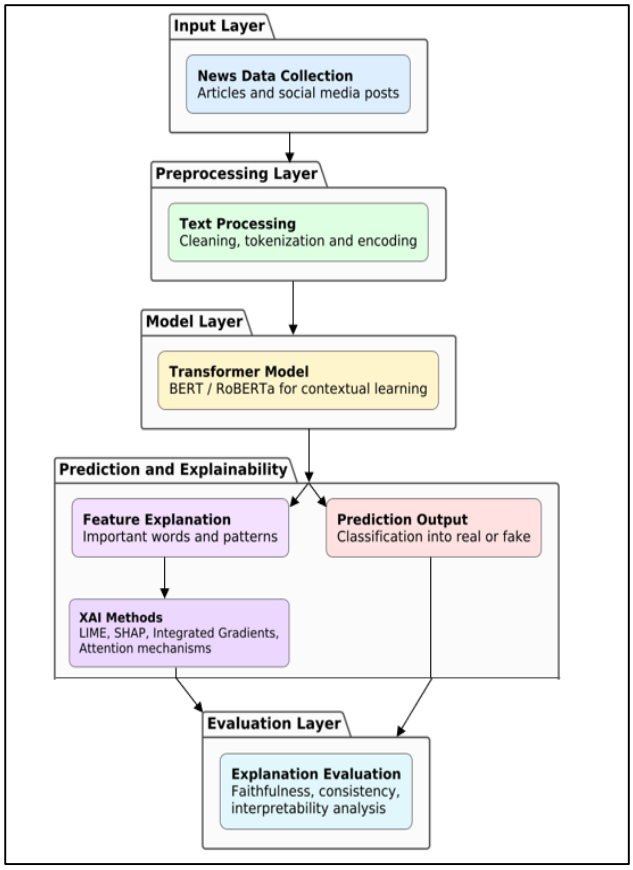


Fig. 2: Conceptual Framework

Figure 2 presents the conceptual framework for explainable fake news detection. The framework integrates data preprocessing, transformer-based classification, explainability modules, and evaluation mechanisms to ensure both accurate prediction and interpretable decision-making.

**VI. Evaluation Strategy of the Proposed Framework**

The proposed framework is evaluated using two complementary perspectives: (i) classification performance of transformer-based fake news detection models and (ii) quality of explanations generated by explainable AI techniques. This dual evaluation ensures that the system achieves both predictive accuracy and interpretability.

**Model Performance Evaluation**

This section evaluates the **classification model**.

**1. Common Evaluation Metrics**

Use standard ML metrics:

Example equation definitions:

Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision

$$Precision = \frac{TP}{TP + FP}$$

Recall

$$Recall = \frac{TP}{TP + FN}$$

F1-score

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

These are standard metrics used in **fake news detection studies**.

**Explainability Evaluation**

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 2460-2465

As our framework is about **explainable fake news detection**, so the evaluation quality metrics can be given as in the table below:

**Table II. Explanation Quality Metrics**

| Metric | Description |
|---|---|
| Faithfulness | Explanation matches model behaviour. Measures how accurately the explanation reflects the decision-making process of the model. |
| Consistency | Similar inputs produce similar explanations |
| Interpretability | Humans can easily understand explanations |

**2. Comparative Analysis of Explainability Techniques**

The explainability module of the proposed framework integrates multiple XAI techniques. These techniques differ in terms of explanation scope, model dependency, interpretability level, computational complexity, and theoretical foundation. A comparative analysis of widely used XAI techniques is presented in Table III as given below:

Table III: The explainability module of the proposed framework

| Method | Explanation Scope | Model Dependency | Explanation Mechanism | Interpretability Level | Computational Cost | Advantages | Limitations |
|---|---|---|---|---|---|---|---|
| **LIME** | Local | Model-agnostic | Builds a simple surrogate model around a specific prediction | High | Medium | Easy to understand explanations, works with any model | Instability due to random perturbations |
| **SHAP** | Local + Global | Model-agnostic | Uses Shapley values from cooperative game theory to estimate feature contribution | High | High | Theoretically grounded, consistent explanations | Computationally expensive for large models |
| **Integrated Gradients** | Local | Model-specific | Computes gradient-based feature attribution along a path from baseline input | Medium | Low–Medium | Satisfies axioms such as sensitivity and implementation invariance | Requires differentiable models |
| **Attention Visualization** | Global | Model-specific | Uses attention weights to highlight important tokens in transformer models | Medium | Low | Provides intuitive visual explanation for NLP tasks | Attention weights may not always correspond to true explanations |

The framework allows comparative analysis of different explainability techniques to determine which methods provide more reliable and interpretable explanations for fake news detection models. The values in Table X are determined based on established properties of XAI methods reported in prior research literature.

**3. Experimental Setup**

Even if implementation is future work, we should still describe **how experiments will be conducted**

**Experimental Setup**

Include:

**Dataset**

Example datasets used in fake news research:

- FakeNewsNet
- LIAR dataset
- ISOT Fake News dataset

**Model**

Transformer models used:

- BERT
- RoBERTa

**Tools**

Implementation environment:

- Python
- PyTorch
- HuggingFace Transformers
- SHAP / LIME libraries

**4. Expected Results / Validation Strategy**

Even if experiments are not yet done, still we can explain **what results will demonstrate**:

The experimental evaluation will demonstrate the effectiveness of the proposed framework in generating interpretable predictions while maintaining high classification performance. Comparative analysis of different explainability techniques will provide insights into their reliability and usefulness for fake news detection tasks.

**Benefits of our framework**.

• improves transparency in fake news detection
• integrates multiple explainability techniques
• provides systematic evaluation of explanations
• helps researchers analyze model behaviour

## VII. Critical Analysis of Existing Approaches

Although significant progress has been made in fake news detection using deep learning and transformer-based models, several limitations remain. Traditional machine learning approaches rely heavily on feature engineering and fail to capture complex contextual relationships. While transformer models such as BERT and RoBERTa achieve high accuracy, their black-box nature limits interpretability. Explainable AI techniques such as LIME and SHAP provide useful insights into model predictions; however, these methods often generate approximate explanations that may not accurately reflect the true decision-making process of the model. Additionally, explanation methods can be unstable, producing different interpretations for similar inputs. Attention-based explanations, although intuitive, do not always correlate with actual feature importance. Furthermore, there is a lack of standardized evaluation metrics for assessing explanation quality, making it difficult to compare different XAI techniques. These limitations highlight the need for more robust, reliable, and human-aligned explainability methods in fake news detection systems.

## VIII. Limitations and Future Work

Despite providing a comprehensive review, this study has certain limitations. The analysis is restricted to published literature and may not include unpublished or emerging research. Additionally, the comparison of explainability techniques is based on reported findings rather than direct experimental evaluation. Future research should focus on developing standardized evaluation frameworks for explainable AI methods, improving the robustness and stability of explanations, and designing techniques that better align with human reasoning. Moreover, integrating multimodal data and real-time explainability in fake news detection systems remains an open research challenge.

