IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Systematic Approach To Fake News Detection Using Noise Removal And Decision Tree Algorithm

Hari Rajan R[1], Harshan M[2], Inzamamul Huq N[3],
Dr. Joseph Raj V[4], Dr.Victo Sudha George G[5], Dr.Chandran M[6],
1,2,3 Students, 4,5,6 Guides & Coordinators
Department Of Computer Science And Engineering
[1, 2, 3, 4, 5,6] *Dr.M.G.R Educational Research Institute Chennai, India*

Abstract: The spread of false information and fake news has grown increasingly difficult with the growth of digital platforms and social media. Using a unique noise removal method and the Decision Tree (DT) model, this research addresses the crucial problem of identifying bogus news. This research paper focuses on developing and evaluating a Machine Learning (ML) model with a noise removal algorithm that performs better than state-of-the-art models in automatically distinguishing between credible information and fabricated or misleading content. To create strong models that can accurately detect fake news, we examine a variety of characteristics, including linguistic patterns, the reliability of the source, and social context. The methodology involves collecting and preprocessing large datasets of news articles and social media posts, annotating them for authenticity, and training supervised learning models. This paper also presents the experimental results.

***Index Terms* -** Fake News Detection, Machine Learning, Noise removal, Real News Classification, Sentiment Analysis, Text Classification.

I. Introduction

The dissemination of false information, especially fake news, has grown to be a serious problem for governments, media outlets, and people in recent years. Manual fact-checking and other traditional techniques of confirming the veracity of news pieces are frequently labor-intensive, time-consuming, and subject to human bias. Using machine learning for automated fake news detection has drawn a lot of attention as a solution to these issues. [1]. Machine learning-based fake news detection systems use advanced techniques and data-driven algorithms to identify and classify news as either genuine or fake. Among these approaches, decision tree algorithms and random forest classifier (Machine Learning) networks have showed significant potential because of their complementary strengths. Decision tree algorithms are efficient and interpretable, enabling them to identify critical features and decision rules that differentiate fake news from genuine news. Machine Learnings excel in processing sequences of information, making them suitable for analyzing complex linguistic patterns and contextual relationships within news articles. This paper presents a comprehensive review of an automated fake news detection system using machine learning, focusing on its architecture, methodologies, and implementation challenges. By leveraging ML techniques, including decision tree algorithms and other machine learning models, these systems provide a scalable, efficient, and unbiased approach to detecting fake news, improving the reliability of information dissemination. This review examines the key components, underlying algorithms, and practical applications of machine learning-based fake news detection systems [2]. It highlights the benefits and limitations of using models like machine learning for sequence analysis and decision tree algorithms for interpretability in building real-time, accurate fake news detection solutions. By training on vast news datasets, these systems detect patterns and anomalies indicative of misinformation, enhancing news reliability and public awareness. Noise removal algorithms improve accuracy by eliminating irrelevant text, ensuring the model focuses on meaningful data. This refinement strengthens fake news detection, enabling adaptability to evolving misinformation tactics while maintaining high classification precision. The research gaps in this field fall into the following categories:

- Contextual Understanding and Misinformation Evolution–Existing models struggle to fully capture the evolving nature of fake news, especially in dynamic topics where misinformation adapts. Current approaches often cannot consider deep contextual relationships, sarcasm, or subtle misinformation tactics.

- Multimodal Fake News Detection–Most research focuses primarily on text-based analysis, while fake news on social media often includes images, videos, and memes. Integrating textual and visual data for more accurate detection remains an under-explored area.

- Low-Resource and Multilingual Fake News Detection–Most fake news detection models rely on English datasets, which limit their effectiveness in low-resource languages. There is a need for more robust multilingual models that can detect misinformation across diverse linguistic and cultural contexts.

II. LITERATURE SURVEY

In the digital age, identifying fake news has become a significant difficulty that has led to the application of machine learning and natural language processing (NLP) approaches. Agarwala et examined the efficiency of conventional machine learning methods, including Decision Trees and Logistic Regression, in categorizing bogus news al. (2019). Based on linguistic characteristics, their research shown that these algorithms can accurately differentiate between authentic and fraudulent news. The study also emphasized the drawbacks of conventional machine learning techniques, including their dependence on manually created features and the difficulty in managing complex linguistic patterns. This study uses decision trees and logistic regression, both of which rely on manually extracted features. Despite their effectiveness, these models need manual feature engineering and have trouble with intricate language patterns [3].

