International Research Journal of Education and Technology

Peer Reviewed Journal, **ISSN: 2581-7795**

# AI-Based Fake News Detection Using Transformers

Anuja Beatrice B[1] , Sruthi S[2]

Head and Associate professor[1], BCA Student[2]

Sri Krishna Arts and Science College Coimbatore[1],[2]

## Abstract

The proliferation of misinformation on digital platforms has emerged as a significant societal challenge, as individuals increasingly rely on online news sources for information and decision-making. The rapid spread of fabricated or misleading content through social media and online publishing platforms has made traditional manual fact-checking methods and rule-based systems insufficient for effectively combating fake news. To address this issue, this report presents an AI-based fake news detection framework that integrates advanced Natural Language Processing (NLP), deep learning techniques, and graph-based credibility analysis. The proposed system utilizes transformer architectures such as BERT, RoBERTa, and XLNet to capture deep contextual relationships within textual data and accurately distinguish between authentic and misleading information. In addition, hybrid neural components and graph-based models are incorporated to analyze both linguistic patterns and relationships between news sources. Explainable Artificial Intelligence (XAI) techniques are employed to improve model transparency by identifying key textual indicators that influence classification decisions. Furthermore, adversarial robustness mechanisms are integrated to enhance the system's ability to resist manipulation attempts and misleading content variations. Experimental evaluation using benchmark datasets such as LIAR and FakeNewsNet demonstrates that the integration of semantic, stylistic, and discourse-level features significantly improves detection accuracy and reliability. The results highlight the effectiveness of combining transformer-based models with explainability and robustness strategies to build scalable and trustworthy misinformation detection systems.

## Introduction

The tremendous growth of online news consumption has profoundly impacted the way information is generated, disseminated, and consumed. The advent of online platforms, social media sites, and news publishing tools has enabled the rapid spreading of news among the global population within a matter of seconds. Although this has helped the proliferation of information sharing, it has also led to the spreading of fake news and misinformation, which is a major threat to the health, political, and social stability of the country, along with the results of the democratic

elections, where people rely on online news to form opinions and make critical decisions in their lives.

The proliferation of misinformation is beyond the control of manual verification techniques, as the rate at which fake news is spreading online is beyond the capabilities of manual verification techniques. Artificial Intelligence (AI) techniques, especially Natural Language Processing (NLP) and machine learning algorithms, have been proposed to solve this critical problem.

The latest developments in the field of Deep Learning (DL) technologies, such as the use of transformer models, have shown considerable promise in the ability of AI systems to evaluate complex linguistic patterns in the context of textual information. Techniques such as BERT, RoBERTa, and XLNet utilize attention models to evaluate contextual information in the detection of semantic cues that are often used in the development of fake news content. These models are more efficient in the detection of fake news when compared to other models that are used in the evaluation of textual information patterns.[1]

The study proposes the development of a hybrid model for the detection of fake news that utilizes the capabilities of transformer models, graph learning models, and Explainable Artificial Intelligence (XAI) models. The proposed model is more efficient in the evaluation of the relationships that exist in the news sources that are used in the development of fake news content. The use of these models in the development of the AI system is more reliable in the evaluation of the detection of fake news in the context of the digital information environment.

**Problem Statement**

The increasing popularity and growth of digital media and social networking sites have led to the spread of misinformation and fake news content on the internet. Misinformation and fake news could affect the masses, the political scenario, and the overall social scenario, and in some instances, it could affect the health of the people. Though researchers and developers have tried to develop systems to detect fake news using artificial intelligence techniques, there are a number of critical challenges that are yet to be addressed in the current systems to develop a reliable and trustworthy misinformation detection system.

**Scalability:**

The huge amount of information on the internet is a critical issue in the detection of misinformation and fake news. There are millions of news articles, social media posts, and other multimedia content pieces posted on the internet every day. Considering the huge amount of information, it is practically difficult to rely on a manual fact-checking system. Thus, it is essential to develop a system that could efficiently handle the huge amount of information on the internet with a high level of accuracy.