## REFERENCES

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

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 2460-2465

[2] W. Y. Wang, " 'Liar, Liar Pants on Fire': A new benchmark dataset for fake news detection," in *Proc. 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017, pp. 422–426.
[3] N. Ruchansky, S. Seo, and Y. Liu, "CSI: A hybrid deep model for fake news detection," in *Proc. ACM Conference on Information and Knowledge Management (CIKM)*, 2017, pp. 797–806.
[4] E. Tacchini, G. Ballarin, M. L. Della Vedova, S. Moret, and L. de Alfaro, "Some like it hoax: Automated fake news detection in social networks," *arXiv preprint arXiv:1704.07506*, 2017.
[5] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.
[6] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media," *Big Data*, vol. 8, no. 3, pp. 171–188, 2020.
[7] J. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Computing Surveys*, vol. 53, no. 5, pp. 1–40, 2020.
[8] S. Kumar and N. Shah, "False information on web and social media: A survey," *Communications of the ACM*, vol. 63, no. 2, pp. 60–68, 2020.
[9] Z. Yang, J. Lu, and C. Meeng, "A survey on fake news detection: Fundamental theories, detection approaches and challenges," *IEEE Transactions on Knowledge and Data Engineering*, 2020.
[10] L. Hu, S. Man, and D. Zhao, "Deep learning for fake news detection: A comprehensive survey," *IEEE Access*, vol. 10, pp. 113–130, 2022.
[11] A. B. Athira, M. S. Reddy, and P. V. Reddy, "Explainable artificial intelligence for fake news detection: A systematic review," *Expert Systems with Applications*, vol. 225, 2023.
[12] A. Bandi, A. Dighe, and S. Khandekar, "Explainable AI methods for misinformation detection: A systematic review," *Engineering Applications of Artificial Intelligence*, vol. 123, 2023.
[13] V. Gongane, "A survey of explainable AI techniques for detection of fake news and hate speech," *Journal of Computational Social Science*, 2024.

[14] Y. Shen, X. Li, and H. Wang, "Fake news detection on social networks: A survey," *Applied Sciences*, vol. 13, no. 1, 2023.
[15] R. Jain, V. K. Gupta, and P. Sinha, "Multimodal fake news detection: A survey," *IEEE Access*, vol. 12, pp. 45500–45520, 2024.
[16] K. Shu, S. Wang, and H. Liu, "Understanding user profiles on social media for fake news detection," in *Proc. IEEE International Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2018.
[17] R. Rashkin, E. Choi, J. Jang, S. Volkova, and Y. Choi, "Truth of varying shades: Analyzing language in fake news and political fact-checking," in *Proc. EMNLP*, 2017.
[18] D. de Beer and A. Matthee, "Approaches to identify fake news: A systematic literature review," in *Proc. International Conference on Information Communication Technologies*, 2020.
[19] H. Moalla, A. Ben, and S. Kacem, "Dual deep learning models for fake news detection: A survey," *Informatica*, vol. 49, 2025.
[20] M. Nasser, H. El-Sayed, and M. H. Ismail, "Deep learning for multimodal fake news detection: A systematic review," *Information Fusion*, vol. 98, 2025.
[21] F. A. Alshuwaier, "Fake news detection using machine learning and deep learning: A comprehensive review," *Computers*, vol. 14, 2025.
[22] P. Mishra, R. Singh, and A. Verma, "Meta-analysis of automated fake news detection studies," *Applied Sciences*, vol. 12, no. 2, 2022.
[23] S. Tejedor and S. Cervi, "Fact-checking and fake news: A systematic review," *Journalism and Media*, vol. 4, no. 1, 2024.
[24] M. Sultan et al., "Susceptibility to online misinformation: A systematic meta-analysis," *Computers in Human Behavior*, vol. 146, 2024.
[25] D. Castells, "Explainability in neural attention models for fake news detection," *Neural Computing and Applications*, 2022.
[26] S. Gupta and P. Rao, "Interpretable fake news detection using LIME and SHAP," *Journal of Network and Computer Applications*, 2023.
[27] J. Ahmed, "Algorithms and datasets for fake news detection: A systematic review," *Applied Intelligence*, vol. 53, no. 6, 2023.
[28] A. Geto et al., "Multimodal Amharic fake news detection using CNNs and dataset creation," *Scientific Reports*, 2025.
[29] K. Shu and H. Liu, "Explainability and trust in fake news detection systems: Challenges and perspectives," *ACM Computing Surveys*, 2023.
[30] N. Hoy, "Detection of fake news articles: A systematic review," *Future Internet*, vol. 13, 2021.