

AI-Based Fake News Detection: A Review

R. Baby, S. Harini Anandhi, M. Jackulin, S. Kalpana, A. Raja Dharsan, A. Jegatheswaran

Department of Information Technology, J.P. College of Engineering, Ayikudy, Tamil Nadu, India

Abstract- The rapid growth of digital platforms and social media has transformed the way people create, access, and share information. Although online platforms provide fast access to news and public information, they have also contributed to the rapid spread of misleading and fabricated content. Fake news can influence public opinion, create confusion, damage reputations, and reduce trust in legitimate information sources. Traditional fact-checking methods often require significant human effort and may not be sufficient to process the large volume of information generated online. Artificial Intelligence (AI), Machine Learning (ML), Natural Language Processing (NLP), and Deep Learning techniques have therefore emerged as important approaches for automated fake news detection. This review examines the major AI-based techniques used for identifying fake news, with particular emphasis on text preprocessing, feature extraction, machine learning algorithms, deep learning models, and natural language processing. Traditional machine learning approaches such as Naive Bayes, Logistic Regression, Support Vector Machines, and Random Forest have been widely explored, while deep learning architectures such as Convolutional Neural Networks, Recurrent Neural Networks, Long Short-Term Memory networks, and transformer-based models provide more advanced methods for capturing contextual and linguistic patterns. The review also discusses commonly used datasets, evaluation metrics, challenges related to misinformation, dataset quality, language diversity, evolving news patterns, and model generalization. Finally, future research directions including multilingual detection, explainable AI, multimodal analysis, and real-time misinformation detection are discussed. This review highlights the potential of AI-based approaches to support automated and scalable fake news detection systems while emphasizing the need for reliable datasets, continuous model updating, and human fact-checking.

Keywords: Fake News Detection, Artificial Intelligence, Machine Learning, Natural Language Processing, Deep Learning, Misinformation, Social Media, Text Classification, Neural Networks, Transformer Models.

I. INTRODUCTION

The Internet and social media have become major sources of news and information for people around the world. News can now be created, published, and shared within seconds through social networking platforms, online news portals, messaging applications, and other digital media. This rapid exchange of information provides significant benefits, but it also creates opportunities for false, misleading, and fabricated information to spread widely. Fake news refers broadly to fabricated or misleading information presented in a news-like form with the potential to deceive readers. It may be intentionally created to influence public opinion, generate financial benefits, promote particular viewpoints, or attract online attention.

The large-scale distribution of misleading content can have social, economic, and political consequences and can reduce public trust in reliable

information sources. The problem becomes more difficult because users often encounter a large amount of information from different sources and may not have sufficient time or resources to verify every claim. Manual fact-checking requires human expertise and can be time-consuming, especially when large volumes of online content are produced continuously. Automated approaches can therefore assist human fact-checkers by screening large collections of news articles and identifying content that requires further verification.

Artificial Intelligence has provided several techniques for automated misinformation detection. Machine Learning algorithms can learn patterns from labelled news datasets and classify articles according to learned characteristics. Natural Language Processing enables systems to analyse textual information, including word usage, sentence structure, linguistic patterns, and semantic relationships. Traditional approaches commonly use

algorithms such as Naive Bayes, Logistic Regression, Support Vector Machines, and Random Forest for text classification. These methods generally depend on feature representations such as term frequency, TF-IDF, and linguistic characteristics. Although these approaches can provide effective results on suitable datasets, their performance may be affected by changes in language, writing style, topic distribution, and dataset quality.

Deep Learning has expanded the capabilities of automated fake news detection by learning complex representations directly from text. Convolutional Neural Networks can identify local textual patterns, while Recurrent Neural Networks and Long Short-Term Memory networks can model sequential dependencies. More recently, transformer-based language models have provided powerful contextual representations that can be applied to text classification tasks. However, fake news detection remains challenging because misinformation changes rapidly and may resemble legitimate news in writing style. In addition, the same information can appear in multiple languages, formats, and media types. A system trained on one dataset or specific domain may not perform equally well when applied to new topics or platforms. This paper presents a review of AI-based approaches for fake news detection. It discusses traditional machine learning, natural language processing, deep learning, datasets, evaluation measures, major challenges, and future research directions. The objective is to provide an overview of existing techniques and identify important areas for developing more reliable and adaptable fake news detection systems.

II. LITERATURE REVIEW

Research on automated fake news detection has developed significantly with the growth of social media and online information platforms. Early approaches mainly focused on manually designed linguistic features and traditional machine learning algorithms. These approaches typically transformed news text into numerical representations and used classification algorithms to distinguish between reliable and unreliable content. Natural Language Processing plays an important role in fake news

detection because most news articles contain textual information that can be analysed computationally. Common preprocessing operations include converting text to a consistent format, removing unnecessary symbols, tokenization, stop-word removal, stemming or lemmatization, and feature extraction. TF-IDF and n-gram representations have traditionally been used to represent textual characteristics. Machine learning algorithms such as Naive Bayes, Logistic Regression, Support Vector Machines, and Random Forest have been widely applied to text classification. These algorithms can learn relationships between extracted features and predefined news labels.

Their advantages include relatively simple implementation and lower computational requirements compared with large deep learning models. Deep learning methods have subsequently been introduced to learn more complex textual representations. Recurrent Neural Networks are designed to process sequential information, while LSTM networks can retain information over longer sequences. CNN architectures have also been applied to text classification by learning local patterns from word representations. The development of attention mechanisms and transformer architectures has significantly influenced natural language processing. Transformer-based models can capture contextual relationships between words and have demonstrated strong performance across a wide range of language tasks. Models based on transformer architectures can therefore provide more advanced representations for misinformation and fake news classification.

Another important research direction is the use of social context. Instead of relying only on the textual content of an article, some approaches analyse information such as user interactions, propagation patterns, comments, sharing behaviour, and network structures. These additional signals can provide useful information because misleading information may spread through social networks differently from reliable news. Multimodal fake news detection is also receiving increasing attention. Modern misinformation may contain combinations of text, images, videos, and other media. A system that

analyses only text may fail when visual information contradicts the accompanying statement. Multimodal approaches attempt to combine information from multiple sources to improve detection. Despite significant progress, several research gaps remain. Many datasets are limited to specific topics, languages, or platforms. Models may perform well on benchmark datasets but experience reduced performance when applied to real-world information. Dataset imbalance, evolving misinformation strategies, multilingual content, and lack of explainability remain important challenges. This review therefore focuses on the development of AI-based fake news detection methods and discusses how machine learning, deep learning, NLP, and emerging transformer-based approaches can contribute to more reliable misinformation detection.

III. METHODOLOGY

As a review paper, the methodology focuses on identifying, organizing, and analysing existing research related to AI-based fake news detection. The review considers major AI techniques, textual processing methods, datasets, evaluation metrics, and research challenges.

The review process consists of five major stages:

1. Identification of relevant research
2. Classification of detection approaches
3. Analysis of datasets and features
4. Comparison of AI techniques
5. Identification of research gaps and future directions

Research Identification

Relevant research studies related to fake news detection, misinformation detection, machine learning, deep learning, natural language processing, and social media analysis are considered. Research papers, review articles, conference publications, and other reliable academic sources can be examined to understand the development of automated detection techniques.

Text Preprocessing

Most text-based fake news detection systems require preprocessing before classification. Common preprocessing operations include:

- Text normalization
- Tokenization
- Removal of unnecessary symbols
- Stop-word removal
- Stemming or lemmatization
- Handling of punctuation
- Feature representation

These operations reduce unnecessary variation in the input text and prepare the content for machine learning or deep learning models.

Feature Extraction

Feature extraction converts textual information into a numerical representation that can be processed by AI models.

Common approaches include:

- Bag-of-Words
- TF-IDF
- N-grams
- Word embeddings
- Contextual embeddings

Traditional representations such as TF-IDF focus on word frequency, whereas embedding-based approaches represent words or sentences in a continuous vector space.

Machine Learning Approaches

Several traditional machine learning algorithms can be applied to fake news classification.

1) Naive Bayes

Naive Bayes is a probabilistic classification algorithm that can be effective for text classification. It estimates the probability of a news article belonging to different classes based on the observed textual features.

2) Logistic Regression

Logistic Regression can be used for binary or multi-class classification. It estimates the probability of a particular class based on the extracted features.

3) Support Vector Machine

Support Vector Machine attempts to identify a decision boundary that separates different classes. It has been widely used for high-dimensional text classification.

4) Random Forest

Random Forest combines multiple decision trees to produce a classification result. It can handle different feature types and may provide useful performance for structured feature representations.

Deep Learning Approaches

Deep learning methods can learn complex representations from textual data.

1) Convolutional Neural Network

CNNs can identify local patterns in text and have been applied to sentence and document classification.

2) Recurrent Neural Network

RNNs process sequential information and can model relationships between words across a text sequence.

3) Long Short-Term Memory

LSTM networks are a type of recurrent architecture designed to retain relevant information over longer sequences. They have been widely used for sequential language processing.

4) Transformer-Based Models

Transformer architectures use attention mechanisms to model relationships between tokens and generate contextual representations. Transformer-based language models have become important

approaches for modern NLP applications, including text classification and misinformation analysis.

Social Media-Based Analysis

Fake news detection can also incorporate social context. Useful information may include:

- User interaction patterns
- Sharing behaviour
- Comments
- Propagation structure
- Source characteristics
- Network relationships

Combining textual and social information can provide additional signals for detecting misinformation.

Evaluation Metrics

Fake news classification systems are commonly evaluated using:

- Accuracy
- Precision
- Recall
- F1-Score
- Confusion Matrix
- ROC-AUC, where appropriate

Precision measures the proportion of predicted positive instances that are actually positive, while recall measures the ability to identify positive instances. F1-score provides a combined measure of precision and recall.

IV. COMPARATIVE ANALYSIS

The major AI approaches used for fake news detection can be compared according to their representation methods, advantages, and limitations.

Table 1: Comparison of AI-Based Fake News Detection Approaches

Approach	Common Representation	Major Advantage	Major Limitation
Naive Bayes	TF-IDF / Bag-of-Words	Simple and computationally efficient	Limited contextual understanding
Logistic Regression	TF-IDF / N-grams	Effective baseline for text classification	Depends strongly on feature quality
SVM	TF-IDF / N-grams	Effective for high-dimensional text	Training and tuning can become difficult for large datasets
Random Forest	Statistical / linguistic features	Handles diverse features	May not capture deep semantic relationships
CNN	Word embeddings	Learns local textual patterns	Limited long-range contextual modelling

Approach	Common Representation	Major Advantage	Major Limitation
LSTM	Word embeddings	Models sequential dependencies	Higher computational cost
Transformer Models	Contextual embeddings	Strong contextual representation	Higher computational and resource requirements

The comparison indicates that traditional machine learning models provide relatively simple and efficient solutions, while deep learning and transformer-based models offer more advanced representation capabilities. However, the appropriate method depends on dataset size, language, computational resources, and application requirements.

V. DISCUSSION

AI-based fake news detection has become an important research area because the volume and speed of online information continue to increase. Traditional manual verification cannot always process the large amount of content generated across social media and digital news platforms. Automated AI systems can assist by identifying suspicious content and prioritizing information for further human verification.

Traditional machine learning approaches remain useful because they are relatively simple and can perform effectively when suitable textual features are available. TF-IDF-based representations combined with algorithms such as Logistic Regression, Naive Bayes, or SVM can provide strong baseline systems. However, these approaches may have difficulty understanding deeper semantic relationships and contextual meanings.

Deep learning approaches address some of these limitations by learning representations from data. CNN-based models can capture local textual patterns, while LSTM networks can model sequential information. However, deep learning models generally require larger datasets and greater computational resources.

Transformer-based models provide more advanced contextual understanding and have become

important in NLP research. Their attention mechanisms allow relationships between different parts of a text to be modelled effectively. Despite their advantages, transformer models can require significant computational resources and may still be vulnerable to changes in domain, language, and misinformation strategies.

Another important issue is dataset reliability. A model can only learn useful patterns when the training data contain accurate labels and sufficient diversity. If datasets contain biases, duplicate content, limited topics, or unreliable labels, the resulting model may not generalize well to real-world information.

Language diversity is another major challenge. Fake news may be produced in multiple languages and may contain code-mixed or informal language. Systems developed primarily using English-language datasets may not perform equally well for regional or multilingual content.

Explainability is also important for misinformation detection. Users and fact-checkers may require information about why an AI system considers a particular article suspicious. Explainable AI techniques can provide insights into important words, sentences, or features contributing to a prediction. This can improve transparency and support human verification.

Overall, AI can provide valuable assistance in identifying potentially misleading content, but automated classification should not be treated as a replacement for professional fact-checking. A reliable system should combine automated detection with source verification, human expertise, and continuous evaluation.

VI. FUTURE WORK

Future research can focus on improving the accuracy, generalization, transparency, and real-world applicability of fake news detection systems.

1. Multilingual Fake News Detection

Future systems can support multiple languages and regional language varieties to improve misinformation detection across diverse populations.

2. Explainable AI

Explainable AI techniques can be integrated to identify the words, sentences, sources, or other features influencing the model's decision.

3. Multimodal Fake News Detection

Future research can combine text, images, videos, and other media signals to detect misinformation that cannot be identified through text alone.

4. Real-Time Detection

AI systems can be optimized to analyse rapidly spreading online content and identify potentially misleading information in near real time.

5. Cross-Domain Generalization

Future models should be evaluated across different topics, platforms, languages, and news sources to determine whether they generalize beyond benchmark datasets.

6. Human-AI Collaboration

AI-based systems can be designed as decision-support tools for professional fact-checkers rather than completely replacing human verification.

7. Advanced Language Models

Future research can investigate modern transformer architectures and efficient language models to improve contextual understanding while reducing computational requirements.

8. Dataset Development

Large, diverse, accurately labelled, and regularly updated datasets can improve the reliability of fake

news detection models and help address changing misinformation patterns.

VII. CONCLUSION

Fake news has become a major challenge in the digital information environment because misleading content can be created and distributed rapidly through online platforms. Artificial Intelligence provides several approaches for addressing this problem by automatically analysing textual, linguistic, social, and multimodal characteristics of online information.

This review examined traditional machine learning algorithms, natural language processing techniques, deep learning architectures, and transformer-based approaches for fake news detection. Traditional methods such as Naive Bayes, Logistic Regression, SVM, and Random Forest provide efficient and useful baseline solutions, while CNN, RNN, and LSTM architectures can learn more complex patterns from textual information. Transformer-based models provide advanced contextual representations and have become important approaches for modern NLP-based misinformation detection.

Despite the progress of AI-based methods, several challenges remain, including dataset quality, class imbalance, language diversity, changing misinformation strategies, cross-domain generalization, and model interpretability. A system that performs well on a specific benchmark dataset may not necessarily maintain the same performance when exposed to new topics, platforms, languages, or forms of misinformation.

Future systems should therefore focus on multilingual and multimodal analysis, explainable AI, real-time detection, robust dataset development, and human-AI collaboration. Automated detection should be considered a supportive mechanism for identifying potentially misleading information rather than a complete replacement for human fact-checking.

Overall, AI-based fake news detection represents a promising research area for improving the reliability

of digital information systems. Continued development of robust, transparent, and adaptable AI methods can contribute to reducing the impact of misinformation and supporting more trustworthy online information environments.

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.
2. 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.
3. R. Oshikawa, J. Qian, and W. Y. Wang, "A Survey on Natural Language Processing for Fake News Detection," *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 6086–6093, 2020.
4. V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, "Automatic Detection of Fake News," *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3391–3401, 2018.
5. N. Ruchansky, S. Seo, and Y. Liu, "CSI: A Hybrid Deep Model for Fake News Detection," *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 797–806, 2017.
6. Y. Wang, M. McKee, A. Torbica, and D. Stuckler, "Systematic Literature Review on the Spread of Health-Related Misinformation on Social Media," *Social Science & Medicine*, vol. 240, 2019.
7. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.
8. A. Vaswani et al., "Attention Is All You Need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
9. T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *arXiv preprint arXiv:1301.3781*, 2013.
10. Y. Goldberg, "A Primer on Neural Network Models for Natural Language Processing," *Journal of Artificial Intelligence Research*, vol. 57, pp. 345–420, 2016.
11. S. Vosoughi, D. Roy, and S. Aral, "The Spread of True and False News Online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.
12. X. Zhou and R. Zafarani, "A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities," *ACM Computing Surveys*, vol. 53, no. 5, Article 109, 2020.
13. X. Zhou, A. Jain, V. V. Phoha, and R. Zafarani, "Fake News Early Detection: A Theory-Driven Model," *Digital Threats: Research and Practice*, vol. 1, no. 4, 2020.
14. K. Shu, S. Wang, and H. Liu, "Beyond News Contents: The Role of Social Context for Fake News Detection," *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pp. 312–320, 2019.
15. H. Allcott and M. Gentzkow, "Social Media and Fake News in the 2016 Election," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–236, 2017.