**• Contextual Complexity:**

Fake news is designed in a way that it is credible and persuasive. Most fake news articles are written in a sophisticated style, and the context in which the information is presented is misleading. In most fake news, there is a mixture of truths and falsehoods. Traditional linguistic approaches, which are based on simple linguistic features like word frequency, keyword matching, and sentiment analysis, are not able to detect the complex semantic relationships in fake news. Also, the

narratives of misinformation are dynamic and change over time.[5]

**• Adversarial Vulnerability:**

Attackers are always finding new ways to evade detection systems. Some of the adversarial attacks include synonym replacement, text paraphrasing, sentence rearrangement, and making slight changes to the wording. These attacks have the potential to mislead machine learning systems, which rely heavily on text features. Moreover, the problem of misinformation can arise if several sources publish the same false information, but with slight variations.

**• Interpretability:**

Even though deep learning models such as neural networks and the transformer model have shown high accuracy in the detection of fake news, most of these models are complex systems that function in a "black-box" manner. This means that the decision-making mechanism of these models is not easily understandable or explainable by the human mind.

This could pose a problem in terms of the level of trust that journalists, fact-checkers, and other users place in these models when it comes to the detection of fake news. This is because it is not possible to explain the reason behind the detection of the news as fake or real.

**• Source Credibility and Propagation Analysis:**

One other problem that arises in the detection of fake news is the assessment of the credibility of the information source. Moreover, the propagation analysis of the fake news is also another problem. Fake news is usually propagated through the coordinated actions of several accounts or through an unreliable information source.

Generally, the text-based models that have been developed for the detection of fake news usually consider the content of the news article.

However, other important factors, such as the credibility of the information source, the propagation of the information, etc., have also been ignored. Therefore, this is another problem that arises in the detection of fake news.

Thus, the above challenges have emphasized the need for the development of sophisticated fake news detection systems that can effectively address the challenges through the incorporation of strong contextual analysis, adversarial resistance, credibility analysis, etc.[5]

**Related Work**

The area of fake news detection has witnessed tremendous growth over the past decade, from conventional linguistic and statistical approaches to highly sophisticated deep learning techniques. The conventional approaches to fake news detection were based on classical Natural Language Processing (NLP) techniques, including Term Frequency-Inverse Document Frequency (TF-IDF), n-gram models, and sentiment analysis, to name a few, to identify the patterns associated with fake news content.

The conventional approaches were based on manual feature extraction techniques combined with conventional machine learning algorithms, including Support Vector Machines (SVM), Naive Bayes, Random Forest, and many more. Although the conventional approaches were able to achieve reasonable results in detecting fake news content, the results were limited due to the conventional nature of the techniques, which were unable to understand the deep semantics associated with the content of the news articles.[4]

With the advent of deep learning techniques, researchers were able to adopt various deep learning techniques to achieve high accuracy results in fake news detection. The conventional deep learning techniques, including Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) Networks, were able to achieve high accuracy results in fake news detection, especially for long content associated with news articles.

**Methodology**

The proposed methodology for the detection of fake news is based on an efficient multi-stage pipeline that utilizes the latest advancements in the field of Natural Language Processing, Deep Learning, and Explainable Artificial Intelligence. Each step of the proposed methodology aims at making the detection of fake news more reliable, explainable, and robust. The entire process of the proposed methodology can be divided into several steps: data collection, data preprocessing, feature extraction, building the hybrid model, and the incorporation of robustness and explainability.[9]

**1. Data Collection:**

The first step of the proposed methodology for the detection of fake news is based on the collection of reliable datasets that contain both genuine and fake news. For the purpose of the proposed approach, the LIAR dataset, FakeNewsNet, and ISOT Fake News Dataset have been utilized.

These datasets contain thousands of news statements and articles that have been collected from various sources, including social media, fact-checking websites, online news portals, etc. These datasets contain useful information regarding the news text, credibility, context, etc. Using multiple datasets for the proposed approach helps in making the detection system more robust.

