IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Data-Driven ML Approach for Fake News Detection Using DGLRF

Krish Goyal[1], Devansh Tomar[2], Jai Badachiya[3], Navneet Jaguri[4], Daksh Dhamija[5]

[1-5]University Institute of Engineering, Chandigarh University, Mohali, India

Abstract: Fake news is a significant global concern, distorting facts and influencing public opinion. This study leverages machine learning techniques, including Logistic Regression, Decision Trees, and Deep Learning [1], achieving over 99% accuracy in fake news detection. By prioritizing data preprocessing, such as noise reduction and feature selection, our approach addresses data imbalance and enhances model reliability. This research offers a scalable solution to misinformation, contributing to the mitigation of its societal impact.

Keywords: Fake news detection, Machine learning algorithms, Data preprocessing, Misinformation identification, Deep learning methods

1. INTRODUCTION

The rise of the internet has revolutionized the way information is disseminated, making it an indispensable tool for communication, knowledge sharing, and global connectivity. However, this rapid technological advancement has also given rise to a parallel problem, an unprecedented surge in the spread of unregulated and unverified information, particularly through the popular social media platforms like X, Instagram etc. False news, often referred to as "fake news," has the potential to influence public perception, skew decision-making, and even destabilize societal structures. The COVID-19 pandemic served as a critical example, where the rampant spread of misleading information about the virus, its origins, and treatments contributed to confusion, public health risks, and misguided actions. This issue underscores the pressing need for reliable and effective solutions capable of distinguishing between authentic news and fabricated content.

Major social media platforms, such as Facebook, Twitter, and others, have become major conduits for the dissemination of fake news due to their vast user base and the ease with which information can be shared. With millions of daily active users, these platforms have accelerated the spread of misinformation, often without adequate fact-checking mechanisms. As a result, false claims and inaccurate information are frequently circulated, amplifying their reach and impact. This challenge has drawn significant attention from researchers and policymakers alike, driving the need for solutions that utilize technological innovations to combat the problem.

Machine learning (ML) has emerged as a promising approach for fake news detection, offering automated methods to sift through vast datasets, identify patterns, and classify news content as either reliable or false. In this research, the DGLRF methodology applies four distinct machine learning classifiers individually for fake news detection: Decision Tree Classifier (D), Gradient Boosting Classifier (G), Linear Regression (L), and Random Forest (F). The Decision Tree Classifier (D) [2] uses a tree-like structure for decision-making based on input features, offering a clear and interpretable classification approach. The Gradient Boosting Classifier (G) builds weak learners sequentially to correct previous errors, thereby enhancing overall predictive accuracy. Linear Regression (L), although traditionally used for regression, assists in feature analysis and provides a benchmark for comparing more complex models. The Random Forest (F) model constructs more than one decision trees and aggregates for their predictions to improve accuracy and mitigate overfitting. By evaluating each of these models independently, the research aims to assess their performance and effectiveness in detecting fake news, offering a detailed understanding of each model's contribution to the classification task.

The significance of this work extends beyond just improving detection accuracy. Fake news has far-reaching implications, affecting not only public health but also political stability, economic conditions, and global security. The widespread availability of false information can erode trust in democratic institutions, skew election results, and foster divisive ideologies. Therefore, the development of a reliable fake news detection system serves a critical role in safeguarding information integrity and ensuring the availability of truthful, verified content in the digital landscape.

The following sections of this research explore the effectiveness of machine learning techniques in detecting fake news, with a particular focus on their application to real-world data. The experiments conducted provide insights into the strengths and limitations of each algorithm, and the results serve as a foundation for future advancements in the field of misinformation detection.

1.1. Relevant Contemporary Issues

The COVID-19 infodemic poses a serious threat to public health and impedes pandemic response efforts because of false information about the virus's diagnosis, treatment, and prevention. Fake news also contributes to political polarization and manipulation, which has an impact on global governance and elections. These difficulties highlight the need for cutting-edge machine learning techniques to quickly identify and dispel false information. Scholars notably Shaikh and Patil (2020) have demonstrated how machine learning may be used to detect and reduce false news, underscoring the significance of this subject in tackling modern problems.

