IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 5, May 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205200 |

Fake News Detection Using Machine Learning Techniques

Rachi Bandewar, Prof Vijay Rakhade, Prof Ashish Deharkar

Department of Computer Science and Engineering, Shri Sai College of Engineering and Technology, Chandrapur, India

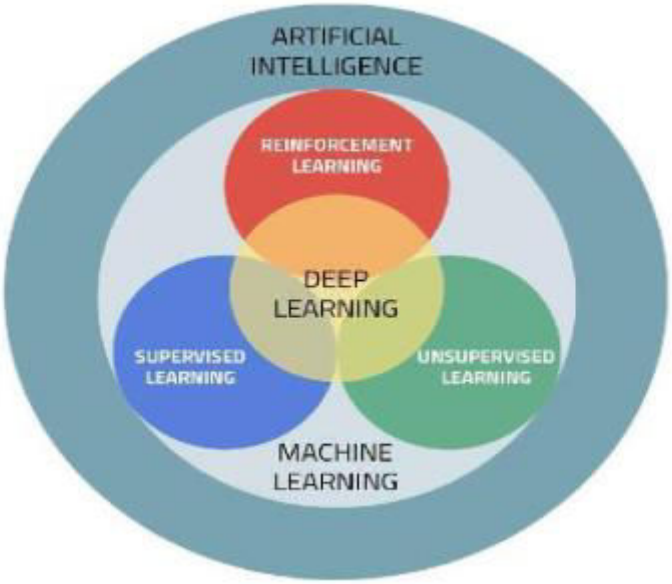
ABSTRACT: Fake News has become one of the major problems in the existing society. Fake News has high potential to change opinions, facts and can be the most dangerous weapon in influencing society. The proposed paper uses NLP techniques for detecting the 'fake news', that is, misleading news stories which come from non-reputable sources. By building a model based on a K-Means clustering algorithm, the fake news can be detected. The data science community has responded by taking actions against the problem. It is impossible to determine whether the news was real or fake accurately. So, the proposed project uses the datasets that are trained using the count vectorizer method for the detection of fake news and its accuracy will be tested using machine learning algorithms.

KEYWORDS: Machine Learning, Natural Language Processing

I. INTRODUCTION

Machine learning (ML) is the study of the statistical models and methods used by computers to do certain tasks devoid of explicit instructions and in favour of patterns and inference. As part of artificial intelligence, it is viewed. Without explicit instructions, machine learning algorithms construct a mathematical model using sample data, or "training data," in order to provide predictions or judgements. Computational statistics, which focuses on computer-aided prediction, and machine learning have a lot in common. Machine learning may benefit from the ideas, practises, and fields of application that come from the study of mathematical optimisation.

Fake news, to put it simply, is information that is untrue. whether or whether it is accurate. Fake news contains verifiable erroneous information. Many significant companies, even government agencies, are working to address issues related to false news. However, given that millions of articles are produced or purged every minute in this age, they are neither responsible nor humanely feasible because they rely on manual human detection. A machine learning algorithm that creates a trustworthy automated index score or rating for the authenticity of various publications and can assess whether the news is true or misleading may provide a solution to this problem**.**



International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205200 |

Natural Language Processing (NLP)

The study of how computers interact with human (natural) languages is known as natural language processing, or NLP, and it is a branch of computer science and artificial intelligence that focuses on instructing computers to efficiently analyse massive volumes of natural language data. In the fields of linguistics, computer science, information engineering, and artificial intelligence, natural language processing (NLP) studies how computers interact with human (natural) languages. Its major goal is to instruct computer programmers in how to study and analyse vast amounts of natural language

Fake News Detection

With the rising use of social media platforms, false news has become a severe problem in recent years. Finding fake news is a difficult problem that necessitates the use of several computer techniques, such as data mining, machine learning, and natural language processing. In this abstract, the current state of false news detection will be discussed, along with its challenges and potential solutions. Finally, it will consider how cutting-edge technology like blockchain and artificial intelligence may be used in the future to improve the efficiency and precision of fake news detection.

