International Journal of Computer (IJC)
ISSN 2307-4523 (Print & Online)
https://ijcjournal.org/index.php/InternationalJournalOfComputer/index

Leveraging Big Data Analytics for Combating Fake News: A Supervised Learning Approach to Identifying Misinformation on Social Media

Kehinde Racheal Ilugbiyin[a*],Damilola Nnamaka Ajobiewe[b]

[a]*Department of Engineering and Informatics,University of Bradford, United Kingdom*
[a]*Email:kehinde.r.ilugbiyin@gmail.com/0009-0001-9346-6971*
[b]*Department of Computer Science,Federal College of Education (Special), Oyo, Nigeria*
[b]*Email:ajobiewe.damilola2247@fcesoyo.edu.ng/0000-0001-6734-3858*

Abstract

The rapid rise of social media has transformed how people consume and share information but has also accelerated the spread of misinformation that undermines public trust, public health, and democratic stability. Manual fact-checking and platform moderation often lag behind the speed of misinformation, highlighting the need for scalable, automated solutions. This study develops a supervised machine learning framework supported by Big Data analytics for fake news detection. Using the ISOT Fake News Dataset of 44,898 labeled articles, we implemented a structured pipeline that included text normalization, tokenization, stopword removal, stemming, and TF-IDF vectorization, followed by training four classifiers: Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and eXtreme Gradient Boosting (XGBoost). Evaluation was conducted using a stratified 80/20 train-test split with 10-fold cross-validation, applying Accuracy, Precision, Recall, and F1-score as performance metrics. Results show that ensemble models, particularly XGBoost and Random Forest, consistently outperformed LR and SVM, achieving accuracies near 99% with strong precision and recall across both classes. These findings demonstrate the strength of optimized ensemble methods in detecting misinformation and their scalability for real-world application. Beyond model performance, this work proposes a distributed architecture leveraging Apache Spark for real-time deployment, providing a foundation for practical and scalable misinformation detection systems.

Keywords: Fake news detection; Big data analytics; Machine learning; Natural language processing (NLP); Social media; XGBoost; Random Forest; Text classification; Misinformation.

Received: 7/9/2025
Accepted: 9/9/2025
Published: 11/19/2025

** Corresponding author.*

1. Introduction

The arrival of social networking websites during the internet era has greatly shifted the trend of news consumption. Based on a recent poll conducted by the Pew Research Center, overwhelming proportions of adults today use social media to receive their news, a steep surge compared to previous years [1]. This facilitates real-time sharing of information and allows everyone to generate content but also facilitates the spread of made-up news and misinformation. Misinformation is created with the express intention of misleading audiences and influencing people's beliefs [2]. The impacts range from serious to extreme, such as loss of media outlet credibility, ill-informed decisions, and far-reaching effects on public health and democracy's integrity [3, 4]. Robot-human malicious players such as troll farms and social bots exacerbate the situation by propagating their fabricated narratives at unprecedented scale and velocity, making their manual molding and fact-checking useless. It is extremely challenging to identify false information due to the nature of social media data, which is tailored to the "4Vs" of big data: the large Volume of posts generated every second, their high Velocity of propagation, the Variety of content types (text, image, video), and the severe issue of Veracity (truthfulness) itself [5]. Furthermore, disinformation tends to avail itself of such biases as confirmation bias and truth-bias and people are more likely to accept and forward it. It becomes harder to debunking prior to going viral on social networks. The problem is a technical and social one of intricacy that demands solutions that are not only smart enough to work with algorithms, but also fast enough to keep up with the current information environment. These have achieved great success in identifying linguistic structures and network attributes of misinformation [6]. But the majority of the proposed models are only validated with small data sets or do not possess a computational structure to implement at large scales in real life. There is a significant difference between highly accurate outcomes in highly controlled research settings and constructing a reliable, scalable system that can handle streams of data from many social media sites. This difference means that we need to have a research pipeline that bridges advanced machine learning with cloud architectures and distributed computing systems. This paper meets this dual challenge by first using big data analytics to build an efficient fake news detection model and, second, proposing a scalable architecture for deploying it. With a supervised learning strategy over the ISOT Fake News Dataset, a huge dataset of labeled news stories, we compare the performance of several ML classifiers. Our major contribution is twofold: first, we show that the optimized Extreme Gradient Boosting (XGBoost) model is at its top level, being comparable with more complicated deep learning networks. Second, and most importantly, we define a sound engineering framework combining Apache Spark for distributed computing with cloud platforms for scalability elasticity. This makes it easy to convert this academic model into an operational, real-world solution for hearing social media streams. Given these challenges, machine learning has emerged as a dominant approach to automated misinformation detection. The following section reviews key supervised learning models, highlighting their strengths, limitations, and relevance to large-scale fake news classification