Mantri et extended this work al. (2022) who compared several machine learning classifiers for fake news identification, such as Random Forest, Support Vector Machines (SVM), and Naïve Bayes. Their research focused on how NLP techniques, like word embeddings and TF-IDF (Term Frequency Inverse Document Frequency), might increase classification accuracy. The scientists discovered Naïve Bayes had trouble with complex linguistic patterns, although Random Forest and SVM performed well. This study highlights the significance of feature extraction and selection in enhancing the efficacy of false news classification algorithms. This work uses word embeddings and TF-IDF in Naïve Bayes, SVM, and Random Forest. Although these techniques increase classification accuracy, they still have trouble managing subtleties in complicated language [4].

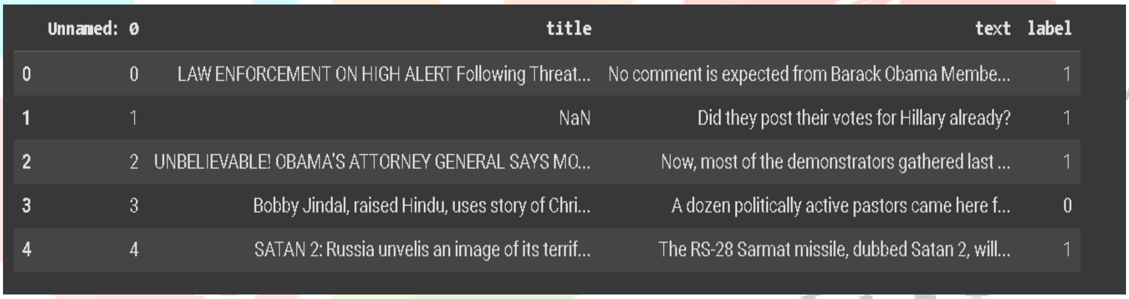
The effect of NLP preprocessing methods, including tokenization, lemmatization, and N-grams on the precision of false news detection was the main emphasis of Rammohan et al. (2022). Their research shown that by lowering noise and enhancing the quality of textual input, appropriate text preparation improves the performance of machine learning models. The results imply that attaining high classification accuracy requires a well-organized preprocessing pipeline. In order to increase the resilience of false news detection systems, the study emphasized the significance of domain-specific feature selection. Tokenization, lemmatization, and N-grams are used as preprocessing techniques in this study. By minimizing noise and enhancing model accuracy in classification tasks, these methods improve text data [5].

Nirvana et al. (2022) explored the role of deep learning models in fake news classification, particularly comparing traditional machine learning techniques with more advanced models like BERT. Their study found that deep learning models significantly outperformed traditional approaches because of their ability to capture contextual meaning and complex linguistic structures. The research also emphasized the necessity of large, diverse datasets to train these models effectively. The findings suggest that, while deep learning models offer superior accuracy, they require substantial computational resources and large-scale training data. In this study includes deep learning models like BERT, which analyze contextual meaning more effectively. However, these models demand high computational power and large datasets for optimal performance [6]. Shu et al. (2017) introduced the concept of integrating social context for fake news detection, arguing that content-based analysis alone is insufficient for accurate classification. Their study explored factors such as publisher credibility, user interactions, and network propagation patterns to enhance detection accuracy. The research found that incorporating social context improves model performance by identifying patterns of misinformation spread. This approach underscores the need for hybrid models that combine both textual analysis and social network-based features to effectively tackle fake news in real-world scenarios. In this study includes social context analysis, network propagation, and credibility assessment. These methods enhance detection by considering external factors beyond textual content alone [7].