1.2. Identification of Problem

The spread of false information poses serious problems for social cohesiveness, politics, and public health. False information has the power to divide society more sharply, mislead people, and distort public opinion. Confusion and possible injury were brought on by false information on treatment and preventive measures during the COVID-19 pandemic. To ensure the dissemination of accurate information and protect the public's well-being, this circumstance necessitates the development of trustworthy machine learning models to recognize and handle fake news.

The goal of the research is to create and apply cutting-edge machine learning techniques to accurately detect and categorize bogus news. This entails using a variety of algorithms, including ensemble techniques, decision trees, and logistic regression, to evaluate news and separate reliable sources from fraudulent ones. Data preparation is essential for reducing false positives and negatives, improving model performance, and fine-tuning datasets. The goal is to develop trustworthy prediction models that can identify false news and lessen the negative consequences it has on society. The research also focuses on modifying these models to stay up to date with changing disinformation tactics.

1.3. Problem Description and Contribution

To identify and classify fake news, prior studies have used machine learning techniques like Support Vector Machines (SVM), passive-aggressive classifiers, clustering algorithms, and deep learning techniques. These approaches face challenges due to limited datasets reliance on textual data alone, and high mortality. This research addresses this gap by developing and evaluating a model using cleaning and pre-processing techniques to achieve more accurate and better performance in detecting fake news on various types of data.

1.4. Related Work

Current research on supervised learning for fake news detection using machine learning offers different perspectives and approaches.

Logistic Regression is a classification algorithm commonly used to detect fake news. It is often used for its simplicity and clarity. Analysis [1] of Menard's logic hierarchy is a very good discussion and application guide for this algorithm.

Decision Tree is less sensitive to false positives due to its robustness in handling difficult decisions that are used to search. process Safavian and Landgrebe [3] provide an overview of decision tree classifiers, and Lyu and Lo [4] thoroughly investigate fake news detection using decision trees and find them useful, but they can be added.

Gradient Boosting Classifier is an advanced machine learning technique used in a variety of applications, including fake news detection. Although not directly covered in the references provided, [5] it has been found useful for classification tasks due to its robustness and ability to handle large datasets.

Random Forest Classifier is another well-known ensemble learning method. Accuracy and ability to process various types of data. Segal [6] provides a study of Random Forest techniques, which can detect fake news.

Existing research explores various machine-learning techniques for fake news detection. Logistic Regression and Decision Trees are commonly used for their interpretability. However, ensemble methods like Random Forest and Gradient Boosting offer improved accuracy by leveraging multiple models. Studies by Shaikh and Patil [7] and Ahmad et al. [8] demonstrate the superiority of ensemble approaches. Our work builds on these findings by incorporating advanced preprocessing techniques to further enhance model robustness and reliability.

Sharma et al. [9] contribute by reviewing different tools for fake news detection and emphasizing deep learning methods such as LSTM and BI-LSTM, which can complement supervised learning cycles. In their systematic assessment of machine learning techniques, Manzoor et al. [10] provided a detailed examination of the benefits and drawbacks of different models.

The research emphasizes the use of decision trees, random forests, logistic regression, and gradient boosting—often in

combination—to enhance the identification of false news. Researchers can create more accurate and reliable models for spotting bogus information by carefully processing data.

1.5. OBJECTIVES

A lot of people are concentrated on accomplishing their goals:

i. Improve the accuracy of fake news detection using advanced machine learning models such as logistic regression, decision trees, gradient clustering, and random forest classification.
ii. Examine how different data formats, such as text, can be used to recognize false information.
iii. Evaluate and contrast various machine learning models to determine the best technique for spotting false information.
iv. Uses cleaning and preprocessing techniques to improve model reliability and fake detection performance.

1.6. DESIGN CONSTRAINTS:

The importance of data quality and having access to resources like Kaggle—which provides a balanced dataset with 5,320 real news stories and 5,377 fake news pieces—should be emphasized in the document on fake news detection. Accurate detection results and efficient model training depend on this data equilibrium. Furthermore, the complexity of models and the degree of parameter optimization may be limited by limitations on processing power and time. During the design process, ethical considerations like upholding secrecy and preserving trust must be carefully considered.

1.6.1. FEATURE SELECTION:

The implementation of machine learning (ML) in a project is heavily influenced by feature selection, which plays a key role in determining model performance. Choosing the right features for training is crucial, as it directly impacts the accuracy and efficiency of the model. Redundant or irrelevant features can introduce noise and unnecessary complexity, which may lead to reduced performance. By focusing on selecting the most relevant data, feature selection enhances the model's ability to generalize and reduces computation time [11].