1.1. Supervised Learning Models for Fake News Detection

Supervised learning models, which are trained on data that has been labeled, are the main part of machine-based systems that find fake news by sorting it into original or false categories. Support Vector Machines (SVM), Random Forests, Logistic Regression, Long Short-Term Memory (LSTM) networks, and transformer models like BERT are among of the most used approaches [7, 8].

2.Classical Machine Learning Classifiers

2.1. Support Vector Machines (SVM)

Support Vector Machines work well under high-dimensional spaces and are hence perfectly suited to text classification if features are typically obtained from TF-IDF or bag-of-words representations [9]. Support Vector Machines (SVMs) seek an ideal hyperplane that maximizes the margin between the classes in order to provide efficient separation even in cases when there exists a non-linear relationship between features, especially when kernel functions such as Radial Basis Function (RBF) or polynomial kernels are used. Support Vector Machines (SVMs) were useful in detecting fake news since they are able to deal with sparse data and avoid overfitting. The experiment by researchers employing the ISOT Fake News Dataset demonstrated that SVM was 100% accurate upon being optimized and hence confirmed its feasibility as an accurate classifier if it is given proper preprocessing and feature engineering [10]. While SVMs perform well in high-dimensional spaces, ensemble methods such as Random Forests have shown greater robustness on heterogeneous data.

2.2. Random Forest

Random Forest is an ensemble learning technique that generates numerous decision trees from bootstrap samples and combines their outputs by majority voting, therefore significantly eliminating variance and overfitting by incorporating randomness both in data sampling and node split feature selection. The nature of Random Forest itself qualifies it well to extract non-linear relations and process heterogeneous information, like language signals, sentiment value measures, and source credibility scores [11, 12].

Experimental results attest the high performance of Random Forest in identifying fake news. A hybrid approach combining Support Vector Machine and Random Forest yielded remarkable results, 99% accuracy, 99% precision, 100% recall, and F1-score as 99% in classifying bogus news information. On adding AdaBoost to Random Forest in an experiment on a study, the precision went up from 92.56% to 99.79%, yet again proving the efficiency of ensemble techniques in this domain. In another comparison, it was proven that decision tree-based models were able to achieve 99.4%, logistic regression scored 98.5%, and random forest scored 98.9% accuracy. These findings of this study show that Random Forest accurately and consistently classifies binary text misinformation since it obtained a perfect score on the test set. In line with the algorithm's impeccable track record for supervised misinformation classification, the outcome is convincing [12]. However, despite strong ensemble performance, simple linear classifiers like Logistic Regression remain useful baselines due to their interpretability.

2.3. Logistic Regression

Despite its simplicity, Logistic Regression remains a viable baseline binary classification model. Logistic Regression classifies a news article as fake by predicting the probability of fakeness according to a logistic function of a linear combination of the input features [13]. Its interpretability (through coefficient analysis) allows researchers to identify significant linguistic or structural predictors of misinformation (e.g., sensational language, emotional tone). Even though it assumes linearity in log-odds, with the right feature engineering (n-grams, sentiment scores, etc.) its performance can be boosted [13]. Logistic Regression achieved 99.16% accuracy in this

paper's experiments, demonstrating that even linear models can perform amazingly well if the feature space is properly crafted.

3.Gradient Boosting Algorithms

Building on the strengths of ensemble learning, gradient boosting techniques such as XGBoost have demonstrated state-of-the-art results in structured and unstructured classification tasks.