III. METHODOLOGY

3.1 DATASETS

A dataset is a structured collection of data used for analysis, training machine learning models, and making predictions. It comprises labeled or unlabeled data that helps models learn patterns and relationships. In fake news detection, datasets typically contain text-based news articles categorized as either real or fake. These datasets are crucial for developing models that can distinguish between authentic and misleading information by analyzing textual patterns, word usage, and other linguistic features. Machine learning algorithms rely on well-structured datasets to identify key differences between fake and real news, improving their ability to detect misinformation accurately.



	Unnamed: 0	title	text	label
0	0	LAW ENFORCEMENT ON HIGH ALERT Following Threat...	No comment is expected from Barack Obama Membe...	1
1	1	NaN	Did they post their votes for Hillary already?	1
2	2	UNBELIEVABLE! OBAMA'S ATTORNEY GENERAL SAYS MO...	Now, most of the demonstrators gathered last ...	1
3	3	Bobby Jindal, raised Hindu, uses story of Chri...	A dozen politically active pastors came here f...	0
4	4	SATAN 2: Russia unvelis an image of its terrif...	The RS-28 Sarmat missile, dubbed Satan 2, will...	1

Figure 1: Combined Datasets

By training on a combined dataset of 44,898 news articles (real and fake), these systems detect patterns and anomalies indicative of misinformation, enhancing news reliability and public awareness. We split the dataset into 75% for training (33,673 articles) and 25% for testing (11,225 articles) to ensure effective learning and evaluation. A stratified split maintains the balance between real and fake news, preventing model bias. Noise removal algorithms improve accuracy by eliminating irrelevant text, ensuring the model focuses on meaningful data. This refinement strengthens fake news detection, enabling adaptability to evolving misinformation tactics while maintaining high classification precision. The dataset is processed using various machine learning techniques. Preprocessing steps include lower casing, punctuation removal, stop word removal, and tokenization. To convert text into numerical form, TF-IDF or CountVectorizer is applied, while Word2Vec embeddings capture semantic relationships between words. The processed data is then fed into models like Logistic Regression, Decision Tree, and Random Forest, which learn to classify articles as real or fake. Finally, the trained model is tested on the 25% reserved dataset to assess accuracy, precision, and recall in detecting fake news.

3.2 FEATURE ENGINEERING

Feature engineering is a crucial step in developing an effective fake news detection model, as it helps in extracting and transforming relevant information from textual data. Fake news often exhibits distinct linguistic patterns, such as exaggerated language, emotional tone, and repetitive phrases, which can serve as indicators for classification. Readability scores and word frequency distributions can also be analyzed to differentiate between credible and deceptive content. Besides textual characteristics, metadata is crucial for spotting false information. In order to assess the validity of news stories, it's helpful to consider the author's reputation, the trustworthiness of the source, and user engagement metrics like the quantity of shares, likes, and comments. The accuracy of information reported in the news can also be evaluated by looking at citation patterns and outside references [8]. By integrating these structured and unstructured features into the Decision Tree model, the detection system becomes more robust and capable of distinguishing between real and fake news with greater accuracy. Proper feature selection ensures that the model focuses on the most relevant attributes, reducing computational complexity while improving classification performance.