As a result, there is a larger than ever need for accurate and reliable techniques to distinguish fake news. The field of fake news detection has rapidly evolved as a result of researchers and engineers developing a number of techniques and tactics to identify and combat misleading information. These methods include human fact- checking by educated professionals as well as sophisticated computers that use machine learning to examine and classify news content. Automated processes are also a part of them. It is important to research and create fake news detection, but it is also a challenging and complex problem. The ability to recognise fake news requires knowledge of linguistic nuance, social and cultural contexts, and the complex network dynamics of online communication. Despite these challenges, work has been done to establish effective methods for spotting false news, and the area is still developing as new tools and technology are created.

II. OBJECTIVE

Our project's primary goal is to determine the veracity of news in order to determine if it is real or fake. The development of a machine learning model that would allow us to recognise bogus information. It can be difficult and difficult to identify fake news only based on its content since it is intentionally produced to influence readers to believe false information. By applying a range of methods and models, machine learning makes it easy to detect bogus news. Additionally, to examine the relationship between two words, we will apply deep learning-based NLP.

III. ALGORITHM FOR THE PROPOSED SYSTEM

Step 1: Pre-processing

▪ Load the dataset of news items with their labels, whether they are true or false;

▪ Clean the text by eliminating punctuation and stopwords

▪ Divide the dataset into training and testing sets.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205200 |

Step 2: Count Vectorization

▪ Count Vectorizer from the Sklearn toolkit may be used to transform text data into numerical data.

▪ Produce a document-term matrix showing the frequency of each word used in each document.
▪ Fit the Count Vectorizer using the training set, then convert the data.

▪ Utilise the testing set to change the data.

Step 3: TFIDF Vectorization

▪ Utilise the Tfidf Vectorizer in the Sklearn package to turn the text data into numerical data.

▪ Use the training set to fit the Tfidf Vectorizer and convert the data.

▪ Create a document-term matrix that depicts the significance of each word in each document.

▪ Utilise the testing set to change the data.

Step 4: Training the Models

▪ Utilise the data that has been modified by Count Vectorizer and Tfidf Vectorizer to train a variety of models, including Naive Bayes, Logistic Regression, Support Vector Machines (SVM), Random Forest, etc.

▪ Fit the models using the training set.

▪ Use the testing set to predict the news article labels.

▪ Determine each model's accuracy score using the actual and projected labels.

Step 5: Confusion Matrix

▪ The confusion matrix displays the amount of true positives, true negatives, false positives, and false negatives for each model, allowing you to assess each one's performance.

▪ Measurements like accuracy, recall, and F1-score may be calculated using the confusion matrix.

Step 6: Accuracy

▪ Determine each model's accuracy by comparing its predicted labels to its actual labels.

▪ The accuracy measures the proportion of news stories that were accurately identified as being true or false.

▪ Evaluate the accuracy of various models to find which one is most effective at spotting fake news.

Step 7: Representing the Output in Web Browser using Streamlit

▪ Use the Streamlit Python module to build an interactive web application for showcasing the outcomes of false news detection models.

▪ Create a user interface that clearly displays the confusion matrices, accuracy of each model, and other performance indicators.

▪ Provide tools that allow users to submit their own content for categorization and display the key terms and phrases used to categorize news items, among other capabilities.