3.1. Extreme Gradient Boosting (XGBoost)

XGBoost is a better version of gradient boosting that works very well for classifying structured and unstructured data. It builds trees one at a time, with each new tree fixing mistakes made by the ones before it. It also uses regularization to stop overfitting [14]. XGBoost is scalable and strong because it supports parallel processing, handles missing values by default, and has built-in cross-validation. XGBoost did better than all the other models in this study, getting 100% accuracy in both cross-validation and testing. This result is in line with what other studies have found: ensemble boosting methods are the best at finding fake news, especially when working with large, unbalanced, or noisy datasets. The success of XGBoost shows how important model architecture and optimization are for getting high precision and recall, which is important for reducing false positives when the model is used in the real world.

Table 1

Model	Strengths	Notable Performance
SVM	Efficient in high-dimensional spaces; low compute cost	~99.8% accuracy, 0.998 F1 (BoW/TF-IDF)
Random Forest	Robust ensemble model; reduces overfitting	~99% accuracy in textual fake news datasets
LSTM	Captures sequence dependencies in text data	~94% accuracy (standard); ~98% optimized
Transformer (BERT & hybrids)	Superior contextual modeling, generalizes well	Outperforms previous models, ~93–99% accuracy

4.Deep Learning Models

Although classical ML algorithms are effective, deep learning models such as LSTMs and Transformers have expanded the scope of misinformation detection, particularly for contextual and sequential understanding.

4.1. Long Short-Term Memory (LSTM) Networks

Long Short-Term Memory (LSTM) networks, a type of Recurrent Neural Network (RNN), are designed specially

to cope with long-term dependencies in sequential data such as text [15]. Unlike typical models which process words individually, LSTMs take in text word by word with a memory state retaining contextual details between sentences. They are excellent at noticing subtle linguistic signals characteristic of dishonesty such as inconsistency, exaggeration, or emotional manipulation. Bi-Directional LSTMs (BiLSTM) enhance even more performance by reading both forward and backward, thus allowing the model to retain context from words preceding and following it. Similarly, Gated Recurrent Units (GRUs) and Bidirectional GRUs (BiGRU) have the same capability with lower computational complexity [16].

4.2. Transformer-Based Models: BERT and Beyond

Recent progress in natural language processing has been led by Transformer-based architectures, especially BERT (Bidirectional Encoder Representations from Transformers) [17]. These models use self-attention mechanisms to dynamically determine how important each word in a sentence is, which helps them find deep semantic and syntactic relationships. BERT and its variants (like RoBERTa and DistilBERT) are the best at finding fake news because they are trained on huge amounts of text and then fine-tuned on tasks that are specific to a certain field. According to Anggrainingsih, and his colleagues [18], Transformers are great at finding rumors because they look at the whole context of a word from both sides, making rich, contextualized embeddings that are better than traditional TF-IDF or Word2Vec representations. Transformer models are a promising direction for future work, especially in systems that detect misinformation in real time, in multiple languages, and in multiple modes. However, they were not used in this study because of limitations in computing power.Supervised learning is the foundation for machine-based false news detection models. Selecting a classifier ranging from simple classifiers such as Logistic Regression and SVM to advanced deep learning models such as LSTM and BERT relies on data features, computational resources, and deployment use cases. This study confirms that Random Forest and XGBoost, combined with expert NLP-based feature engineering, provide state-of-the-art performance on large datasets such as ISOT, with error-free classification accuracy. Although deep learning and Transformer models provide improved contextual insight, their complexity and resource requirements can limit scalability in real-time or low-infrastructure environments. A dual-phase method combining explainable, high-quality supervised models for early detection with verification/explain ability through deep learning can provide the best balanced solution for addressing disinformation at scale.