In text-based data, techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) [12] scores help extract important linguistic features. This process highlights keywords and their frequency within the dataset, improving model focus. Other methods, such as word embeddings, further capture semantic meaning, leading to more accurate predictions. Effective feature selection ensures that the model is robust and well-suited to handle complex datasets efficiently.

1.6.2. FEATURE IMPORTANCE:

Classifying the distribution of values among different features provides insight into how these features affect the model's predictions. Importance ratings offer a high-level assessment of each feature's significance within the model. Analyzing these ratings reveals features with minimal contribution, allowing for the removal of low-scoring features and the retention of those with higher scores. This process not only guides the model's design but also enhances its behaviour and overall efficiency. By examining feature significance, it becomes clear which data aspects have the most substantial impact on the model's performance. For example, specific words or phrases in text data might be pivotal in distinguishing between false and true information. This analysis not only improves model interpretation but also optimizes performance and increases the accuracy of false information detection.

2. LITERATURE SURVEY

A literature review serves as a critical assessment of existing research relevant to the topic at hand. It involves identifying, analyzing, and synthesizing previous studies to gain a comprehensive understanding of the current state of knowledge. This process not only highlights the strengths and weaknesses of prior work but also reveals gaps and inconsistencies that can inform future research directions. In the context of fake news detection, the literature review covers a broad spectrum of machine learning techniques, evaluating their performance, limitations, and potential for improvement. By doing so, it sets the foundation for the subsequent experimentation and advancements presented in this research.

Table 1. Literature Survey on Previous Studies

YEAR	NAME OF PAPER	FINDINGS
2024	Large Language Model Agent for Fake News Detection [13]	LLM-based agents improve accuracy in fake news detection, leveraging internal and external knowledge sources
2023	Advancing fake news detection: hybrid deep learning with fast-text and explainable AI.[14]	AI-based systems can achieve real-time detection of misinformation
2022	Deep Learning for Misinformation in Social Networks [15]	Advanced deep learning models improve accuracy in fake news detection
2021	Evaluation of Tools for Fake News Detection [16]	Models such LSTM and BI-LSTM as enhance the detection of false information
2020	Fake News Detection Using Machine Learning [17]	ML approach can reduce the spread of misinformation
2020	Fake News Detection Using ML Ensemble Methods [18]	Ensemble methods improve performance in classifying false information
2019	Fake News Detection Using ML: A Systematic Review [19]	Explores various ML techniques and their effectiveness in identifying fake news

This research is motivated by the need to address shortcomings by integrating comprehensive data preprocessing techniques and employing both traditional and deep learning algorithms. By focusing on these aspects, the research seeks to improve model accuracy, adaptability, and scalability, ensuring more reliable fake news detection in real-world applications.

3. FLOWCHART

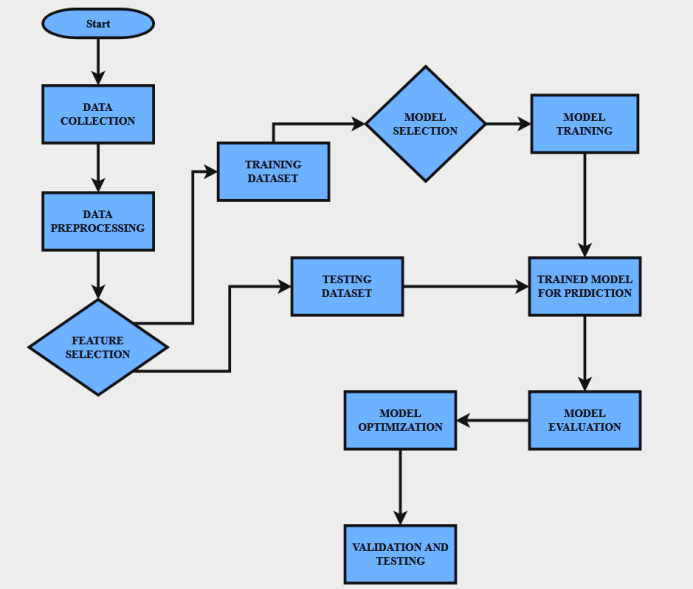

Figure 1: Flowchart

The flowchart illustrates the methodology used in this research for fake news detection. It starts with data collection from reliable sources, followed by data preprocessing, which involves noise reduction and feature selection. The prepared data is then input into various machine learning models, including Decision Tree, Gradient Boosting, Linear Regression [1], and Random Forest,[2] each assessed for its effectiveness. The process concludes with performance analysis, evaluating models based on accuracy, precision, recall, and F1 scores. This visual representation provides a clear overview of the research methodology