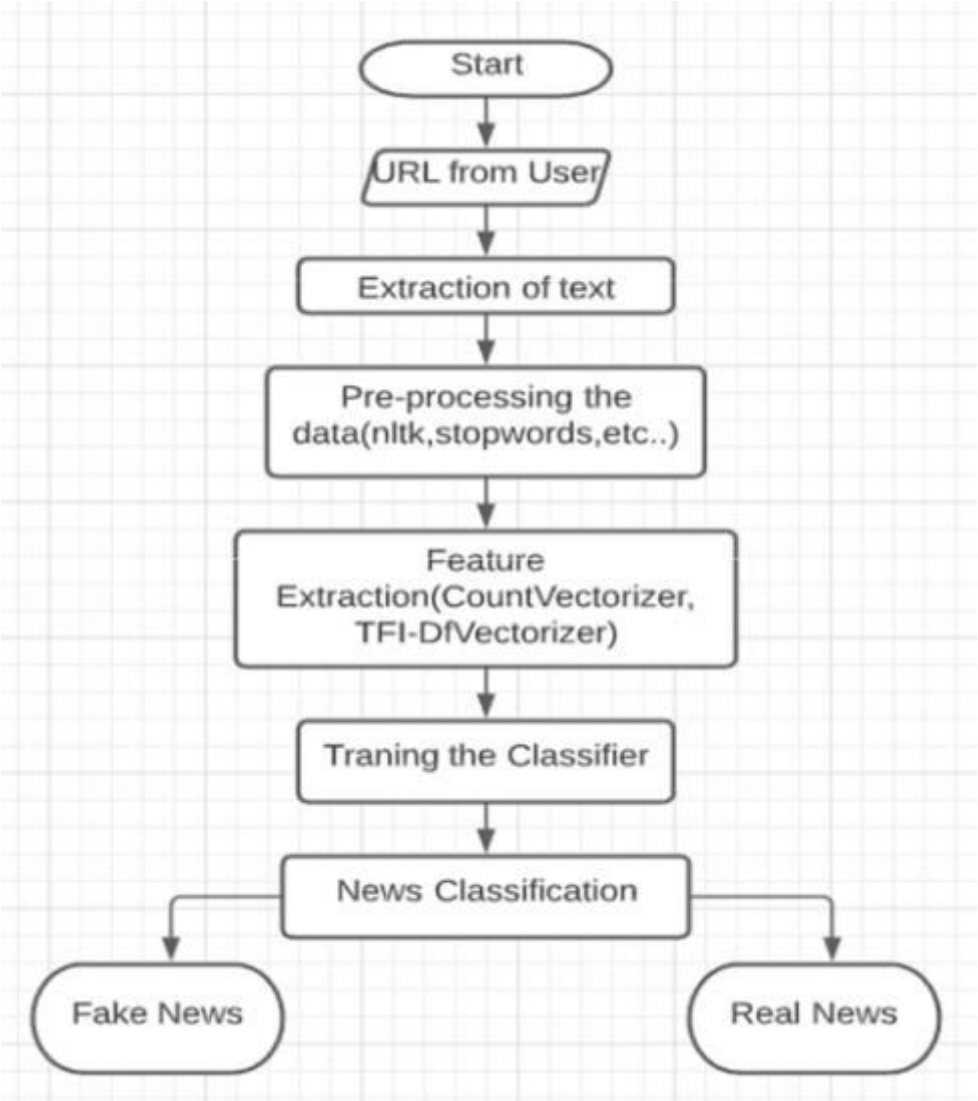
International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205200 |

IV. FLOWCHART



V. DESIGN OF PROJECT

Dataset: The first step is to collect or obtain a dataset of news articles, labeled as "fake" or "real". This dataset will be used to train and evaluate the performance of different fake news detection models.

Preprocessing: The dataset must now be cleaned up by eliminating any extraneous or irrelevant data, including stop words, punctuation, and digits. Additionally, the text may need to be normalised by making all characters lowercase and eliminating any special characters or symbols.

Count Vectorizer (BOW): The Bag-of-Words (BOW) format can be used to transform textual data into numerical characteristics after preprocessing the text. This entails building a matrix where each row represents a news item and each column represents a distinct term from the dataset. The value in each cell indicates how often the term appears in the related art.

Train-Test Split: Once we have the BOW matrix, we can split the data into training and testing sets. The training set will be used to train the fake news detection model, while the testing set will be used to evaluate the model's performance on new, unseen data.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

