Development of a Web Application for Fake News Classification using Machine learning

Ankit Gupta[1], Muskan Pethiya[2], Diksha Wankhede[3], Harshal Gajbhiye[4], Nilima Ghuguskar[5], Mrudula Nimarte[6], Hrishikesh Panchabudhe[7]

[1,2,3,4,5] Student, [6,7] Professor, Department of Computer Science and Engineering, S.B. Jain Institute of Technology, Management and Research, Nagpur

---***---

Abstract - *The spread of fake news on social media and other platforms is a serious worry because it has the potential to have a negative influence on society and the country. On finding it, there has already been a lot of research. The automatic detection of false content in news stories is the main topic of this research. Westar by introducing a dataset for the false news detection job. We provide a thorough explanation of the pre-processing feature extraction, classification, and prediction procedures. To categories bogus news, we applied language processing methods based on logistic regression. Tokenizing, stemming, and exploratory data analysis, including response variable distribution and data quality checks (i.e., null or missing values), are some of the tasks carried out by the pre-processing algorithms. Simple feature extraction methods include the usage of n-grams, bag-of-words, and TF-IDF. As a classifier for fake news identification with probability of truth, the logistic regression model is used.*

Keywords: **Fake news detection, Logistic regression, TF-IDF vectorization.**

I. INTRODUCTION

As the spread of false information online accelerates, particularly in media channels like social media feeds, news blogs, and online newspapers, fake news detection has recently drawn increasing interest from the general public and researchers. Fake accounts, posts, and news are a problem on social media and the internet. The goal is frequently to deceive readers and/or persuade them into making erroneous purchases or beliefs. Therefore, a system like this would help in some small way to solve a problem.

When reading a sentence or a paragraph, a person can comprehend the words in the context of the entire document. In this project, we use machine learning and prediction classifiers like the logistic regression to train a system how to read and grasp the differences between real news and fake news. These techniques will predict if an item is true or false.

A. Goals or Objectives:

- To classify whether the news is fake or real.
- To build model interface credibility by labelling of new articles whether it is real or fake.
- To aware people about fake news.

II. LITERATURE SURVEY

Fake news can generally be divided into three categories. Fake news, or news that is wholly made up by the authors of the pieces, is the first category. The second category is phony satire news, which is made primarily with the intention of making readers laugh. The third category consists of badly written news items that contain some genuine news but are not totally accurate. In essence, it refers to news that fabricates entire stories while quoting political people, for instance. This form of news is typically intended to advance a particular goal or a prejudiced opinion.

Zhou, X., Zafarani, R.: A survey of fake news:fundamental theories, detection methods, and opportunities' Comput. Surv. (2020). They studied analyses and assesses strategies for identifying fake news from four angles: the inaccurate information it contains, the writing style, the ways it spreads, and the reliability of the source. Based on the review, the poll also suggests a few interesting study projects. In order to promote interdisciplinary study on false news, we specifically identify and describe similar core theories across multiple fields [1].

Agarwal, Aarush, and Akhil Dixit. "Fake News Detection: An Ensemble Learning Approach." 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE, 2020. In this study, they present a model for identifying false news by assessing a report's correctness and determining its veracity. This model creates an ensemble network to learn the depictions of news items, authors, and titles simultaneously through feature extraction and believability score construction from the textual input. Precision, recall, and F1-Score were used to assess the efficiency and effectiveness of classifiers. The application of several algorithms

demonstrates the dataset's effectiveness [2].

Tanveer Khan, Adnan Akhundzada: "Fake news outbreak 2021: Can we stop the viral spread?" Accepted 19 May 2021 by Tampere University, Finland. They thoroughly examine numerous approaches for the early identification of fake news in the literature and present their findings in this overview study. They specifically look at Machine Learning (ML) models for detecting and categorizing false news, online fake news detection contests, statistical results, and the benefits and drawbacks of some of the accessible data sets. They next assess the current web browsing tools available for spotting and preventing fake news, and they outline some unresolved research problems [3].

Hadeer Ahmed, Issa Traore, and Sheriff Saad in their paper they proposed a false news detection programmed that makes use of machine learning and n-gram analysis. Six distinct machine classification algorithms were examined and compared, as well as two different features extraction strategies. With an accuracy of 92%, the experimental evaluation's best results are obtained utilizing the feature extraction method Term Frequency-Inverted Document Frequency (TF-IDF) and the classifier Linear Support Vector Machine (LSVM) [4].

In the article published by Kai Shu, Amy Sliva Suhang Wang, Jiliang Tang, and Huan Liu, by analyzing the current literature in two stages characterization and detection—they investigated the issue of false news. They introduced the fundamental ideas and tenets of fake news in both traditional and social media during the characterization phase. They analyzed existing fake news detection techniques from a data mining standpoint during the detection phase, including feature extraction and model building [5].

Muhammad Ovais Ahmad ,[2020] in this paper, they advocate the automated classification of news articles using a machine learning ensemble technique. They investigate many linguistic characteristics that can be utilized to tell bogus information from authentic. These attributes are used to train a variety of machine learning algorithms using various ensemble approaches, and their performance is assessed using four real- datasets. Experimental analysis demonstrates that they suggested ensemble learner technique outperforms individual learners [6].

Kumar; Kumar; Yadav; Meghna Bagri(IEEE)[2021], this paper examines various methods for spotting fake news and a study of papers from 2017 to 2021. This study provides a comprehensive overview of historical and present studies on the identification of false news using various ML algorithms [7].

Z Khanam et al 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1099 012040 , In order to develop a model of a product with supervised machine learning algorithm, which can classify fake news as true or false by using tools like Python Scikit-Learn, NLP for textual analysis, this paper analyses the research on fake news detection and explores the best traditional machine learning models. They suggest utilising the Python scikit-learn module to perform tokenization and feature extraction of text data because it has helpful tools like Count Vectorizer and Tiff Vectorizer. This method will result in feature extraction and vectorization. Then, using the results of the confusion matrix, they applied feature selection techniques to explore and pick the best-fit features to get the highest degree of precision [8].

2019 IEEE International Conference on Industrial Internet (ICII)In order to develop a model of a product with supervised machine learning algorithm, which can classify fake news as true or false by using tools like Python Scikit-Learn, NLP for textual analysis, this paper analyses the research on fake news detection and explores the best traditional machine learning models. They suggest utilizing the Python scikit-learn module to perform tokenization and feature extraction of text data because it has helpful tools like Count Vectorizer and Tiff Vectorizer. This method will result in feature extraction and vectorization. Then, using the results of the confusion matrix, they applied features election techniques to explore and pick the best-fit features to get the highest degree of precision [9].

III. PROPOSED WORK

A. Flow of the System:

After being accessed by the user, the website requests input from them in the form of text, sentences, etc. With our dataset, this arbitrary data is preprocessed. After training the classifier with inputs like text, sentences, etc. using the technique

Logistic Regression, we next perform procedures like feature extraction, etc. After the aforementioned procedures, we examine database results for our most current output. If this output is discovered or matches our dataset, the accuracy of these text and sentence matches in the dataset is checked. Our website will publish the output as real/true news if this accuracy percentage is deemed to be above 90%; otherwise, it will display as false news. If current output cannot be located or does not match our dataset, our website will display it as fake news.

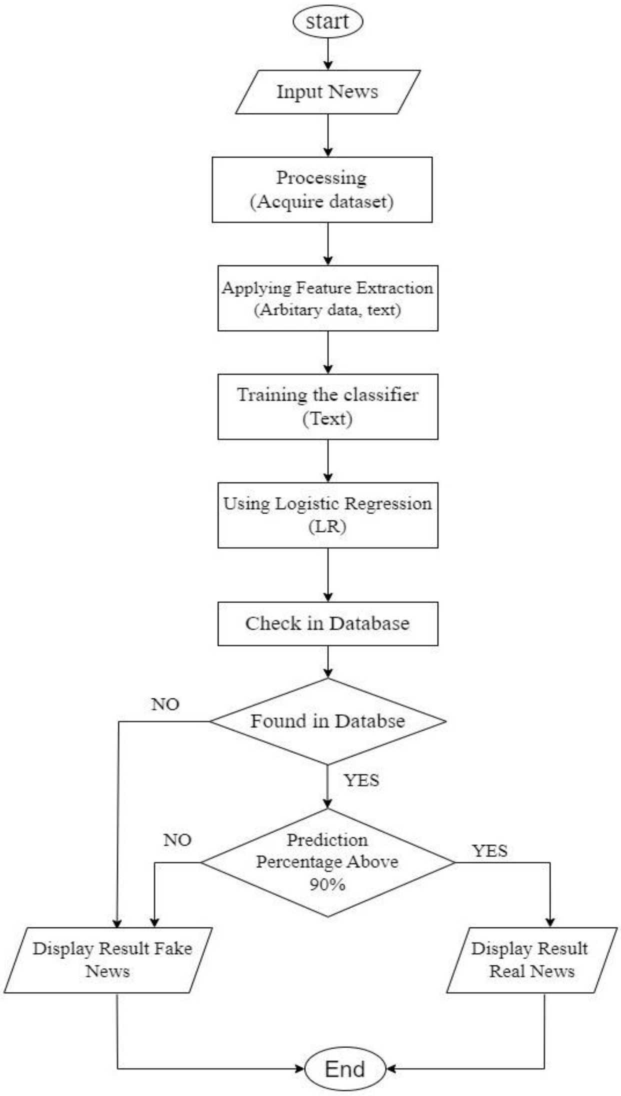

start
Input News
Processing (Acquire dataset)
Applying Feature Extraction (Arbitary data, text)
Training the classifier (Text)
Using Logistic Regression (LR)
Check in Database
Found in Databse
NO
YES
Prediction Percentage Above 90%
NO
YES
Display Result Fake News
Display Result Real News
End

Fig.3.1 Proposed flow

B. Functional Modules:

The whole system is divided into three modules. They are Dataset Collection and Pre-processing, Trained the Dataset, Classification and Prediction.

1) *Dataset Collection and Pre-processing*:

We gather data for this module from the Kaggle website and news websites. On a raw dataset of text and random data, we conduct feature extraction. By using text, articles, and news headlines, we train our classifier.

2)*Trained the Dataset:*

We trained the dataset using the Logistic Regression (LR) algorithm. News API was applied.

3)*Classification and Prediction:*

After classifying the dataset, we predict the end result, or prediction of user input, by comparing the amount of text supplied by users to the news' original text.

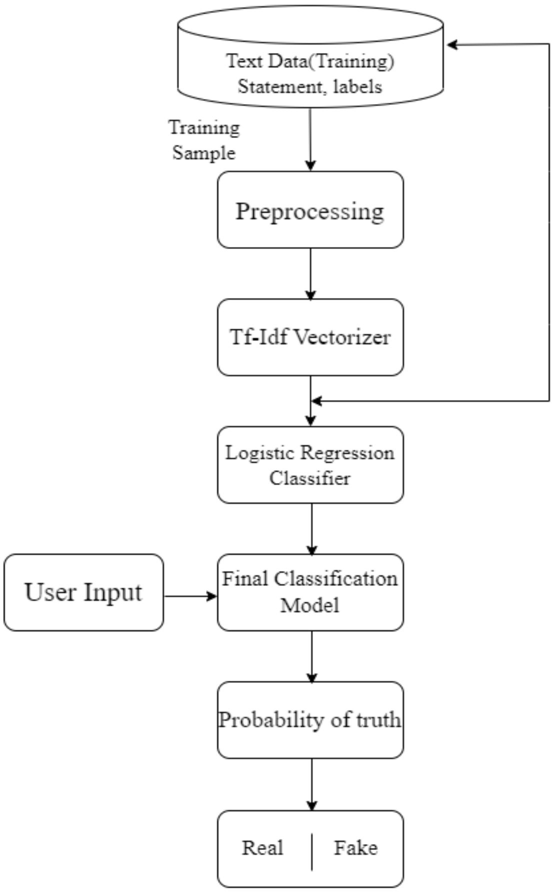

Text Data(Training) Statement, labels
Training Sample
Preprocessing
Tf-Idf Vectorizer
Logistic Regression Classifier
User Input
Final Classification Model
Probability of truth
Real
Fake

Fig.3.2 Preprocessing Flow

C. Classifier:

- TF-IDF Vectorizer:

The term frequency of a word in a document is determined by how frequently it appears in that document. When a term is part of the search terms, a greater number indicates that it occurs more frequently than others, indicating that the document is a good match. IDF (Inverse Document Frequency): Terms that frequently appear in one document but not in many others may not be meaningful. IDF is a metric for gauging a term's importance across the board. The TfidfVectorizer creates a matrix of TF-IDF features from a set of source documents.

- Logistic Regression:

To estimate the likelihood of a categorical dependent variable, a machine learning classification approach is used. The dependent variable in logistic regression is a binary variable with data coded as 1 (yes, success, etc.) or 0 (no) (no, failure, etc.). In other words, P(Y=1) is predicted by the logistic regression model.

D. System Architecture

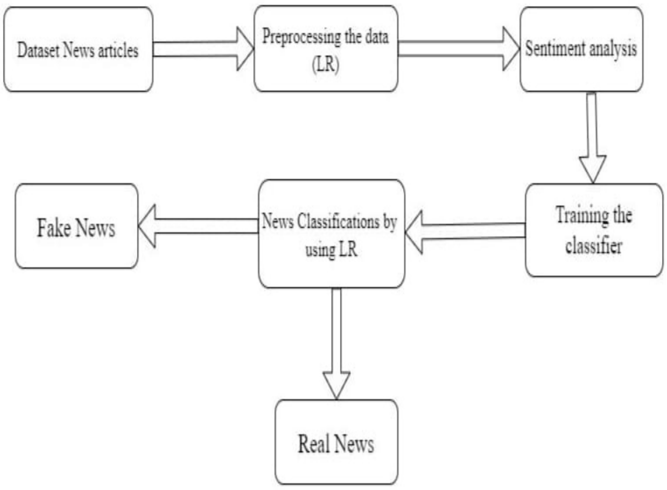


Fig.3.3 System Architecture

The static portion of the fake news detection system's architecture is rather straightforward and was designed with the basic machine learning process flow in mind. The system design is self-explanatory and is depicted above. The second search field on the website asks for keywords to be looked up online, and it then returns an appropriate result with the likelihood of that term really being in an article or an article with similar content that makes use of those keywords. It will then categorize whether the news is fake or not.

IV. APPLICATION

1. Stock market / Stock exchange - False news detection techniques are used in this industry, for example, if someone spreads false information about shares or the stock market, it may result in financial loss for them.

2. Business Trade - When someone buys a stake in a business, it can occasionally cause a big problem for the company since they will either make a loss or a profit.

3. Political Voting - In the political process, news is crucial. If the news is false, it will have an impact on the local population, which will be unaware of who will be picked as their leader.

V. RESULT

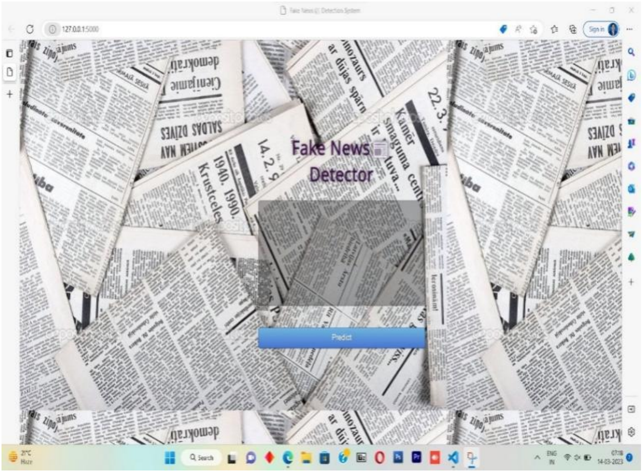
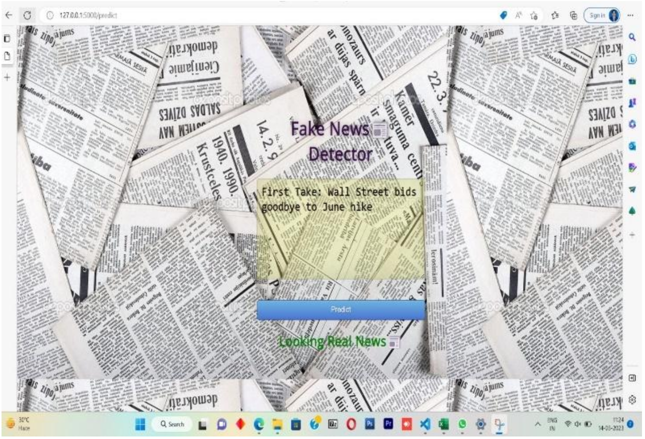


Fig.5.1 Home Page

This is the home page of our website. Above fig shows the input screen to the user where user will give the input such as news articles and some related headlines. After that user will click on the predict button then he/she will get output as whether the news is fake or spam.



First Take: Wall Street bids goodbye to June hike

Looking Real News

Fig.5.2 Fake News Detection Page

Fig.5.2 shows the result of the given input to the user , it shows the news as real to the user.

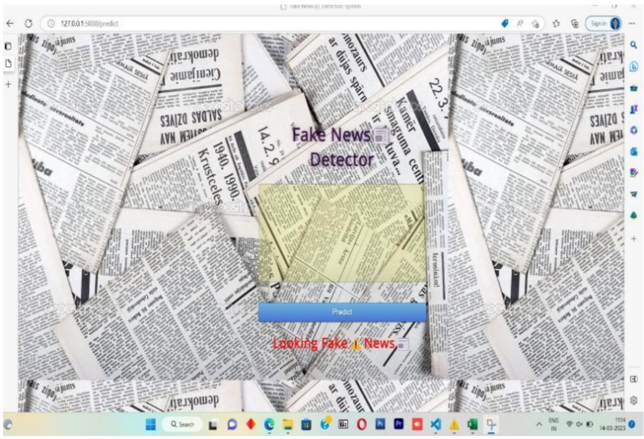


Fig.5.3 Fake News Detection Page

Above fig shows the result of the given input to the user , above fig shows the news as spam/fake to the user.

IV. CONCLUSION

This project helps in developing an approach for classification of Fake News. It has helped to explore the various approach that has been previously developed. With this survey and study, we have proposed an efficient approach for detection of news whether it is real or fake using Logistic Regression Algorithm while implementing this, it gives the accuracy of news upto 95%. Our approach also has a feature extraction of matching the text, sentences with our dataset and predict the result and that will be our final output. This system in future will help a person to find out news whether its real or fake.

REFERENCES

[1] Zhou, X., Zafar ani, R.: A survey of fake news: fundamental theories, detection methods, and opportunities. ACM Compute Survey [2020].

[2] Agarwal, Aarush, and Akhil Dixit. "Fake News Detection: An Ensemble Learning Approach."2020 4th International the Conference on Intelligent Computing and Control Systems (ICICCS). IEEE, [2020].

[3] Tanveer Khan, Adnan Akhundzada: “Fake news outbreak 2021: Can we stop the viral spread?” Accepted 19 May 2021 by Tampere University, Finland.

[4] Hader Ahmed, Issa Traore, and Sheri Saad, International Conference on Intelligent, Secure, and Dependable Systems in Distributed and cloud Environments, October[2017].

[5] Kai Shu, Amy Silva Suhang Wang, Jiliang Tang, and Huan Liu. Fake News Detection on social media: A Data Mining Perspective [Submitted on 7 Aug 2017 (vl), Last revised 3 Sep 2017 (this version,v3)].

[6] Muhammad Ovais Ahmad [2020].WILEY-HINDAWI,2020. Vol. 2020, article id 8885861.

[7] Shailendra; Subodh Kumar; Pooja Yadav; Meghna Bagri (IEEE) A Survey on Analysis of Fake News Detection Techniques [2021].

[8] Z Khanam et al 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1099 012040. IOP Conference Series: Materials Science and Engineering, the Volume 1099,International Conference on Applied Scientific Computational Intelligence using Data Science (ASCI 2020) 22nd-23rd December 2020, Jaipur, India.

[9] IEEE International Conference on Industrial Internet (ICII). Fake News Detection in Social Networks Using Machine Learning and Deep Learning: Performance Evaluation [2019].