4. RESULT ANALYSIS AND VALIDATION

4.1. PARTICULARS PREPROCESSING

Previous research has been instrumental in enhancing the effectiveness of supervised learning models for detecting fake news. Before the application of machine learning algorithms, missing values that could impact model performance were systematically addressed. For instance, attributes such as “material weight” and “flow size” exhibited 17% and 28% missing values, respectively. Analysis of the relationships between functionally similar attributes revealed overlaps that required resolution, leading to the removal of redundant elements to clarify the model. Additionally, invalid values, such as "LF" and "reg" in the object's fat content attribute, were appropriately processed and replaced. A value of 0 in the “object simplicity” condition also suggested potential heterogeneity in data collection.

Table 1: Summary of Data Pre-processing

Features	Category	Binary Encoding
News Recourse	Twitter	1
News Recourse	Facebook	0
News	Real	1
News	Fake	0
News Sentiment	Positive	1
News Sentiment	Negative	0

4.2. CONFUSION MATRIX:

To judge the performance of various ML models, confusion matrices were generated for each approach. These matrices illustrate the relationship between reliable and predictive rankings of fake news within the ideal models. The confusion matrix for the decision tree model demonstrated a marginally improved false positive and false negative rate [20] compared to the logarithmic model, showing more reductions in errors and indicating better prediction accuracy. For classification tasks, the Forest model outperformed others, exhibiting the fewest false positives and false negatives. This model processes data without including headers.

Table 2: Confusion Matrix

Total	Class1 (Predicted)	Class1 (Predicted)
Class1(actual)	True Positive	False Negative
Class2(actual)	False Positive	True Negative

4.3. MODEL EVALUATION MATRIX:

Several assessment indicators must be taken into account when evaluating machine learning models' efficacy in identifying false news. These consist of F1 score, recall, accuracy, and precision. The percentage of accurate predictions the model produces is known as accuracy. Recall gauges the model's capacity to recognize every pertinent event, whereas precision evaluates the calibre of positive predictions.[21] The F1 score offers a thorough understanding of the model's overall performance by combining recall and precision. These metrics aid in evaluating a model's accuracy in text classification.

4.4. ACCURACY SCORE:

The percentage of accurate predictions—including true positives and true negatives—as a fraction of the total data set is measured as accuracy. Our objective in this project is to optimize accuracy to deliver exact and trustworthy information classification [22].

$$\text{Accuracy} = \frac{|T\,P|+|TN|}{|T\,P|+|TN|+|F\,P|+|F\,N|}$$

4.5. PRECISION:

The precision measures the percentage of expected positive cases that turn out to be positive. To ensure that the model correctly identifies real instances of fake news, high precision is essential for minimizing false positives. Models including random forest classifiers, gradient boosting, and decision trees demonstrated their efficacy in separating fake news from real news in our study by achieving a perfect precision score of 1.00 [22].

$$\text{Precision} = \frac{|\text{T P}|}{|\text{T P}|+|F\ P|}$$

4.6. RECALL:

Recall assesses the model's ability to identify all significant instances, such as actual fake news, in a data set. High recall means that the model effectively captures the majority of positive cases and minimizes false negative cases. In our project, models such as decision tree, gradient boosting classifier and random forest classifier outperformed recall, each with a perfect score of 1.00. [22]

$$\text{Recall} = \frac{|\text{T P}|}{|\text{T P}|+|\text{F N}|}$$

4.7. F1 SCORE

The F1 score is a metric for assessing the performance of a machine learning classification model, especially when dealing with imbalanced class distributions. It balances precision and recall, providing a unified measure that encompasses both aspects of the model's effectiveness. [22]