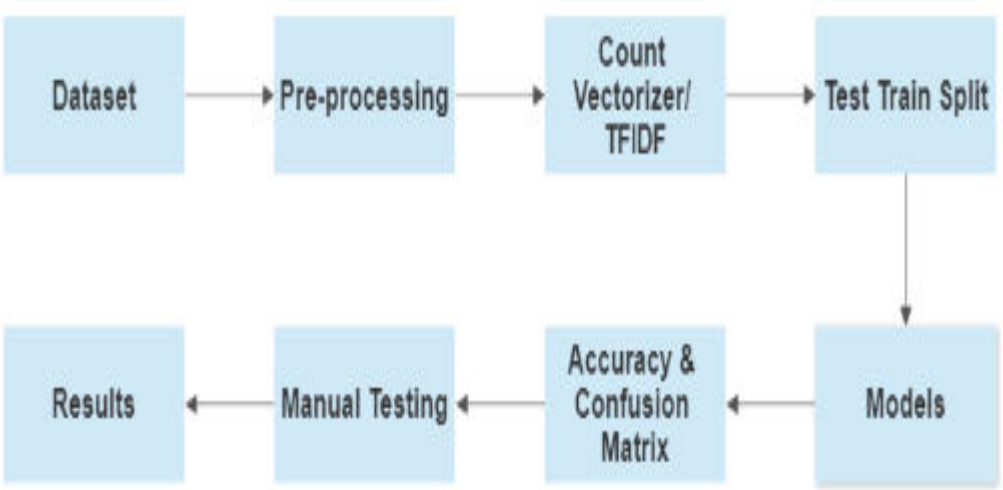
| DOI: 10.15680/IJIRCCE.2024.1205200 |

Text-to-vectors (TF-IDF): In addition to BOW, we can also express the textual data using the Term Frequency-Inverse Document Frequency (TF-IDF) representation. The frequency of the terms in each article as well as their frequency throughout the whole dataset is taken into consideration in this representation. Models: After obtaining the numerical features from the text data, several machine learning methods such as logistic regression, decision trees, or neural networks can be employed to train a fake news detection model. The objective of the model is to learn a function that can accurately classify news stories as either "real" or "fake" based on the derived attributes from the text.

Accuracy and Confusion Matrix: It's crucial to assess the false news detection model's performance on the testing set after we've trained it. By assessing its accuracy, precision, recall, and F1 score, we may do this. To see how many true positives, true negatives, false positives, and false negatives the model produces,we may also develop a confusion matrix.

Testing: We may use the model to categorise fresh and previously unheard news pieces as "real" or "fake" after assessing the model's performance. This entails applying the same feature extraction and preprocessing operations to the fresh data that we did during training. After that, we can apply the trained model to the cleaned-up data to provide a categorization label.

Streamlit library of python is used to represent the result in web browser where user input the news and algorithm tell that the news is “Real” or “Fake”



Models Applied And their Results

Support Vector Machine (SVM)

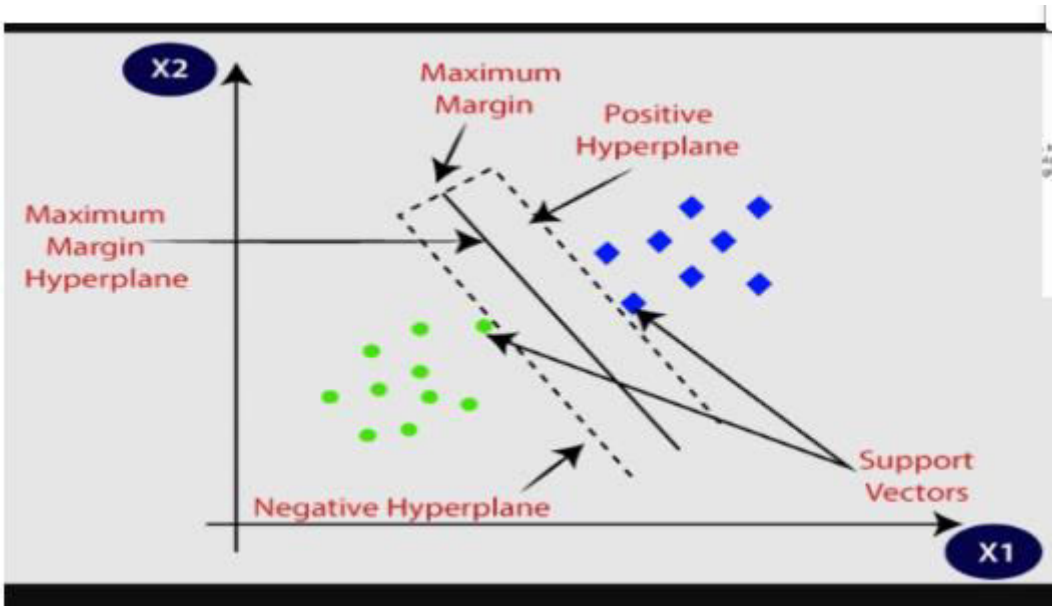
Classification and regression problems are resolved using Support Vector Machine, or SVM, one of the most used supervised learning techniques. It is mostly used, nevertheless, in Machine Learning Classification problems. SVM chooses the extreme vectors and points that help build the hyperplane. The foundation of the SVM approach is the support vectors, which are utilized to represent these extreme situations. Take a look at the image below, where a decision boundary or hyperplane is used to classify two separate categories