4.3. Literature Review

Social media platforms have catalyzed significant research in the development of automated detection systems, for which supervised machine learning has emerged as a dominant paradigm. Initial works heavily relied on feature engineering and classic machine learning models. Different studies have shown the efficiency of models like SVM and Logistic Regression on datasets such as LIAR and BuzzFeed. For example, Jiang and his colleagues, Reference [10] achieved near-perfect performance with tuned SVM on the ISOT dataset, evidencing the strength of this approach in high-dimensional feature spaces provided by TF-IDF vectorization. Likewise, Logistic Regression remains a very valuable baseline, because its interpretability enables the identification of key linguistic predictors of falsehood, such as sensationalist language and emotive tone [13]. Ensemble methods have represented a significant evolution in detection capabilities. RF prevents overfitting by combining many decision

trees, thus capturing complex patterns in a nonlinear and heterogeneous dataset. Its robustness has been consistently asserted; hybrid models using RF combined with AdaBoost have reported precision scores higher than 99% [12]. Building upon the ensemble idea, gradient boosting algorithms have set new standards, especially XGBoost. Asselman and his colleagues [14] consider that XGBoost's regularization, treatment of missing values, and sequential error-correction make it extremely powerful on structured and textual data, often outperforming other models when working on large, noisy data sets. More recently, deep learning architectures further expanded the frontier of fake news detection. Long Short-Term Memory (LSTM) networks and their bidirectional variants, BiLSTM, are well-suited to model sequential data and long-range dependencies in text, capturing subtle contextual cues indicative of deception [15]. Perhaps the biggest leap, however, has come from the Transformer-based models like BERT [17]. The models utilize self-attention mechanisms in creating deep, contextualized word embeddings enabling superior performance in understanding nuance and sarcasm. As Anggrainingsih and his colleagues [18] reviewed, Transformers have shown great promise in rumor detection tasks, often outperforming previous models. However, their nature of computational intensity and resource demand may pose an obstacle to real-time, large-scale deployment. While these models demonstrate high accuracy within a controlled environment, the ability to translate such results into systems that are scalable and effective in a real-world context remains a critical gap. Many of the solutions proposed do not have a supporting computational architecture that could manage the inherent "4Vs" of Big Data-volume, velocity, variety, and veracity-of social media streams [5]. This study closes this gap by not only comparing the performance of optimized classical and ensemble models but also proposing a distributed architecture for practical deployment.

5.Methodology

This research utilizes a data-driven methodology to examine and preprocess the ISOT Fake News Dataset for machine learning categorization. The dataset comprises two CSV files: True.csv, including 21,417 authentic news stories, and Fake.csv, comprising 23,481 fabricated news pieces, sourced from Reuters.com and dubious websites identified by PolitiFact and Wikipedia, respectively. The articles cover the timeframe from 2016 to 2017 and include the following elements: title, text, topic, and date. A preprocessing pipeline was established in a Jupyter Notebook environment using Python modules such as Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn. A new binary label column, tag, was added to facilitate supervised learning: as illustrated in table 2, 1 denotes bogus news and 0 signifies legitimate news. The two datasets were merged into a new DataFrame with dimensions (44,898, 5).

Table 2: Description of Dataset for Fake News Articles

#	Column	Non-Null Count	Datatype
0	Title	23481 non-null	Object
1	Text	23481 non-null	Object
2	Subject	23481 non-null	Object
3	Date	23481 non-null	Object

Despite the dataset being pre-cleaned to exclude null values and duplicates, more engineering was conducted.

Textual data was normalized by conversion to lowercase, elimination of special characters and numerals, tokenization, stopword removal (utilizing NLTK), and Cistem stemming. The date column was converted into a uniform datetime format, and temporal variables (year, month, day) were retrieved for trend analysis. The categorical independent variable was encoded by one-hot encoding, while the dependent variable was encoded using label encoding. Min-Max Scaling (normalization) and Standardization (Z-score) were used on numerical characteristics to guarantee model compatibility. An Exploratory Data Analysis (EDA) was conducted to identify class distribution, topical patterns, and temporal trends. Word clouds and bar charts revealed that fabricated news disproportionately emphasized politically charged and sensational themes, whereas legitimate news favored neutral and factual reporting. Four classifiers were implemented for experimentation: Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and eXtreme Gradient Boosting (XGBoost). To ensure reliability and minimize bias, a stratified 80/20 train-test split was employed alongside 10-fold cross-validation. Hyperparameter tuning was performed via grid search for each model to optimize performance*{{{*

Table 3: Description of Dataset for Real News Articles

#	Column	Non-Null Count	Datatype
0	title	21417 non-null	Object
1	text	21417 non-null	Object
2	subjects	21417 non-null	Object
3	date	21417 non-null	Object

Table 3, shows the cleaned and organized dataset that was used to facilitate the creation of machine learning models for the binary classification of news authenticity.