4.8. ANALYSIS:

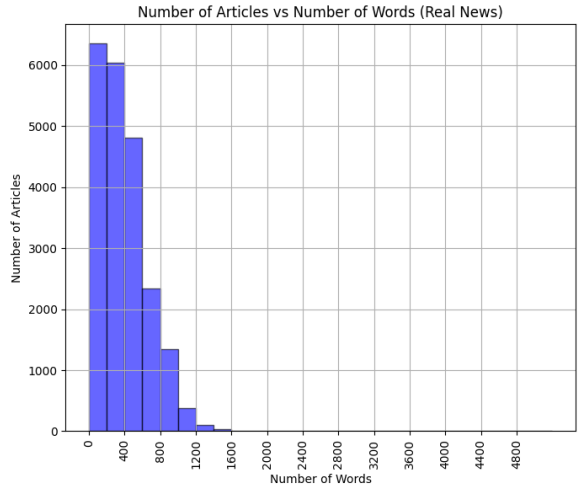

Figure 2: Real News Text Length Counts

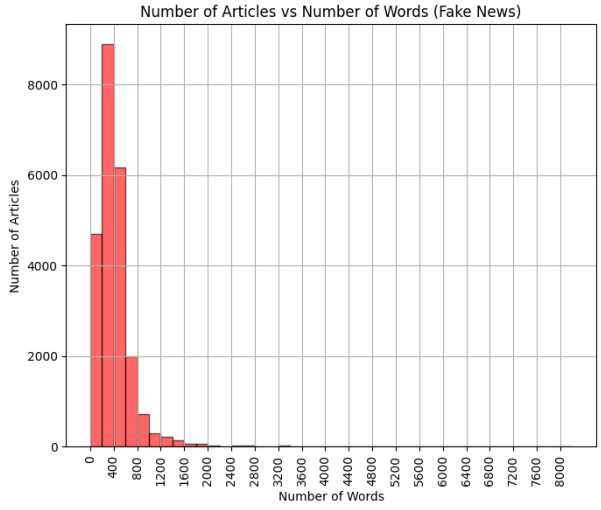

Figure 3: Fake News Text Length Counts

4.9. METRICS:

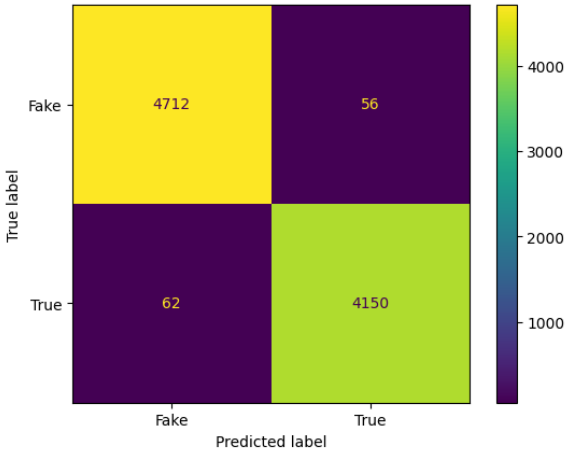

Figure 4: Forecast & Real Labels of Confidence Matrices for False & True News Information using LR

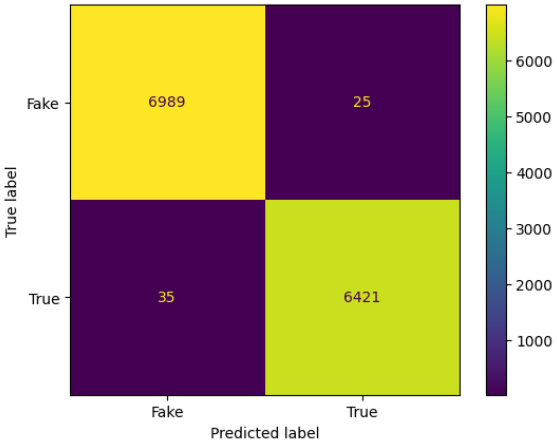

Figure 5: Forecast & Real Labels of Confidence Matrices for False & True News Information using DT

