FAKE NEWS DETECTION

T.G.Ramnadh Babu[1], Guttikonda Ramya, Anna J.V.Asritha[3], Kummaragunta Mounika [4]
[1] Professor, [2, 3 & 4] Student
[1]baburamnadh@gmail.com [2] ramya709376@gmail.com [3] ajvasritha17@gmail.com, [4] kummaraguntamounika@gmail.com
Department of Computer Science and Engineering,
Narasaraopeta Engineering College, Narasaraopet, Andhra Pradesh, India

ABSTRACT- The dissemination of misinformation presents a significant threat to societal well-being and stability. The rapid expansion of social media platforms has intensified the spread of fake news, exacerbated by the absence of effective countermeasures. This study aims to tackle this issue by investigating a range of machine learning techniques customized for analyzing and detecting fake news. Leveraging the WELFake dataset, which encompasses diverse textual data sources, we explore various methodologies, including Logistic Regression. Our research is solely focused on textual data, with the primary objective of constructing robust models capable of accurately identifying the propagation of fake news across social media platforms.

KEYWORDS: Fake News, Social Media Platforms, Machine Learning, Logistic Regression, Textual Data

I INTRODUCTION

In today's digital landscape, social media has become the primary channel for sharing news, both true and false. The ease of sharing personal opinions has led to the rapid spread of misinformation, posing risks to individuals and even national security. To combat this, automated systems using machine learning algorithms[1] like Logistic Regression and Random Forest are being developed. These systems aim to detect fake news and protect users from false content. Additionally, user interfaces are being designed using frameworks like Flask to make these systems accessible and effective in real-world scenarios.

II LITERATURE SURVEY

There have been several studies conducted on fake news detection in various statistical and machine learning techniques. Here are some of the notable literature survey for False news identification:

Farzana Islam etal.[2] study on Navie Bayeis classifier technique to find fake stories. They used count vectorization and TF-IDF vectorization approaches to extract features. This approach was used to examine two publicly available false news datasets from various sources.

The experiments on a false news dataset obtained from Kaggle.com.by Zongru etal.[3]The study used RNN, RF, and Naïve Bayes classifiers. They investigated several feature extraction techniques, such as Bigram and Trigram.

III PROPOSED SYSTEM

The below flow chart Fig.1 defines our methodology. It represent in five steps as shown in Fig.

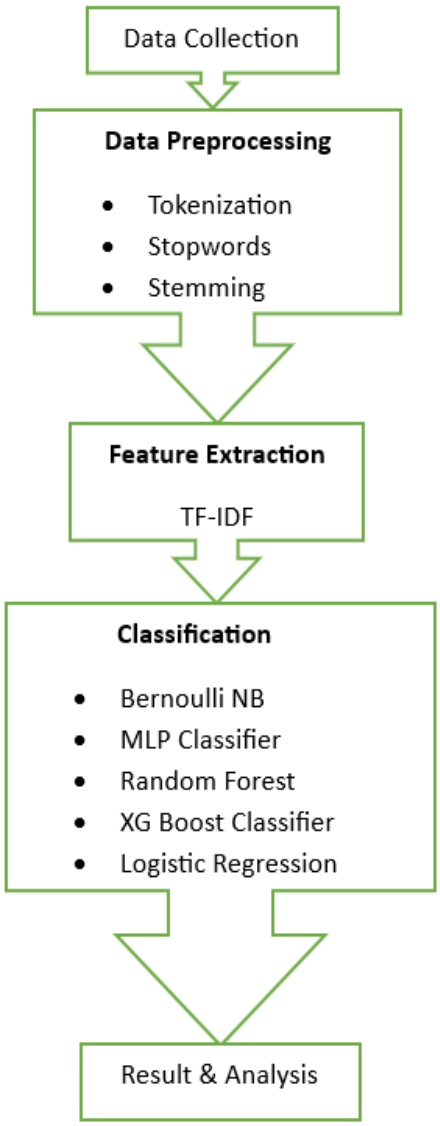

Fig.1. General Approach for our methodology

Our Model is Proposed based on certain criteria as follows:

A.Dataset Collection

B.Preprocessing Techniques

C.Feature Extraction

D.Model Creation and Evaluation

E.Result & Analysis

A.Data Collection:

consider the WELFake dataset[4], consists of 72,134 records with 35,028 labeled as true , 37,106l labeled as false.True data indicate with'0' and false data indicate with '1' Fig. 2 represent the count of each label data in dataset.