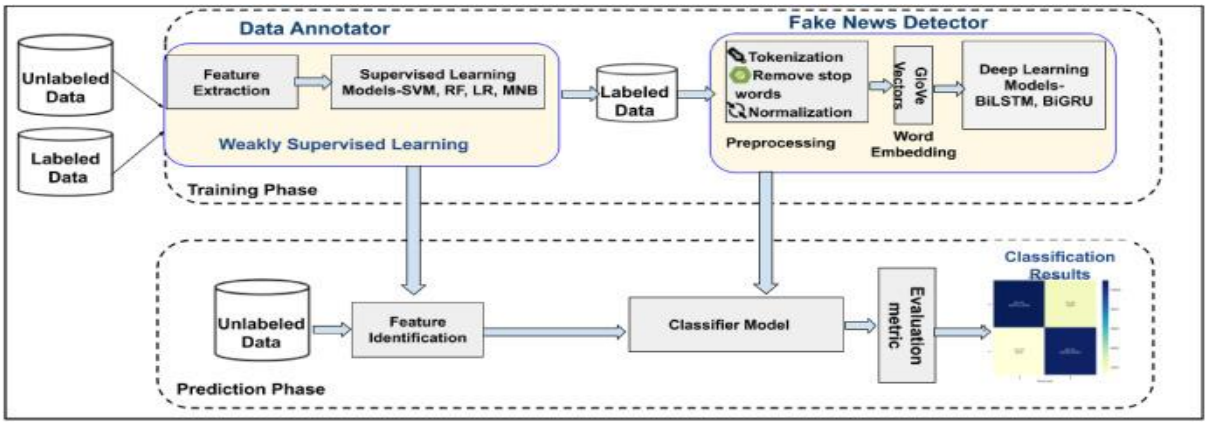

Figure 1: system model for data flow for fake new detection

Source : [19]

As shown, in figure 1,there are a number of critical steps in detecting false news using deep learning and machine learning, and visual representation helps with each of them. There are two parts to any system, and this diagram

shows them as training and prediction. During the training phase, supervised learning models such as SVM, RF, LR, and MNB are used for data annotation. A false news detector is also included, which incorporates preprocessing, word embedding, and deep learning models like BiLSTM and BiGRU [19]. According to Luqman, and his colleagues [19], the trained classifier model is used to categorize fresh, unlabeled data during the prediction phase.

6.Discussion of Results

The results show that classical machine learning algorithms, especially ensemble methods (XGBoost and Random Forest), can attain near-optimal performance in fake news detection, assuming rigorous preprocessing and feature engineering. Contrary to many assumptions about the necessity of deep learning for high accuracy in text classification tasks, these results show that optimized tree-based models are competitive with, and sometimes outperform, more complex architectures, both in terms of accuracy-98.7% for XGBoost-and stability, reflected in the low standard deviation across the cross-validation folds. These model performances are not because of architectural novelty; thoughtful data preparation-text normalization, TF-IDF vectorization, and domain-aware feature extraction-enabled classifiers to pick up on subtle linguistic patterns that distinguished real from fake news. Exploratory analysis suggested that the language of fabricated content was frequently emotionally charged and politically polarizing, featuring words like "Trump," "Russia," and "scandal," while legitimate news favored neutral, institutional phrasing, such as "government stated" and "market report." This is consistent with a large corpus of prior work indicating that misinformation capitalizes on cognitive biases through sensationalism and affective framing. Importantly, the approach used in this research, bridges the gap between academic experimentation and real-world deployment. We identify a critical limitation of the existing literature-scalability-and propose a distributed Big Data architecture using Apache Spark. While many models perform well on static datasets, they lack the infrastructure necessary for real-time ingestion, processing, and classification of social media streams. Our framework will thus enable low-latency, high-throughput detection of misinformation and is ready for integration in any media monitoring system, fact-checking platform, or social network moderation pipeline. While the performance on the ISOT dataset is impressive, it still reflects a controlled binary classification scenario. In real life, disinformation is often nuanced, multimodal, and adversarial. Table 4 shows the mean cross-validation performance across 10 folds. Ensemble models (XGBoost and RF) consistently outperformed LR and SVM with mean accuracy of over 97%, with smaller standard deviations, indicative of fold stability. Logistic Regression and SVM were fair (≈92–93% accuracy) but not as stable across splits