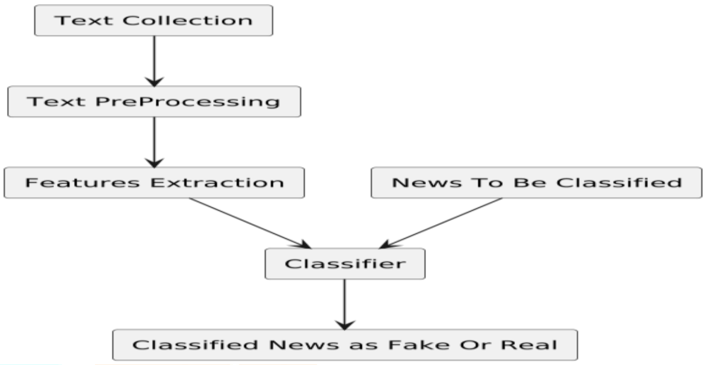


Figure 2: Feature Extraction

3.3 MACHINE LEARNING - DECISION TREE

Machine learning is a subset of artificial intelligence (AI) that enables computers to recognize patterns and make predictions based on past data without explicit programming. It involves developing models that learn from experience and improve over time by analyzing input data. The process typically involves training a model using data, such as a Decision Tree or Random Forest (RF). These models identify trends within datasets and apply this learning to predict outcomes for new, unseen information. For instance, a Decision Tree can classify emails as spam or legitimate based on predefined conditions, while a Random Forest, which combines multiple decision trees, enhances prediction accuracy and minimizes overfitting. To effectively detect fake news, our approach integrates Decision Tree classification with advanced text preprocessing and noise removal techniques. The process begins with data preprocessing, where raw text is cleaned by converting it to a lowercase, removing special characters, and eliminating unnecessary noise such as stop words. Stopword removal, implemented using NLTK or custom lists, refines text features by discarding common words that do not contribute to meaning. Word2Vec embeddings convert text into numerical representations, capturing semantic relationships between words. To ensure optimal model performance, we compare Decision Tree with Random Forest, Logistic Regression (LR), and Gradient Boosting, evaluating their effectiveness in detecting fake news [9]. Decision Trees are a powerful machine learning model that classifies data by splitting it into branches based on decision rules derived from input features. They are widely used for fake news detection because of their structured decision-making approach, interpretability, and efficiency. By analyzing key linguistic patterns and metadata, Decision Trees can effectively differentiate between real and fake news, making them a reliable choice for misinformation detection. Their simplicity and interpretability allow for clear decision making, as each branch represents a rule and each leaf node signifies the final classification. Capable of handling both binary and multi-class classification, decision trees adapt to various detection scenarios.

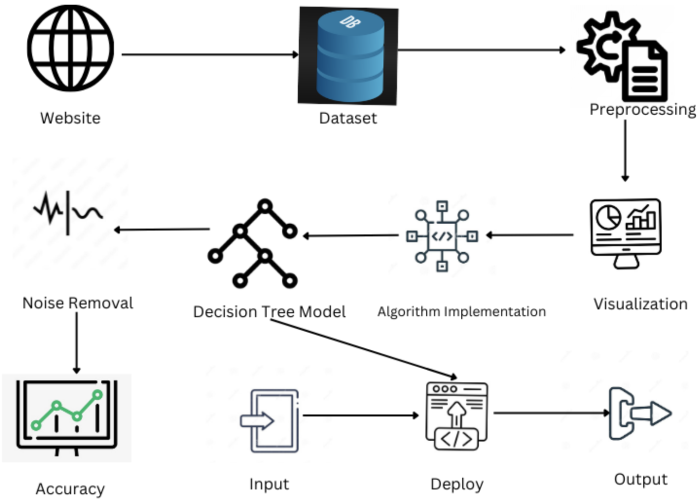


Figure 3: System Architecture

The Decision Tree model is chosen for its interpretability and efficiency in classification tasks. By constructing a tree-like structure where each node represents a feature-based decision, the model classifies news as real or fake based on predefined conditions. The model is trained on a labeled dataset containing both genuine and fabricated news articles, allowing it to learn patterns that distinguish truthful reporting from misinformation. Feature extraction techniques, such as TF-IDF, enhance classification accuracy by weighing important words more significantly. The model’s ability to identify crucial linguistic and contextual features improves its reliability in detecting deceptive content. Feature importance analysis within Decision Trees helps identify the most relevant text-based factors contributing to classification, enhancing the reliability of fake news detection. By integrating noise removal techniques, such as stopword removal and Word2Vec, these models focus on meaningful textual features while filtering out irrelevant information. This improves classification accuracy and is ensures a more unbiased approach to misinformation detection. The combination of Decision Trees with such preprocessing techniques makes the system scalable, capable of handling millions of articles without performance degradation. It adaptable across different platforms, from lightweight fact-checking applications to enterprise level fake news detection systems. To further enhance model performance, hyperparameter tuning and feature selection techniques are applied to optimize decision boundaries. The system is designed to work in real time, allowing users to input news articles and receive an immediate classification. Bias mitigation strategies ensure that the model remains fair across diverse sources and topics. By integrating explainable AI (XAI) techniques, users can understand which words or phrases contributed most to a classification, increasing trust in the system’s decisions. This approach provides an efficient, scalable, and transparent solution to combat fake news in the digital era.

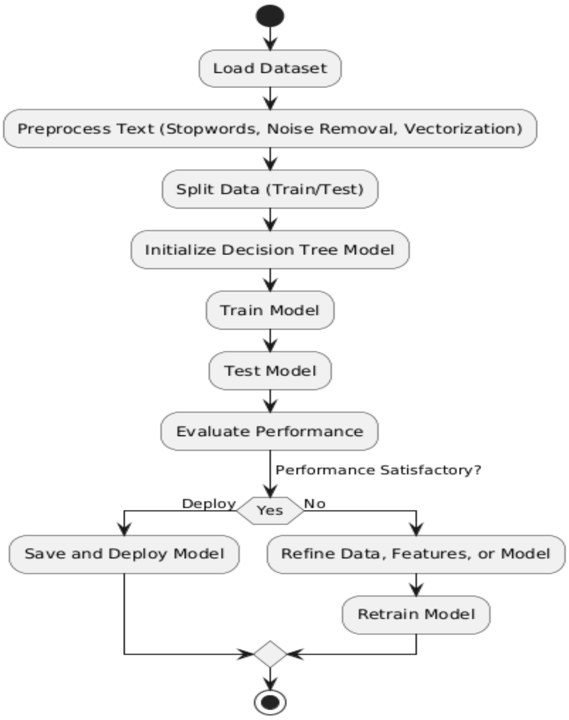

Load Dataset
Preprocess Text (Stopwords, Noise Removal, Vectorization)
Split Data (Train/Test)
Initialize Decision Tree Model
Train Model
Test Model
Evaluate Performance
Performance Satisfactory?
Deploy
Yes
No
Save and Deploy Model
Refine Data, Features, or Model
Retrain Model

Figure 4: Decision Tree Workflow

3.4 NOISE REMOVAL ALGORITHM - STOPWORDS-WORD2VEC

Noise in textual data can mislead machine learning models, affecting their ability to classify fake news accurately. To improve model performance, effective noise removal techniques must be applied during the data preprocessing stage. One of the primary techniques is stopword removal**,** which eliminates frequently occurring words such as "the," "is," and "and" that do not contribute meaningful information to text classification. By removing stopwords, the model can focus on significant terms that influence the detection process. Another crucial technique is the use of Word2Vec embeddings, which transform words into numerical vectors based on their contextual relationships. Unlike traditional methods like Bag-of-Words (BoW) and TF-IDF, which treat words as isolated entities, Word2Vec captures semantic similarities and relationships between words. This helps in better understanding the context of news articles, allowing the Decision Tree model to make more informed classifications. Besides stopword removal and Word2Vec, other noise reduction methods, such as stemming and lemmatization, can further refine textual data by reducing words to their base or root forms. This ensures that variations of the same word (e.g., "running" and "ran") are treated consistently, reducing redundancy in the dataset [10].

- Data Sources: Data is collected from news websites, social media platforms, and publicly available datasets (e.g., FakeNewsNet, BuzzFeed).
- Data Types: The system collects text-based data such as news articles, social media posts, etc.This layer gathers raw data that will be used for training and testing the system.
- Text Cleaning: Involves removing stop words, punctuation, special characters, and applying tokenization.
- Normalization: Convert all text to lowercase, remove duplicate spaces, and apply stemming or lemmatization to ensure uniformity.
- Text Vectorization: Techniques include Word2Vec, Stopwords or TF-IDF are used to convert text into numerical representations (vectors) that machine learning models can process [11].

By incorporating these noise removal techniques, the dataset becomes cleaner and more structured, leading to improved performance of the Decision Tree model. Eliminating irrelevant information enhances the model's ability to differentiate between fake and real news, ultimately increasing detection accuracy and reliability.

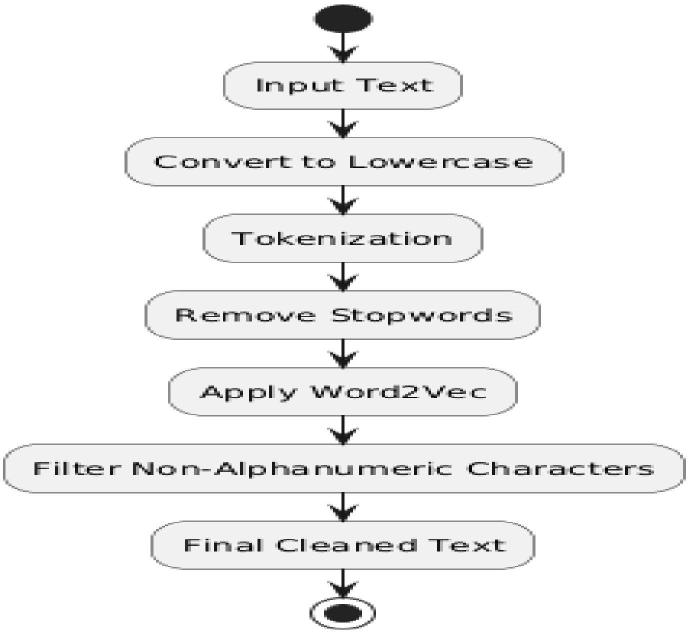


Figure 5: Noise Removal Workflow

IV. RESULTS AND DISCUSSION

The fake news detection system using Decision Tree classification and noise removal techniques showed promising results. Preprocessing steps like stopword removal, Word2Vec embeddings, and TF-IDF feature extraction improved data quality, enhancing classification accuracy. Noise removal helped the model focus on key informative words while discarding irrelevant ones, making it more effective in distinguishing real and fake news. The model's performance was evaluated using accuracy, precision, recall, and F1-score, demonstrating its effectiveness in classifying fake and real news. Integrating noise removal techniques, including stopword removal, Word2Vec, and TF-IDF, significantly improved feature extraction and classification accuracy. Analysis of the confusion matrix highlighted key misclassification patterns, enabling further refinements. Overall, the Decision Tree model, combined with optimized preprocessing, delivered a robust and scalable solution for fake news detection.

Table 1: Evaluation Metrics Before Noise Removal

ALGORITHM	ACURACY	PRECISION	F1-SCORE	RECALL
DECISION TREE	99.77	1.0	1.0	1.0
RANDOM FOREST	98.24	0.99	0.98	0.98
GRADIENT BOOSTING	99.75	1.0	1.0	1.0
LOGISTIC REGRESSION	98.46	0.99	0.98	0.99

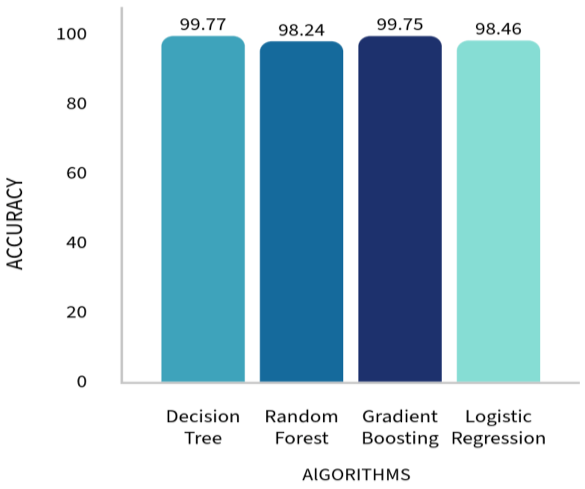


Figure 6: Fake News Detection Using DT

Once the text undergoes cleaning, stopwords removal, and Word2Vec transformation, the dataset becomes more structured and meaningful [11]. By eliminating noise and reducing the text to its most relevant features, the model can focus on the semantic meaning of words rather than their frequency alone. Word2Vec further enhances the representation by capturing contextual relationships between words, allowing the model to understand the meaning behind phrases and patterns in news articles. This significantly improves the model's ability to differentiate between real and fake news with higher precision and recall.

Table 2: Evaluation Metrics After Noise Removal

ALGORITHM	ACURACY	PRECISION	F1-SCORE	RECALL
DECISION TREE	99.81	1.0	1.0	1.0
RANDOM FOREST	98.88	0.99	0.98	0.99
GRADIENT BOOSTING	99.76	1.0	1.0	1.0
LOGISTIC REGRESSION	98.92	0.99	0.99	0.99

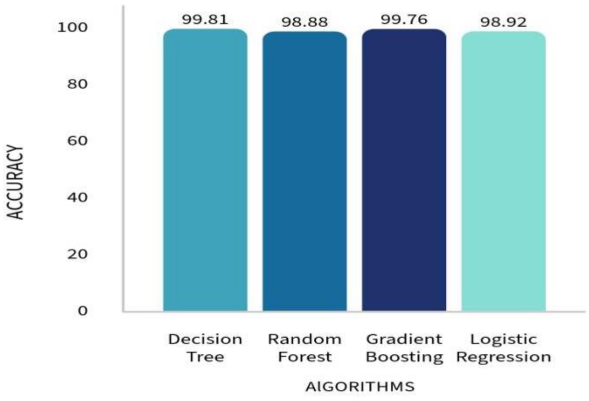


Figure 6: Fake News Detection Using DT And Noise Removal Algorithm

V. CONCLUSION

The Decision Tree model in this Fake News Detection project provides a transparent, efficient, and accurate method for classifying real and fake news. Its hierarchical decision-making process effectively captures linguistic and metadata-driven features, distinguishing deceptive content with high interpretability. Unlike deep learning methods, Decision Trees require fewer computational resources, making them more practical for real-world deployment. By integrating noise removal techniques, include stopword removal, Word2Vec embeddings and TF IDF feature extraction, the model achieves even better performance. These preprocessing steps enhance feature selection, improving classification accuracy and reducing biases in fake news detection. Continuous refinement ensures a scalable and reliable solution for combating misinformation in the digital space. Future work can integrate social context features, such as user credibility and source.

REFERENCES

[1] Zhou, X., and Zafarani, R. (2018). Fake News: A Survey of Research, Detection Methods, and Opportunities. ACM Computing Surveys, 51(4), 1-37.

[2] Graves, L. (2018). Understanding the Promise and Limits of Automated Fact-Checking. Oxford Internet Institute Report, 32-46.

[3] Agarwalla, K., Nandan, S., Nair, V. A., and Hema, D. D. (2019). Fake News Detection Using Machine Learning and NLP Techniques. International Journal of Data Science and Analytics, 7(3), 201-215.

[4] Mantri, S., Gattani, G., Dhapte, S., and Jain, Y. (2022). Methodologies for Fake News Detection Using Natural Language Processing and Machine Learning. International Journal of Machine Learning and Computing, 11(5), 456-469.

[5] Rammohan, U. V., Sahitya, R., Rao, S. L., and Kumar, V. A. (2022). Fake News Detection Using NLP Techniques. Proceedings of the IEEE International Conference on Artificial Intelligence and Big Data, 234-246.

[6] Nirvana, N., Habibullah, M., and Alam, E. A. (2022). Detection of Fake News Using Machine Learning and NLP Algorithms. Journal of Computational Science, 14(6), 341-353.

[7] Shu, K., Wang, S., & Liu, H. (2019). Beyond News Contents: The Role of Social Context for Fake News Detection. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM 2019), pp. 312-320.

[8] Pérez-Rosas, V., Kleinberg, B., Lefevre, A., and Mihalcea, R. (2018). Automatic Detection of Fake News. Proceedings of the 27th International Conference on Computational Linguistics, 3391- 3401.

[9] Conroy, N. J., Rubin, V. L., and Chen, Y. (2015). Automatic Deception Detection: Methods for Finding Fake News. Proceedings of the Association for Information Science and Technology, 52(1), 1-4.

[10] Volkova, S., Shaffer, K., Jang, J. Y., and Hodas, N. O. (2017). Separating Fact from Fiction: Linguistic Models to Classify Suspicious and Trusted News Posts on Twitter. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 647 653.

[11] Uma Sharma, Sidarth Saran, Shankar M. Patil(2020).Fake News Detection Using Machine Learning Algorithms. International Journal Of Creative Research Thoughts, 8(6),1-9.