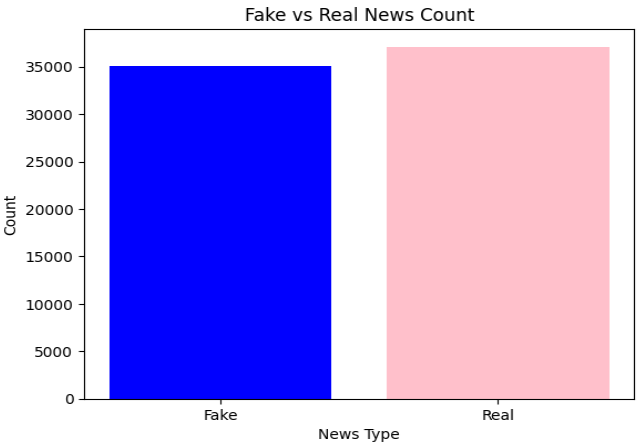

Fig2. Representation of Dataset

B.Preprocessing Techniques:

In our study, we focused on preprocessing unstructured text data to prepare it for analysis. We used various techniques to address text mining challenges:

Remove special charecters, punctuation,and digits using Regular Expression.

Stopword removal[5]: We removed stopwords such as "the" and "with" to improve data categorization by filtering out insignificant words.

Stemming: We applied stemming to simplify words to their root form. For example, "calling" would be stemmed to "callin".

Tokenization: This process involved dividing the text into separate tokens to ensure uniformity in data representation.

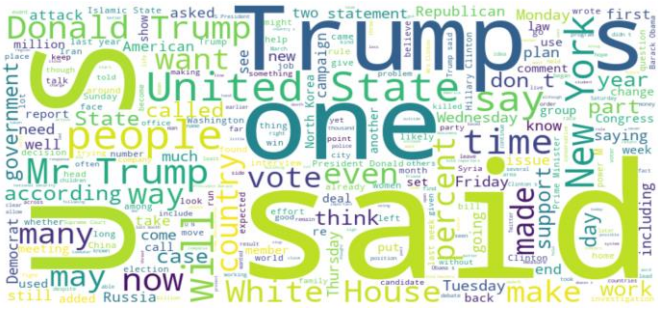

Fig. 3. Word Cloud of Fake Label Data From Dataset

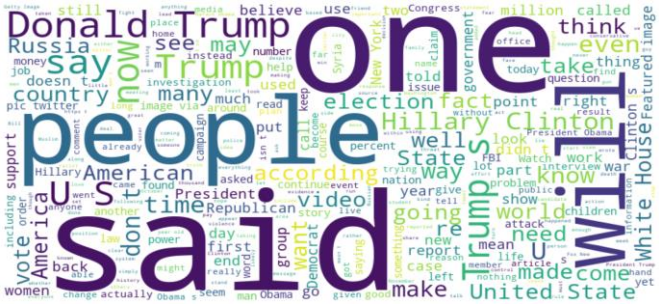

Fig. 4. Word Cloud of Real Label Data From Dataset

Fig 3& 4 represents the word cloud of real and fake labelled data in dataset.

C.Feature Extraction:

In text processing, converting textual data into binary vectors of 0s and 1s is a common technique for training classification algorithms. Each data in the sample text file generates a new vector. One method to create these vectors is using the TF-IDF[4] technique. The TF-IDF vectorizer[6] The TF-IDF vectorizer is a popular approach for feature extraction. It has two stages: Term Frequency: This stage determines the phrase frequency.