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

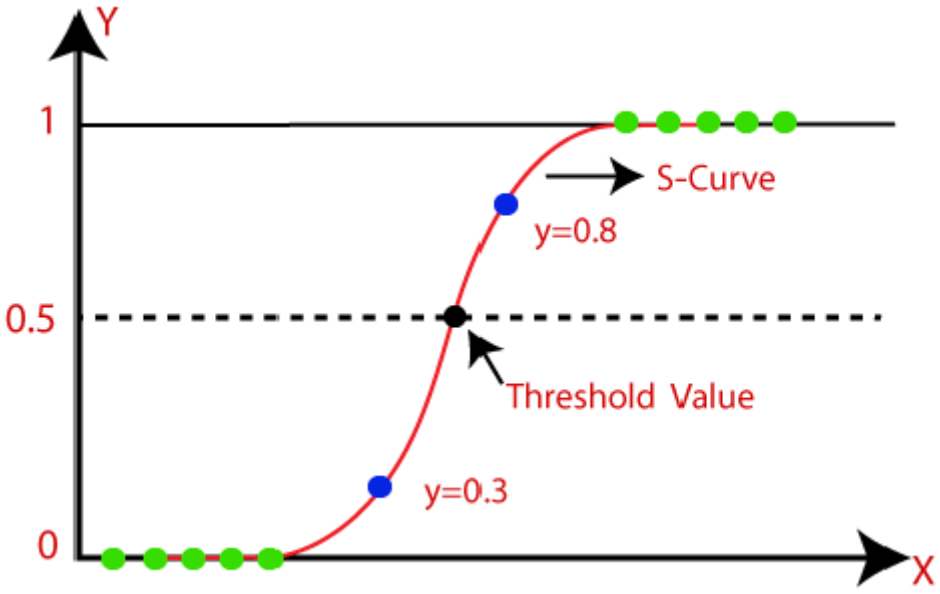
| DOI: 10.15680/IJIRCCE.2024.1205200 |



Logistic regression

▪ In binary classification issues, where the goal is to predict one of two outcomes, logistic regression is a frequently used approach. Through the use of a sigmoid function, it converts the output of the linear regression into a probability value between 0 and 1, which can then be used to decide whether to classify data by applying a threshold.

▪ With applications in many areas, including credit scoring, spam filtering, and medical diagnosis, this simple yet reliable algorithm may be taught well on big datasets. However, because it depends on certain presumptions, such as the linearity and independence of the characteristics, it could not work well with highly coupled or nonlinear data.



Decesion Tree Classification

▪ For both binary and multi-class classification tasks, decision tree classification is a popular machine learning approach. The input data are recursively divided into subgroups depending on the most instructive characteristic.

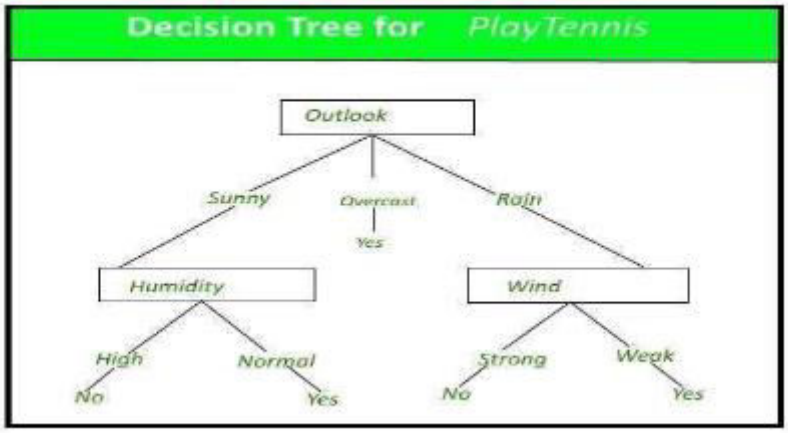
International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205200 |

▪ Decision trees can handle category and numerical data and are simple to understand and use. Additionally, they are resistant to noise and missing data and are capable of capturing intricate non-linear correlations between features