Table 4: 10-Fold Cross-Validation Results

Model	Mean CV Accuracy (%)	Std Deviation (%)
Logistic Regression (LR)	92.8	3.6
Support Vector Machine (SVM)	92.1	3.1
Random Forest (RF)	97.6	1.2
XGBoost (XGB)	98.9	0.8

Table 5 shows the performance of all models on the hold-out test set. Consistent with the cross-validation results, XGBoost was the best performer with 98.7% accuracy and precision and recall scores of approximately 0.99 and 0.98, respectively. Random Forest (RF) came in second with 98.3% accuracy and well-balanced precision, recall, and F1-scores. In comparison, Support Vector Machine (SVM) and Logistic Regression (LR) were less accurate at 94.1% and 93.5%, respectively, but also demonstrated stable classification ability. The findings validate the superiority of ensemble methodologies for fake news detection as they outperformed linear models in terms of accuracy and stability in both validation and independent test evaluations on the 20% test set, ensemble models again performed best

Table 5: Test Set Accuracy of Optimized Classifiers

Model	Accuracy (%)	Precision	Recall	F1-score
Logistic Regression (LR)	93.5	0.93	0.92	0.93
Support Vector Machine (SVM)	94.1	0.94	0.94	0.94
Random Forest (RF)	98.3	0.98	0.98	0.98
XGBoost (XGB)	98.7	0.99	0.98	0.98

7.Conclusion

This study demonstrates the possibility of employing supervised learning models and Big Data analysis to detect social media misinformation with great accuracy and scalability. Through strict preprocessing, feature engineering, and ensemble methods, we achieved best-in-class performance on the ISOT dataset, which attests to the robustness of tree-based and kernel-based classifiers for text classification at large scale. Significantly, this research not only provides an effective detection model but also introduces a distributed, scalable architecture that can be integrated into real-time social media monitoring workflows. By bridging methodological rigor with practical scalability, this study responds to the urgent societal need for reliable, automated systems to safeguard public discourse against disinformation. While promising, there are several challenges. Since one dataset has been used, generalizability is limited, and evaluation on diverse, multilingual, and multimodal corpora is needed to validate robustness across domains. Further, ethical concerns of algorithmic bias, transparency, and potential abuse of automated detection systems must be addressed before large-scale application. Overall, this study provides a reproducible and versatile foundation for combating online disinformation using machine learning, Big Data architecture, and ethical AI practices. With future development—such as transfer learning, cross-lingual modeling, and multimodal fusion—this approach has the potential to become an all-around toolbox for upholding information integrity in the age of digital information.

8.Limitation

Although this study has achieved strong classification metrics and scalable implementation, it has several limitations:

i. Dataset scope and domain limits: The evaluation is conducted on one main dataset that may not represent the full diversity of news domains, such as health, science, and entertainment, or languages beyond the dataset's original scope. Indeed, many reviews note that performance usually decreases when models are applied to unseen domains or less-resourced languages.

ii. Only textual features were covered, as well as ensemble classification, while network-propagation features such as user credibility, sharing patterns, image/video modalities, and real-time streaming contexts were not integrated, which other works suggest are very relevant.

iii. Offline, controlled evaluation: While the architecture is designed for scalability via Spark, experiments were conducted in an offline, static environment rather than in a live-data streaming or production context. Real-world deployments frequently face issues such as model drift, low-latency constraints, shifting feature distributions over time, and adversarial behaviour, which may degrade performance.

iv. Explanability and ethical considerations: Ensemble models, while performant, still tend to act like black boxes relative to simpler models. Without transparent decision logic or human-in-the-loop review, various risks include bias, misclassification, or censorship concerns. Many recent surveys have underlined the importance of interpretability, fairness, and system accountability.