D.Model Creation And Evaluation:

- *BernouliNB:*

  Bernoulli NB[7] enumerate the probability depending on features provided for each class.
  Fig. 5 depicts the model's CM define true positives, false positives, true negatives, and false negatives. It helps to understand the model's performance by demonstrating how well it predicts each class.

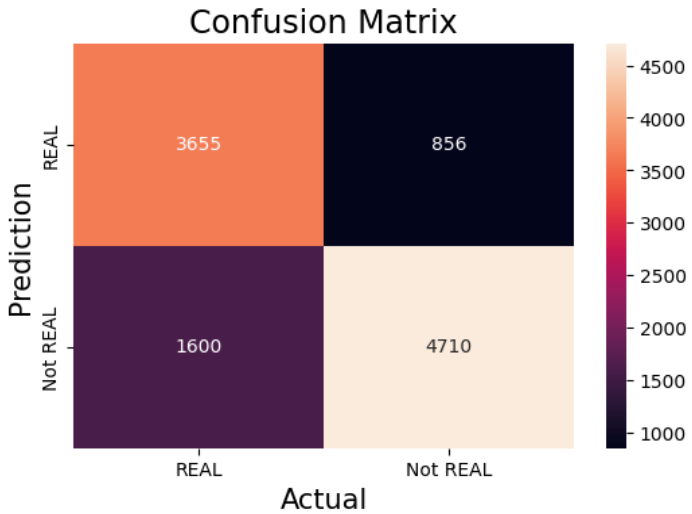

Fig 5. Confusion Matrix of BernouliNB

- *Multilayer Perceptron:*

  The MLP Classifier [8] is a form of Artifical Neural Network is commonly applied in supervised Learning.It is made up of numerous interconnected layer of the nodes,with Each node calculates the weighted sum of its inputs and applies activation function to produced an output.

Fig. 6 represent the confusion matrix of Multi Layer Perceptron classifier.It helps to identify true positives false,positives,false postives,false negatives.

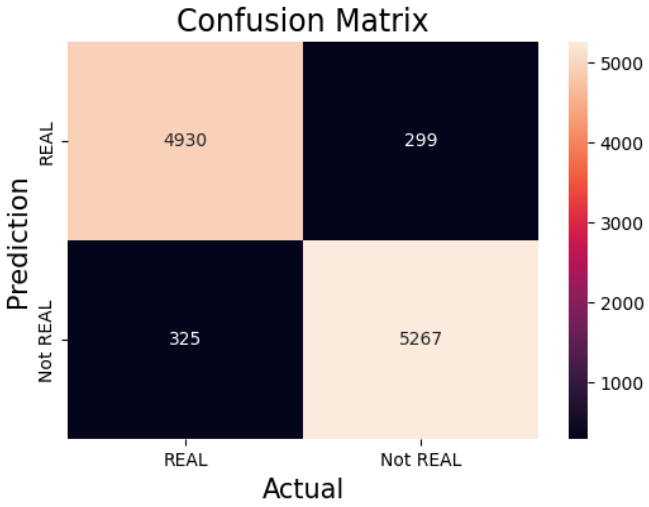

Fig 6. Confusion Matrix of MLP

- *Random Forest:*

The Random Forest [9] is an ensemble machine Learning approach that uses numerous decision trees to increase prediction accuracy and consistency.

Fig.7 represent the confusion matrix of RF classifier.It helps to identify the True Positives, False Positives, True Negatives, False Negatives.