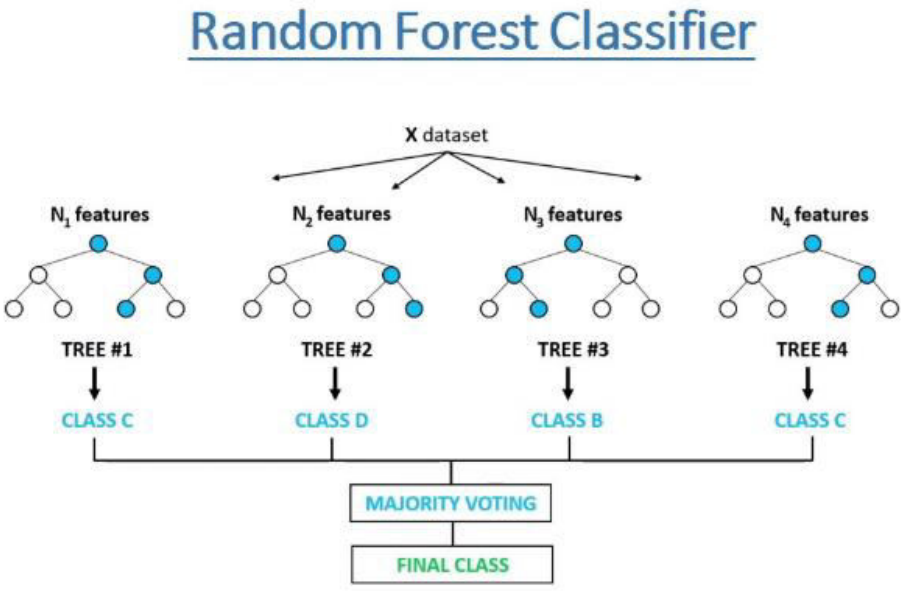
Gradient Boosting Classifier

▪ Gradient Boosting Classifier is a powerful algorithm for both classification and regression problems. It works by combining multiple weak models, such as decision trees, to create a strong ensemble model.

▪ One of the advantages of Gradient Boosting Classifier is that it can handle complex non-linear relationships between features and the target variable. Additionally, it has a built-in mechanism for handling missing data and can automatically select important features for better accuracy. However, it can be computationally expensive and prone to overfitting if not tuned properly

Random Forest Classifier

▪ As the name implies, a Random Forest consists of numerous independent decision trees that work together as an ensemble. Each tree in the Random Forest spits out a class prediction, and the classification that recieves the most votes becomes the prediction of our mode.

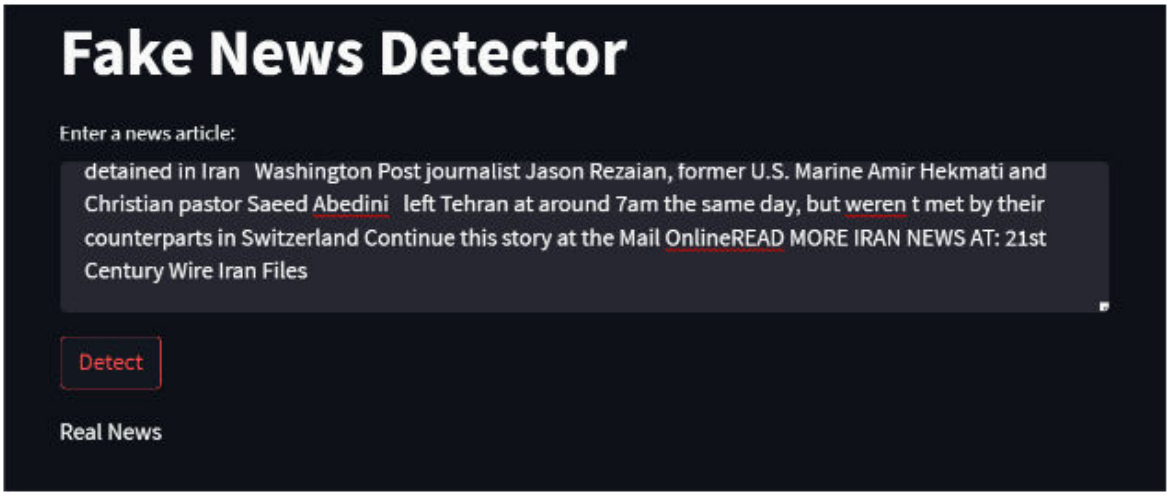


| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205200 |

VI. RESULTS



VII. CONCLUSION

Considering the accuracy scores, we were able to establish for the various models, it appears that all of the models are doing a good job of identifying false news items. The SVM, Decision Tree, and Gradient Boosting classifiers notably achieved a very high accuracy of 99.5%, although the Random Forest Classifier performed just slightly lower, at 98.71%. All things considered, these results suggest that a range of classifiers may be used with equal success rates and that machine learning techniques may be extremely successful in spotting bogus news. It's important to keep in mind that accuracy is only one measure and that the models should be evaluated using multiple metrics including precision, recall, and F1-score in addition to factors like interpretability, scalability, and processing requirements. Investigating different feature extraction and selection methods, classifier types, and ensemble approaches may also be useful to see whether even better results may be produced.