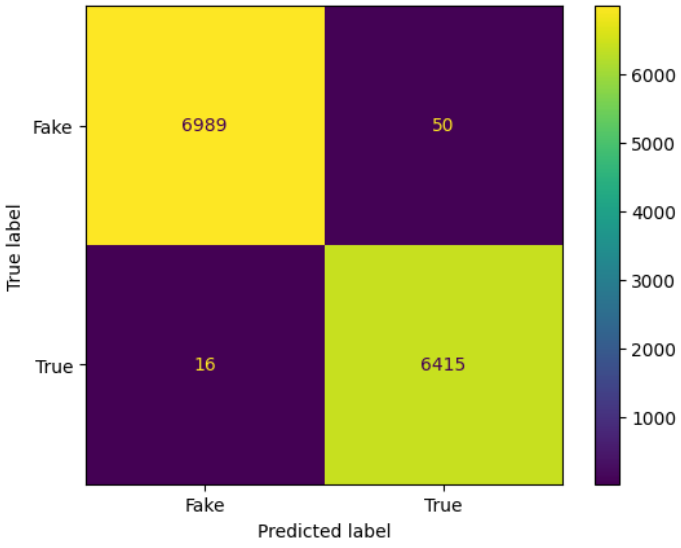

Figure 5: Forecasted & Real Labels Confidence Matrices for False & True News Information using Gradient Boost Classifier

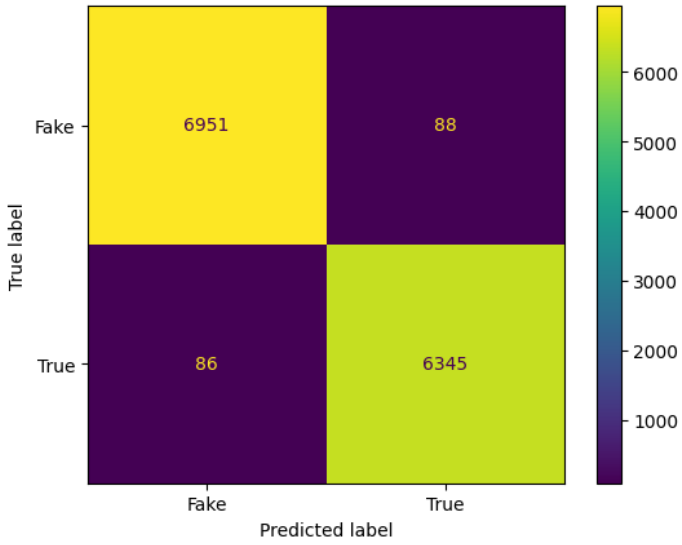

Figure 5: Forecasted & Real Labels Confidence Matrices for False & True News Information using Random Forest Classifier

4.10. VALIDATION:

This study evaluated the effectiveness of various models and classifiers in identifying false news. The Random Forest Classifier demonstrated the highest accuracy at 99.88%, making it the most effective model. The Decision Tree model closely followed with an accuracy of 99.84%.[2] Other models, such as the Gradient Boosting Classifier and Logistic Regression, also showed strong accuracy, making valuable contributions to fake news detection.

Table 3: Model Accuracy Analysis

S.No.	Model Name	Accuracy Score (out of 100)
1.	Logistic Regression	98.645
2.	Decision Tree	99.661
3.	Gradient-Boosting Classifier	99.652
4.	Random-Forest Classifier	98.865

5. CONCLUSION

In this study, machine learning models, such as logistic regression, decision trees, gradient boosting, and random forest [2] classifiers, were employed to develop a supervised learning system for fake news detection. These models achieved superior results, with accuracy rates exceeding 99%. This study successfully demonstrates the potential of machine learning models in fake news detection, achieving remarkable accuracy rates. Our approach, characterized by extensive data preprocessing and model evaluation, sets a new benchmark for reliability in misinformation detection. Future work will focus on expanding dataset diversity and integrating real-time analytics to adapt to evolving misinformation tactics, thereby enhancing the system's applicability and trustworthiness.

6. FUTURE WORK:

Several key improvements can be made to enhance the false news detection project moving forward. Future efforts will expand the dataset to include multilingual sources, enhancing model generalization. Incorporating real-time analytics will enable the timely detection of emerging fake news. Further, exploring novel machine learning techniques and refining feature engineering will potentially boost model performance. Continuous model monitoring and adaptation will ensure relevance in an ever-changing misinformation landscape.

REFERENCES

[1] Goodfellow, I., Bengio, Y., Courville, A. (2016). Deep Learning. MIT Press.

[2] Khan, Aurangzeb, Baharum Baharudin, Lam Hong Lee, and Khairullah Khan. "A review of machine learning algorithms for text-documents classification." *Journal of advances in information technology* 1, no. 1 (2010): 4-20.