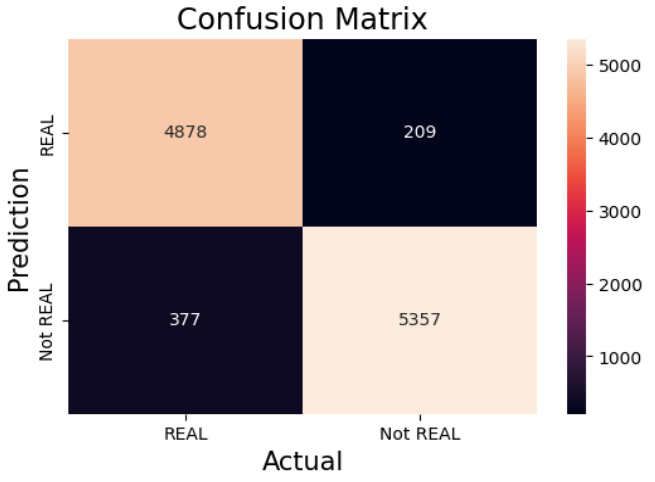

Fig 7. Confusion Matrix of Random Forest

- *XGBoost:*

The Extreme Gradient Boost [10] is a powerful Machine Learning algorithm widely employed in supervised learning scenarios such as classification regression.

Fig 8, Represents the confusion matrix of XG Boost Classifier.It helps to identify the True Positives ,False Positives, True Negatives,False Negatives of the model XG Boost Classifier.

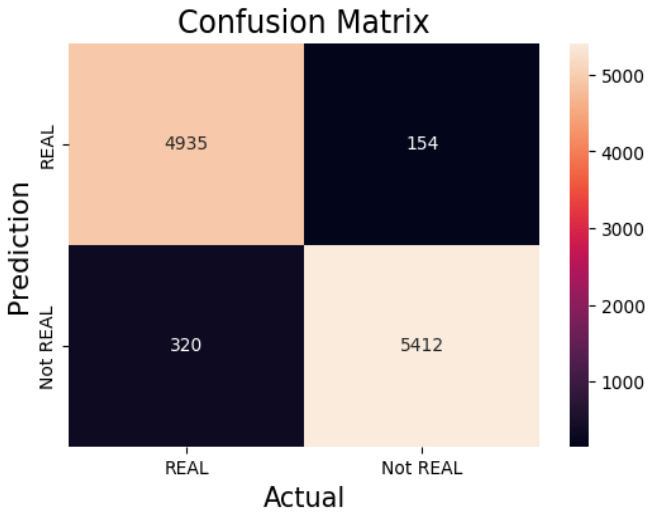

Fig 8 Confusion matrix of XG Boost

- *Logistic Regression:*

Logistic Regression[11] is a type of the regression Analysis that predicts the outcome of the dependent variable using one or more predictors. It is typically applied to binary classification issues in which the dependent variable has two alternative outcomes.

Fig. 9.Represents the confusion matrix of Logistic Regression.

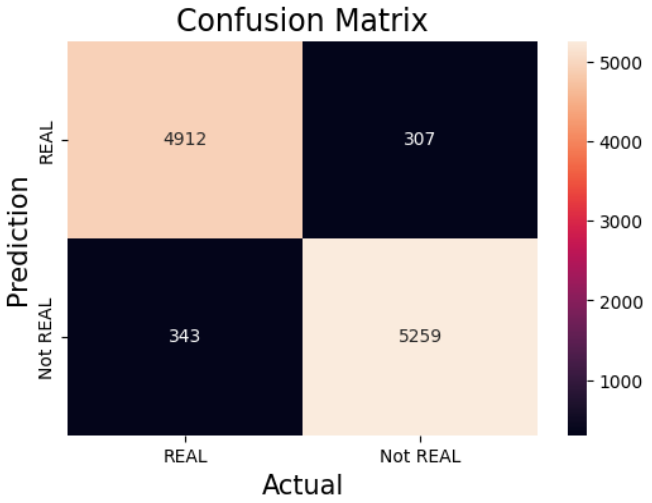

Fig 9. Confusion Matrix of Logistic Regression

E. Result & Analysis:

Among all specified machine learning alogorithms logistic Regression,Random Forest,Bernouli NB,MLP Classifier,XG Boost Classifier.And we get highest accuracy of 96% to Logistic Regression So, we prefereed this as best model.

AUC ROC Curve:

The AUC-ROC[12] curve, is a pictorial representation mostly used for binary tasks to caluclate the performence of a model The Fig. 10 represents the Roc Cureve of Logistic Regression of our model.It gives the True postives rate of Logistic Regression.

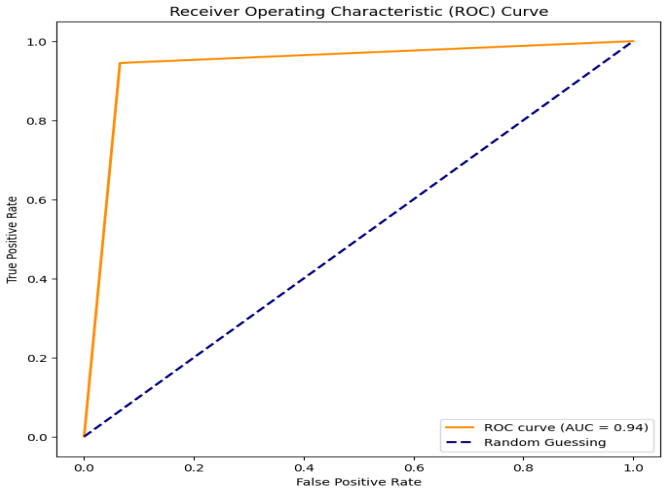

Fig 10. ROC Curve of Logistic Regression

Fig.8 represent the accuracy of all evaluated models.

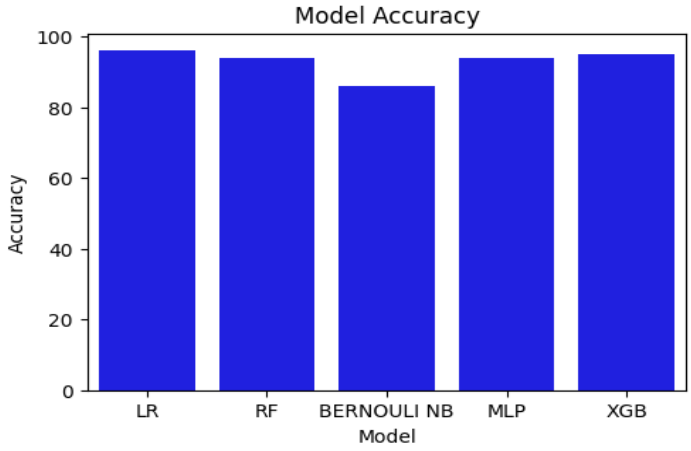

Fig 8. Repersentation of All evaluated models

The below table 1. Represent the comparision[13] of different models like Bernouli NB, MLP Classifier, Random Forest, Logistic Regression

Model	Accuracy	Precision	F1 score	Recall
Bernouli NB	**86%**	**84%**	**86%**	**89%**
MLP Classifier	**94%**	**94%**	**94%**	**94%**
XGB Classifier	**95%**	**94%**	**95%**	**97%**
Random Forest	**94%**	**93%**	**94%**	**96%**
Logistic Regression	**96%**	**95%**	**96%**	**97%**

Table1: Comparision of all models

Fig 9. Represents the comparison all evaluated model of existing system and proposed system.

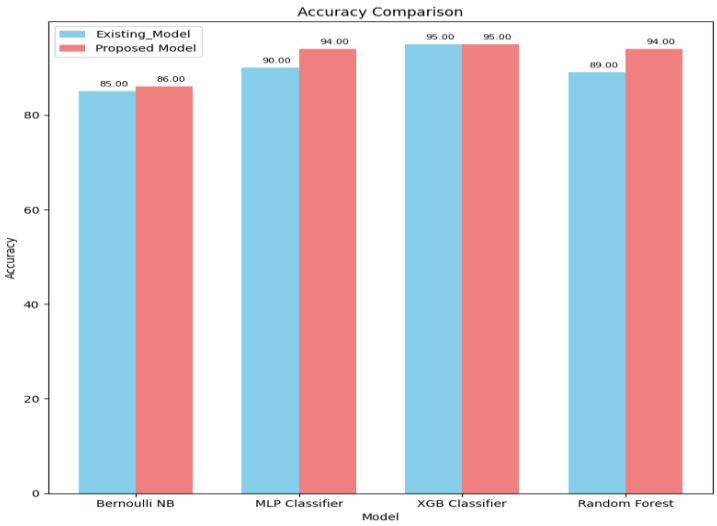

Fig. 9. Comparision Of all models with existing system

IV. CONCLUSION AND FUTURE SCOPE

In conclusion, this study explored the application of Five popular machine learning algorithms,[14].The results obtaine indicate that get highest acuuracy for Logistic Regression so consider as a best model.Future research can expand[15] on this can expand on this work by using more complex features, exploring alternative algorithms, and analyzing the model's performance on a larger dataset of different categories.

V REFERENCES

[1] Archit Guptha,Arnav Batla,etc research on "Comparative Analysis Of Machine Learning Models For Fake News Classification" 2023 3rd International Conference on Intelligent Technologies (CONIT) Karnataka, India. June 23-25, 2023.

[2] A.Farzana Islam, and her team research on "Effect of Corpora on Classification of Fake News using Naive Bayes Classifier." International Journal of Automation, Artificial Intelligence and Machine Learning 1.1.2020

[3] P. Bharadwaj, and Zongru Shao. "Fake news detection with semantic features and text mining." International Journal on Natural Language Computing (IJNLC) Vol 8 -2019.

[4] https://www.kaggle.com/datasets/saurabhshahane/fake-news-classification.

[5] Fahim Belal Mahmud and his team research on "A comparative analysis of Graph Neural Networks and commonly used machine learning algorithms on fake news detection." 2022 7th International Conference on Data Science and Machine Learning Applications (CDMA). IEEE, 2022.

[6] Smitha.N, Bharath.R, "Performance Comparison of Machine Learning Classifiers for Fake News Detection" Proceedings of the Second International Conference on Inventive Research in Computing Applications (ICIRCA-2020).

[7] Baarir and Djeffal's research on "Fake News Detection Using Machine Learning" 2020: Second International Workshop on Human-Centric Smart Environments for Health and Well-Being (IHSH).2020

[8] Jasmine Shaikh and Rupali Patil conducted study on "Fake News Detection Using Machine Learning". 2020 IEEE International Symposium on Sustainable Energy, Signal Processing, and Cybersecurity (iSSSC).2020

[9] T. H. Borkar, T. Ahuja, "Comparative Study of Supervised Learning Algorithms for Fake News Classification," Tirunelveli will host the sixth International Conference on Trends in Electronics and Informatics (ICOEI) in 2022.2022

[10] Rubin, Victoria, L., and their team investigate if fake news is true or not. Using sarcastic clues to identify possibly deceptive news. In: Proceedings of NAACLHLT 2016.

[11] Gahirwal Manisha et. al; International Journal of Advance Research, Ideas and Innovations in Technology ISSN: 2454-132X Impact factor: 4.295

[12] Perez-Rosas, Verónica, Kleinberg, Bennett, Lefevre, Alexandra, and Mihalcea Rada conducted research in 2017 on "Automatic Detection of Fake News".2017

[13] Chaowei Zhang, Ashish Gupta, Christian Kauten, Amit V. Deokar, and Xiao Qin, "Detecting fake news for reducing misinformation risks using analytics approaches," in the proceedings of ELSEVIER, European Journal of Operational Research, 2019.

[14] Z Khanam and his team research on “False News Detection Using Machine Learning Approaches” 2021 IOP Conf. Ser. Mater. Sci. Eng. 1099 012040.

[15] Cristina M Pulido, Laura Ruiz-Eugenio, Gisela Redondo-Sama, and Beatriz Villarejo-Carballido. A new application of social impact in social media for overcoming fake news in health. International journal of environmental research and public health, 17(7):2430,2