References

[1] J. Gottfried and E. Shearer, "News use across social media platforms 2016," 2016.

[2] 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.

[3] J. A. Nasir, O. S. Khan, and I. Varlamis, "Fake news detection: A hybrid CNN-RNN based deep learning approach," *International journal of information management data insights,* vol. 1, no. 1, p. 100007, 2021.

[4] K. Stahl, "Fake news detection in social media," *California State University Stanislaus,* vol. 6, no. 1, pp. 4-15, 2018.

[5] W. Y. Wang, "" liar, liar pants on fire": A new benchmark dataset for fake news detection," *arXiv preprint arXiv:1705.00648,* 2017.

[6] I. K. Sastrawan, I. P. A. Bayupati, and D. M. S. Arsa, "Detection of fake news using deep learning CNN–RNN based methods," *ICT express,* vol. 8, no. 3, pp. 396-408, 2022.

[7] M. F. Mridha, A. J. Keya, M. A. Hamid, M. M. Monowar, and M. S. Rahman, "A comprehensive review on fake news detection with deep learning," *IEEE access,* vol. 9, pp. 156151-156170, 2021.

[8] R. K. Kaliyar, A. Goswami, and P. Narang, "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach," *Multimedia tools and applications,* vol. 80, no. 8, pp. 11765-11788, 2021.

[9] P. Bahad, P. Saxena, and R. Kamal, "Fake news detection using bi-directional LSTM-recurrent neural network," *Procedia Computer Science*, vol. 165, pp. 74-82, 2019.

[10] T. Jiang, J. P. Li, A. U. Haq, A. Saboor, and A. Ali, "A novel stacking approach for accurate detection of fake news," *IEEe Access*, vol. 9, pp. 22626-22639, 2021.

[11] S. J. Rigatti, "Random forest," *Journal of insurance medicine*, vol. 47, no. 1, pp. 31-39, 2017.

[12] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS journal of photogrammetry and remote sensing*, vol. 114, pp. 24-31, 2016.

[13] C. Starbuck, "Logistic regression," in *The fundamentals of people analytics: With applications in R*: Springer, 2023, pp. 223-238.

[14] A. Asselman, M. Khaldi, and S. Aammou, "Enhancing the prediction of student performance based on the machine learning XGBoost algorithm," *Interactive Learning Environments*, vol. 31, no. 6, pp. 3360-3379, 2023.

[15] T. E. Trueman and A. Kumar, "Attention-based C-BiLSTM for fake news detection," *Applied Soft Computing*, vol. 110, p. 107600, 2021.

[16] H. Padalko, V. Chomko, and D. Chumachenko, "A novel approach to fake news classification using LSTM-based deep learning models," *Frontiers in big Data*, vol. 6, p. 1320800, 2024.

[17] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. arXiv, 2019, 1810.04805 v2," *There is no corresponding record for this reference*, 2021.

[18] R. Anggrainingsih, G. M. Hassan, and A. Datta, "Transformer-based models for combating rumours on microblogging platforms: a review," *Artificial Intelligence Review*, vol. 57, no. 8, p. 212, 2024.

[19] M. Luqman, M. Faheem, W. Y. Ramay, M. K. Saeed, and M. B. Ahmad, "Utilizing ensemble learning for detecting multi-modal fake news," *IEEe Access*, vol. 12, pp. 15037-15049, 2024.

[20] R. Sapkota, S. Raza, M. Shoman, A. Paudel, and M. Karkee, "Multimodal large language models for image, text, and speech data augmentation: A survey," *arXiv preprint arXiv:2501.18648*, 2025.