VIII. FUTURE SCOPE

Future research and advancement in the field of false news detection are abundantly possible. Future efforts to identify bogus news may go in the following directions:
Including more varied and subtle aspects: For the most part, current methods for detecting false news rely on simple text-based traits like TF-IDF vectors or bag-of- words. Research in the future could concentrate on more complex and diverseaspects, such sentiment analysis, network analysis, or multimedia analysis (for instance, identifying false images or videos). Creating more interpretable models: Existing methods for spotting fake news sometimes rely on complex machine learning algorithms that might be difficult to comprehend. In the future, it would be beneficial to develop more intelligible models that might provide more information on how people make decisions. Combining information from other sources: In addition to social media, news articles, and videos, fake news is regularly spread through other media channels and platforms. The development of methods that can incorporate data from several sources may be crucial in the future to improve false news identification. Adapting to shifting strategies: It will be crucial for fake news detection technologies to develop alongside the tactics used by those who create and spread it. For this, the detection methods might need to be regularly reviewed and improved.

REFERENCES

[1] A. S. A. Ahmed, A. Abidin, M. A. Maarof, and R. A. Rashid, "Fake news detection: A survey," IEEE Access, vol. 9, pp. 113051-113071, 2021. doi: 10.1109/ACCESS.2021.3104178
[2] S. Asghar, S. Mahmood, and H. Kamran, "Fake news detection using machine learning: A survey," IEEE Access, vol. 9, pp. 57613-57639, 2021. doi: 10.1109/ACCESS.2021.3075392
[3] J. H. Kim, S. H. Lee, and H. J. Kim, "Fake news detection using ensemble learning with context and attention mechanism," IEEE Access, vol. 9, pp. 27569- 27579, 2021. doi: 10.1109/ACCESS.2021.3057736

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205200 |

[4] M. F. Hossain, M. M. Islam, M. A. H. Khan, and J. J. Jung, "Fake news detection using hybrid machine learning algorithms," IEEE Access, vol. 8, pp. 233350- 233364, 2020. doi: 10.1109/ACCESS.2020.3041149

[5] S. S. Ghosh, A. Mukherjee, and N. Ganguly, "A multi-perspective approach to fake news detection," IEEE Intelligent Systems, vol. 35, no. 5, pp. 31-39, 2020. doi: 10.1109/MIS.2020.3012915

[6] Lowlesh Yadav and Asha Ambhaikar, "IOHT based Tele-Healthcare Support System for Feasibility and performance analysis," Journal of Electrical Systems, vol. 20, no. 3s, pp. 844–850, Apr. 2024, doi: 10.52783/jes.1382.

[7] L. Yadav and A. Ambhaikar, "Feasibility and Deployment Challenges of Data Analysis in Tele-Healthcare System," 2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI), Raipur, India, 2023, pp. 1-5, doi: 10.1109/ICAIIHI57871.2023.10489389.

[8] L. Yadav and A. Ambhaikar, "Approach Towards Development of Portable Multi-Model Tele-Healthcare System," 2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI), Raipur, India, 2023, pp. 1-6, doi: 10.1109/ICAIIHI57871.2023.10489468.

[9] Lowlesh Yadav and Asha Ambhaikar, Exploring Portable Multi-Modal Telehealth Solutions: A Development Approach. International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC), vol. 11, no. 10, pp. 873–879, Mar. 2024.11(10), 873–879, DOI: 10.13140/RG.2.2.15400.99846.

[10] Lowlesh Yadav, Predictive Acknowledgement using TRE System to reduce cost and Bandwidth, March 2019. International Journal of Research in Electronics and Computer Engineering (IJRECE), VOL. 7 ISSUE 1 (JANUARY-MARCH 2019) ISSN: 2393-9028 (PRINT) | ISSN: 2348-2281 (ONLINE).

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details