[3] Yang, Zhen, Xiaodong Liu, Tong Li, Di Wu, Jinjiang Wang, Yunwei Zhao, and Han Han. 2022. "A Systematic Literature Review of Methods and Datasets for Anomaly-Based Network Intrusion Detection." Computers & Security 116 (March): 102675–75. https://doi.org/10.1016/j.cose.2022.102675.

[4] Safavian, S.R.; Landgrebe, D. A survey of decision tree classifier methodology. IEEE Trans. Syst. Man, Cybern. 1991, 21, 660–674.

[5] Lyu, S.; Lo, D.C.T. Fake News Detection by Decision Tree. In Proceedings of the 2020 SoutheastCon, Raleigh, NC, USA, 28–29 March 2020; pp. 1–2.

[6] National Library of Medicine. National centre for biotechnology information

[7] Segal, M.R. Machine Learning Benchmarks and Random Forest Regression; Kluwer Academic Publisher: Amsterdam, The Netherlands, 2004. 7896641336

[8] Shaikh, J.; Patil, R. Fake News Detection using Machine Learning. In Proceedings of the 2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC), San Francisco, CA, USA, 16–17 December 2020; pp. 1–5.

[9] Ahmad, I.; Yousaf, M.; Yousaf, S.; Ahmad, M.O. Fake News Detection Using Machine Learning Ensemble Methods. Complexity 2020, 2020, 1–11.

[10] Sharma, D.K.; Garg, S.; Shrivastava, P. Evaluation of Tools and Extension for Fake News Detection. In Proceedings of the 2021 International Conference on Innovative Practices in Technology and Management (ICIPTM), Gautam Buddh Nagar, India, 17–19 February 2021; pp. 227–232.

[11] GeeksforGeeks. 2021. "Feature Selection Techniques in Machine Learning." GeeksforGeeks. January 19, 2021. https://www.geeksforgeeks.org/feature-selection-techniques-in-machine-learning/.

[12] Wikipedia Contributors. 2024. "Tf–Idf." Wikipedia. Wikimedia Foundation. July 26, 2024. https://en.wikipedia.org/wiki/Tf%E2%80%93idf.

[13] Manzoor, S.I.; Singla, J.; Nikita. Fake News Detection Using Machine Learning Approaches A Systematic Review. In Proceedings of the 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 23–25 April 2019; pp. 230–234.

[14] X. Li, Y. Zhang, and E. C. Malthouse, "Large Language Model Agent for Fake News Detection," *arXiv.org*, 2024. https://arxiv.org/abs/2405.01593

[15] E. Hashmi, Sule Yildirim Yayilgan, Muhammad Mudassar Yamin, S. Ali, and M. Abomhara, "Advancing Fake News Detection: Hybrid Deep Learning With FastText and Explainable AI," *IEEE Access*, vol. 12, pp. 44462–44480, Jan. 2024, doi: https://doi.org/10.1109/access.2024.3381038.

[16] Islam, M. R., Liu, S., Wang, X., & Xu, G. (2020, September 29). Deep learning for misinformation detection on online social networks: a survey and new perspectives. Social Network Analysis and Mining. Springer Science and Business Media LLC. http://doi.org/10.1007/s13278-020-00696-x

[17] Garg, S., & Sharma, D. K. (2024, April). Fake news detection in the Hindi language using multi-modality via transfer and ensemble learning. Internet Technology Letters. Wiley. http://doi.org/10.1002/itl2.523

[18] Shaikh, J., & Patil, R. (2020, December 16). Fake News Detection using Machine Learning. 2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC). IEEE. http://doi.org/10.1109/isssc50941.2020.9358890

[19] Ahmad, I., Yousaf, M., Yousaf, S., & Ahmad, M. O. (2020, October 17). Fake News Detection Using Machine Learning Ensemble Methods. (M. I. Uddin, Ed.), Complexity. Hindawi Limited. http://doi.org/10.1155/2020/8885861

[20] Anke Meyer-Baese, and Volker Schmid. 2014. "Foundations of Neural Networks." Elsevier EBooks, January, 197–243. https://doi.org/10.1016/b978-0-12-409545-8.00007-8.

[21] "Precision and Recall in Classification Models | Built In." 2022. Built In. 2022. https://builtin.com/data-science/precision-and-recall#:~:text=In%20machine%20learning%2C%20precision%20and,only%20the%20relevant%20data%20points.

[22] "Classification: Accuracy, Recall, Precision, and Related Metrics." 2024. Google for Developers. 2024. https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall.