Fake News Detection Using ML Algorithms

Kiran Sree Pokkuluri[1](✉), Atluri Dhatri Sai Subha[1], Ch. Phaneendra Varma[1], Ramesh Babu Gurujukota[1], S. S. S. N. Usha Devi N[2], P. B. V. Raja Rao[1], Nagaraju Pamarthi[1], and P. J. R. Shalem Raju[1]

[1] Department of Computer Science and Engineering, Shri Vishnu Engineering College for Women, Bhimavaram, India
drkiransree@gmail.com

[2] Department of Computer Science and Engineering, University College of Engineering, JNTU Kakinada, Kakinada, India

Abstract. In response to the growing threat of fake news in our digital era, this paper aims to propose a machine learning system to identify the fake news. Its primary aim is to accurately discern between authentic news articles and deceptive misinformation. The urgency for such a solution arises from the significant consequences of fake news, including the erosion of trust, societal polarization, and potential safety risks. This paper employs machine learning algorithms to analyses textual and contextual elements within news articles. This entails data collection, pre-processing, and model training, followed by a comprehensive evaluation of the model's performance against established benchmarks. In conclusion, this project is dedicated to fortifying the ongoing struggle against fake news, safeguarding the reliability of information sources, and nurturing a society that thrives on accurate and verified information. The proposed ML system is compared with the baseline methods and results are plotted using various parameters precision, sensitivity, specificity, accuracy and f1score. The proposed model exhibits an accuracy of 96.65% which is promising when compared with existing literature.

Keywords: Machine Learning · Decision Tree · Gradient Boos · Random Forest · Fake News

1 Introduction to Fake News and ML Algorithms

Fake news can be identified as false or deceptive information which may manipulate many readers. This news can be circulated through various media channels, social platforms, and online publications etc. Even Fake news may twist the story or scenario. Let's imagine a fake news story that is created to manipulate public opinion during a heated election season [1]. This may affect the election outcomes by manipulating the people by exploiting their psychological and emotional triggers. Fake news can originate from various sources, including social media, websites, and even traditional media outlets. With the ease of content creation and dissemination on the internet, identifying the credibility of sources has become more complex [2].

P. Pareek et al. (Eds.): IC4S 2024, LNICST 599, pp. 37–44, 2025.
https://doi.org/10.1007/978-3-031-77081-4_4

This misinformation can have serious real-world consequences like promoting harmful practices and also often preys on people's emotions, biases, and fears. So detecting the fake news is so important. The impact of fake news extends beyond individual misinformation. It can influence public opinion, contribute to the spread of fear and panic, and even impact political processes and elections [3]. Addressing fake news is essential for maintaining a healthy information ecosystem. To combat this epidemic of misinformation, various approaches, including machine learning (ML) techniques like linear regression, Decision Tree Classification, Grading Boosting Classifier and Random Forest Classifier have gained prominence in recent years [4].

Machine learning is a branch of artificial intelligence that helps computers to learn from data to make predictions or decisions without being explicitly programmed. The core idea behind this technique is to enable computers to improve their performance [5] on a specific task through experience or exposure to data.

As part of the architecture for detecting fake news, ML models and algorithms are trained to differentiate between authentic and fake news using metadata. Large-scale datasets of news stories, social media messages, and other textual content are analyzed by these algorithms in order to spot linguistic oddities, misleading patterns, and possible sources of false information [6].

2 Literature Survey

Researchers have employed techniques such as Support Vector Machines (SVM) [7], Naïve Bayes classifiers, and Deep Neural Networks [8] to classify news articles as fake or real. For instance many researchers implemented a fake news detection system using Naïve Bayes Classification and achieved an accuracy of up to 93.5%. Similar results have been reported in other studies, highlighting the effectiveness of machine learning. Moreover, ensemble methods like AdaBoost and Gradient Boosting have also demonstrated promising results in fake news detection, combining the strength of multiple weak classifiers to improve overall accuracy [9]. The diversity of algorithms employed in this field underscores the ongoing efforts to explore various approaches to enhance the accuracy and reliability of fake news detection systems [10, 11].

This work investigates data mining methods for identifying false information on social media platforms [12]. It discusses feature engineering, sentiment analysis, and the application of machine learning algorithms. Recurrent neural network (RNN)-based approach to detect rumors on microblogging platforms [13] is discussed in detail. It emphasizes the temporal aspect of information spread for effective rumor detection. This review provides an overview of existing approaches to early detection of fake news [14]. It covers methods based on content analysis, user behavior, and social network structure. The paper presents a comprehensive survey of techniques for identifying and mitigating fake news. It categorizes approaches into content-based, network-based, and hybrid methods [15]. In order to detect fake news, the study investigates the application of deep learning techniques, such as convolutional neural networks (CNNs) and long short-term memory networks (LSTMs), showing how well these methods can capture intricate patterns [3, 10].

3 Design of ML System for Fake News Detection

The block diagram of the design is explained in the Fig. 1.

3.1 Data Preprocessing

We have chosen a labelled dataset containing both genuine and fake news articles. Ensure that the dataset is balanced and representative of the problem. Remove unnecessary symbols, punctuation, and stop words from the text data before processing it. To normalize the text, use lemmatization or stemming.

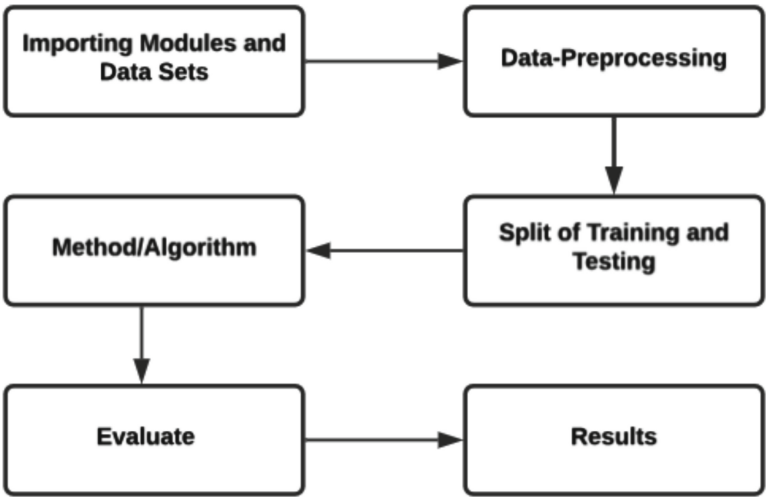

Fig. 1. Design of the Proposed Classifier

3.2 Feature Extraction

We have used Text Vectorization, i.e. Convert the text data into numerical vectors using techniques like TF-IDF (Term Frequency-Inverse Document Frequency) or word embedding's. Consider incorporating additional features such as sentiment analysis scores, source credibility, and other relevant metadata.

3.3 Split of Data

For an accurate assessment of the model's performance, separate the dataset into training and testing sets.

3.4 Methods/Algorithms and Evaluation

This research uses three machine learning methods to distinguish true and fake news. They are Decision Tree Classification, Gradient Boosting Classifier and Random Forest Classifier.

a) **Decision Tree Classification:** For categorization tasks, decision trees are a supervised learning technique. They construct a framework that resembles a tree, with branches standing in for potential outcomes and nodes for features. Decision trees are helpful for making judgements based on a set of conditions since they are interpretable and can handle both numerical and categorical data.
b) **Gradient Boosting Classifier:** Gradient boosting is an ensemble learning strategy that builds a powerful predictive model by aggregating the predictions of several weak learners, usually decision trees. By concentrating on the errors committed in prior rounds, iteratively enhances the model's performance and achieves high predicted accuracy. Gradient boosting has a reputation for being successful in a variety of machine learning contests.
c) **Random Forest Classifier:** Random Forest is another ensemble learning method that builds multiple decision trees and combines their predictions. It introduces randomness by selecting subsets of data and features for each tree, reducing overfitting and improving generalization. Random Forest is robust, works well with large datasets, and provides feature importance scores for understanding variable importance in classification tasks.

Metrics precision, F1 score, accuracy and precision are used for evaluation. Additionally, techniques like cross-validation and confusion matrix analysis help gauge model robustness and performance across different datasets. The effectiveness of algorithms heavily relies on feature selection, dataset quality, and algorithm optimization. Continuous refinement and validation against evolving fake news patterns are crucial for ensuring the reliability of detection systems in countering misinformation.

4 Experimental Results and Comparison

We have collected 48,635 datasets from kaggle [16]. Proposed model is able to classify news stories as true or fraudulent with moderate to high accuracy uptdo 98%. The model's recall, F1-score, and accuracy all demonstrate how well it can identify false positives while lowering false negatives. Models are continuously improved and adjusted in order to handle new forms of fake news and keep strong detection skills in the fight against false information.To implement the proposed method, the different type of libraries used are numpy, pandas, sklearn, re, matplotlib and string. NumPy is used for numerical and mathematical operations in Python. Pandas is used for data manipulation and analysis.

Because it offers data structures like DataFrames and Series, working with structured data, data cleansing, and exploratory data analysis are made simple. A machine learning library called Scikit-learn is used for a number of tasks, such as model evaluation, regression, clustering, and classification. The 're' module is used for pattern matching and text manipulation. It allows you to search, extract, and manipulate text strings based on specific patterns. Matplotlib is a popular data visualization library that enables the

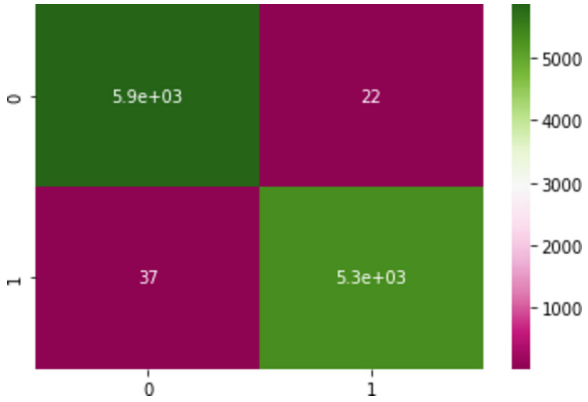

Fig. 2. Decision Tree Classification - Confusion Matrix

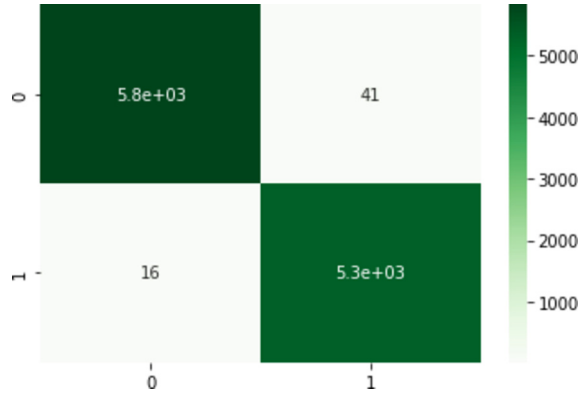

Fig. 3. Gradient Boosting Classifier- Confusion Matrix

creation of various types of plots and charts, such as line plots, bar charts, scatter plots, and more, for data visualization and analysis. The 'string' module provides a collection of string constants and helper functions to work with strings in Python, including operations like formatting, manipulation, and checking character properties.

The definition of the confusion matrix is as follows: True Positive is the quantity of messages that credible messages accurately classify. The number of messages that are accurately identified as unbelievable messages is known as the True Negative. The quantity of messages that believable messages wrongly classify as false positives is known as False Positives. The quantity of communications that are mistakenly categorised as improbable messages is known as False Negative. As shown in Figs. 2, 3, and 4, the precision, recall, F-measure, and accuracy are computed.

Table 1 and Fig. 5 reports the performance of proposed classifier with existing literature. RFA. RFA records highest with parameters Accuracy, Sensitivity and precision. Decision tree depicts a good performance with parameter F1 Score.

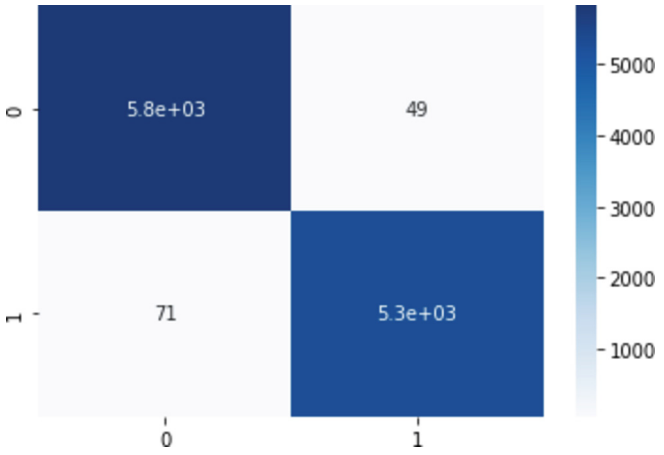

Fig. 4. Random Forest Classifier- Confusion Matrix

Table 1. Comparison of RFA with exiting literature

Method	Accuracy	F1 Score	Sensitivity	Specificity
RFA	96.65	0.9863	0.9963	0.9633
Gradient Boost	94.51	0.9670	0.9803	0.9523
Decision Tree	95.25	0.9623	0.9739	0.9795
Bayesian Class	95.01	0.9867	0.9692	0.9498

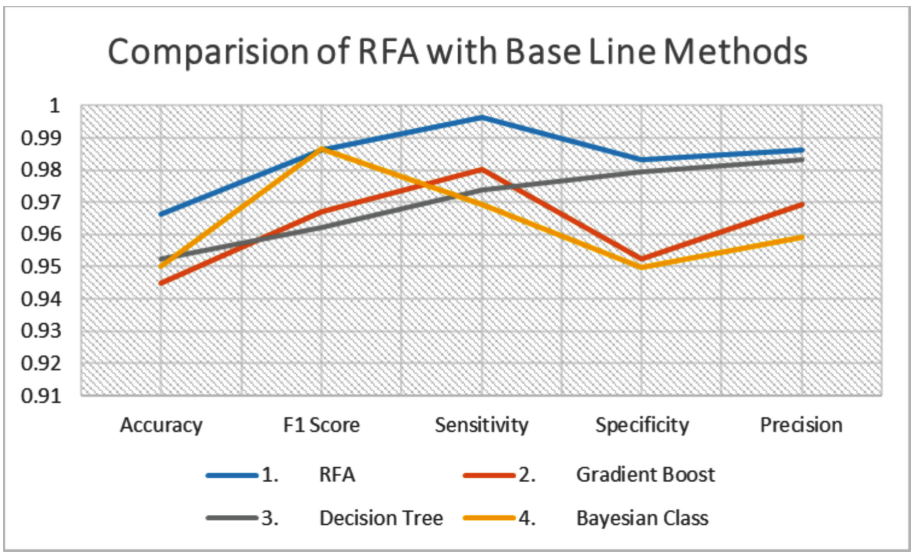

Fig. 5. Comparison of RFA with existing literature

5 Conclusion

Fake news is the difficult problem because it is the rumors it is too hard to identify the fact in contents. Based on the findings of the experiment reported in this research, machine learning techniques can reliably detect fake news. The results of the investigation showed that the system could differentiate between authentic and fake news with an accuracy of up to 0.9665. Future research can increase the accuracy by employing more search engines or methods for scraping, obtaining more information from social media, and utilizing a different algorithm to check for text similarity. Despite the promising results, challenges persist in the dynamic landscape of fake news, including the emergence of new deceptive techniques. Future research could focus on refining the models, incorporating additional features, and exploring real-time detection mechanisms to address the evolving nature of misinformation.

References

1. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: a data mining perspective. ACM SIGKDD Explor. Newsl. **19**(1), 22–36 (2017)
2. Pokkuluri, K.S., Usha Devi Nedunuri, S.S.S.N., Devi, U.: Crop disease prediction with convolution neural network (CNN) augmented with cellular automata. Int. Arab J. Inf. Technol. **19**(5), 765–773 (2022)
3. Altheneyan, A., Alhadlaq, A.: Big data ML-based fake news detection using distributed learning. IEEE Access 11 (2023)
4. Sree, P.K., Ramesh Babu, I., Usha Devi, N.S.S.S.N.: Investigating an artificial immune system to strengthen protein structure prediction and protein coding region identification using the cellular automata classifier. Int. J. Bioinform. Res. Appl. **5**(6), 647–662 (2009)
5. Nasir, J.A., Khan, O.S., Varlamis, I.: Fake news detection: a hybrid CNN-RNN based deep learning approach. Int. J. Inf. Manag. Data Insights **1**(1) (2021)
6. Sree, P.K., Babu, G.R., Raja Rao, P.B.V., Chintalapati, P.V., Prasad, M.: Fake news detection using cellular automata based deep learning. In: 2023 3rd International Conference on Computing and Information Technology (ICCIT), pp. 167–171. IEEE (2023)
7. Ahmad, I., Yousaf, M., Yousaf, S., Ahmad, M.O.: Fake news detection using machine learning ensemble methods. Complexity, 1–11 (2020)
8. Bahad, P., Saxena, P., Kamal, R.: Fake news detection using bi-directional LSTM-recurrent neural network. Procedia Comput. Sci. **165,** 74–82 (2019)
9. Shaikh, J., Patil, R.: Fake news detection using machine learning. In: 2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC), pp. 1–5. IEEE (2020)
10. Kaliyar, R.K., Goswami, A., Narang, P., Sinha, S.: FNDNet–a deep convolutional neural network for fake news detection. Cogn. Syst. Res. **61**, 32–44 (2020)
11. Sree, K., Babu, R., Devi, N.U.: Identification of promoter region in genomic DNA using cellular automata based text clustering. Int. Arab J. Inf. Technol. **7**(1), 75–78 (2009)
12. Pereira-Kohatsu, J.C., Quijano-Sánchez, L., Liberatore, F., Camacho-Collados, M.: Detecting and monitoring hate speech in Twitter. Sensors **19**(21) (2019)
13. Sudheer, M.S., Ganesh Reddy, K., Kiran Sree, P., Purushothama Raju, V.: An effective analysis on various scheduling algorithms in cloud computing. In: 2017 International Conference on Inventive Computing and Informatics (ICICI), pp. 931–936 (2017)

14. Fahad, N., et al.: Stand up against bad intended news: an approach to detect fake news using machine learning. Emerg. Sci. J. **7**(4), 1247–1259 (2023)
15. Mundra, S., Reddy, J., Mundra, A., Mittal, N., Vidyarthi, A., Gupta, D.: An automated data-driven machine intelligence framework for mining knowledge to classify fake news using NLP. ACM Trans. Asian Low-Resour. Lang. Inf. Process. (2023)
16. Fake News. https://www.kaggle.com/c/fake-news/data. Accessed 24 Mar 2024