**2. Data Preprocessing:**

Before the actual model is trained, the collected data in the form of text needs to be cleaned and prepared for the analysis. This step is essential to ensure the quality of the data and is often referred to as data preprocessing. There are various steps involved in the preprocessing of the collected data. The first step is tokenization, where the collected data is split into small pieces of information, often in the form of words or phrases. The second step is the removal of stop words, where common words such as “the,” “is,” “and,” etc., are eliminated from the data, as this information is often irrelevant for the classification task.

The third step is the application of the Lemmatizer, where the collected data is converted into its base form, ensuring that the model is able to classify the data based on the root word. Finally, the Named Entity Recognition (NER) step is applied to the collected data, where relevant entities such as people, organizations, locations, and events mentioned in the news articles are identified.

**3. Feature Extraction:**

Following the preprocessed step, the text data is transformed into numerical data, which can be easily processed by machine learning and deep learning algorithms. Various feature extraction techniques are used to extract different features of the text data.

Some traditional embedding techniques, such as Word2Vec and GloVe, are used to transform the text data into numerical vectors, which can be easily processed by machine learning algorithms. These embedding techniques transform the text data into a numerical representation, showing the semantic relationship between the text data.[8]

Furthermore, some transformer-based techniques, such as BERT and RoBERTa, are used to transform the text data into contextual representations, which can be easily processed by machine learning algorithms. These contextual representations help the system understand the semantic meaning of the text data, which is usually used to create fake news.

**4. Hybrid Modeling:**

For the purpose of attaining higher accuracy in the detection process, the system incorporates the idea of hybrid modeling, in which the system will use a combination of different features. In this context, the system will incorporate lexical features, semantic features, and discourse features, which will help in the overall understanding of the news content. In this context, lexical features will help in understanding the pattern and style in which the words are used, whereas semantic features will help in understanding the context of the news content with the help of the transformer model.

In addition, discourse features will help in understanding the logical consistency of the news content, which may contain contradictions, exaggerations, and emotions. Furthermore, the system will also incorporate sentiment polarity, which is normally used in misleading news.

**System Architecture**

The proposed system will use a multi-dimensional architecture, ensuring the precision and reliability of fake news detection.

**Transformer Core**

The proposed system will use the Transformer-based models, such as BERT, RoBERTa, and XLNet, as the core of the system. The main idea behind using these models is their ability to understand the context of the sentence using self-attention techniques.

Unlike other models, such as RNN and LSTM, the Transformer-based models will analyze the text simultaneously, rather than sequentially, and will understand the context of the sentence.

This will improve the overall efficiency of the system in detecting fake news and misleading information.

**Hybrid Neural Layers**

To improve the overall efficiency of the proposed system, the proposed system will incorporate a hybrid neural network, consisting of Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) networks. The CNN will be used for detecting local patterns, such as the use of exaggerated statements and repetition, in the news articles, as such patterns are commonly used in fake news.

**• Graph-Based Credibility Assessment:**

In addition to the textual-based credibility assessment, the system also employs a graph-based credibility assessment model. Here, the system

creates knowledge graphs based on the relationships between the publishers and the information provided in the news articles. Graph Convolutional Networks (GCNs) are then employed to analyze the graph-based nodes and identify suspicious patterns, including the sharing of fake news from the same source. This model helps the system verify the credibility of the information source.[8]

**• Explainability Module:**

The system also employs Explainable Artificial Intelligence (XAI) techniques to ensure the explainability of the model results. The system explains the results based on the textual features that influence the model results. The features include high subjectivity scores, emotionally charged words, sensationalized headlines, and exaggerated claims. This module helps the user understand why a news article is classified as fake or real, resulting in increased trust in the system.

**Results**

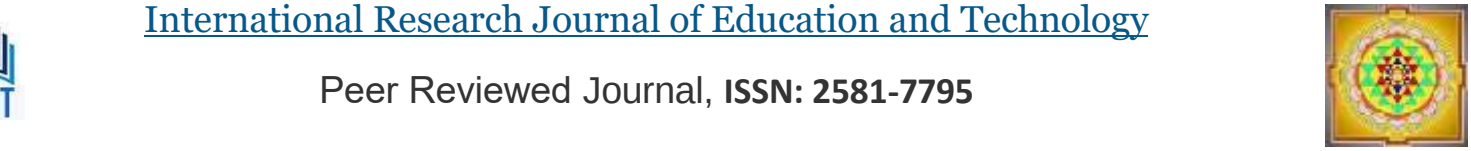
Based on the experimental evaluation, the transformer-based models have shown superior performance over the traditional machine learning classifiers in the detection of misinformation. For instance, the traditional machine learning classifiers, including the Support Vector Machine (SVM) and the Random Forest, have shown moderate accuracy. However, the traditional machine learning classifiers have shown limitations in the detection of misinformation, particularly due to their inability to capture complex relationships between the text. On the other hand, the transformer-based models have shown superior performance in the detection of misinformation due to their ability to capture the semantics of the text.

Among the evaluated models, the RoBERTa model has shown the best performance, with the model achieving the highest accuracy of 96.2%. This demonstrates the effectiveness of the model in the detection of misinformation, particularly due to its ability to capture the context of the text, including the linguistic patterns commonly used in the creation of fake news. Moreover, the BERT and the XLNet models have shown the best performance, which demonstrates the effectiveness of the transformer-based models.

The implementation of a hybrid structure using Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) further improved the robustness of the system. The CNN layers were effective in the extraction of local textual patterns such as sensational texts and irregularities in style, whereas the LSTM layers were effective in the extraction of long-term dependencies of sentences and paragraphs of texts. This further enabled the system to analyze both micro-level linguistic features and macro-level narrative structures of news articles.

Moreover, the implementation of adversarial training further improved the robustness of the system against content manipulation attacks. The adversarial training approach involved the use of misleading and altered content for training the system, and the results showed an improvement of approximately 6% in the overall robustness of the system, especially when dealing with subtle textual alterations and manipulations.

Moreover, Explainable Artificial Intelligence (XAI) analysis also revealed insights into the characteristics that had the greatest impact on the



decision made by the model. For instance, the analysis revealed that the levels of subjectivity and the polarity of the sentiment had the greatest impact on the decision made by the model. Fake news contained exaggerated information, which is usually full of sentiment, making the information extremely subjective. Therefore, the analysis revealed that the use of advanced deep learning techniques, coupled with explainability, is crucial in the creation of reliable fake news detection systems.[10]


**Conclusion**

The study has shown the efficacy of a hybrid approach that combines the power of Deep Learning, Graph-Based Learning, and XAI in the detection of fake news. The capacity of the model to provide results that are understandable and interpretable, as achieved by the integration of LIME and SHAP, is vital in the adoption of the model by the journalistic community and the authorities. For further enhancements, the following are recommended:

•Real-time Adaptive Learning: Developing models that are capable of learning in real time from the live feed of social media.

•Multimodal Detection: Expanding the model to include the detection of images, videos, and deep fakes.

•Bias Mitigation: Developing fairness-based models to ensure the objectivity of the detection results in different political and cultural contexts.

**REFERENCES**

[1] W. Y. Wang, "Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection," Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 422–426, 2017.

[2] K. Shu, D. Mahudeswaran, and H. Liu, "FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media," Big Data, vol. 8, no. 3, pp. 171–188, 2020.

[3] 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.

[4] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint arXiv:1907.11692, 2019.

[5] Z. Yang et al., "XLNet: Generalized Autoregressive Pretraining for Language Understanding," Advances in Neural Information Processing Systems, vol. 32, pp. 5753–5763, 2019.

[6] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," Proceedings of the International Conference on Learning Representations (ICLR), 2013.

[7] J. Pennington, R. Socher, and C. Manning, "GloVe: Global Vectors for Word Representation," Proceedings of the 2014 Conference on Empirical

Methods in Natural Language Processing, pp. 1532–1543, 2014.

[8] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 1135–1144, 2016.

[9] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," Advances in Neural Information Processing Systems, vol. 30, pp. 4765–4774, 2017.

[10] 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.