www.kiet.edu
Delhi-NCR, Ghaziabad

KIET
GROUP OF INSTITUTIONS
Connecting Life with Learning

A

Project Report

on

Fake News Detection Using EnsembleLearning

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY DEGREE

SESSION 2022-23

in

Department of Computer Science and Engineering

By

Mohd Naseem (1900290100085)

Paras Jain (1900290100098)

Prashansa Rai (1900290100105)

Under the supervision of

Pushpendra Kumar

KIET Group of Institutions, Ghaziabad

Affiliated to

Dr. A.P.J. Abdul Kalam Technical University, Lucknow

(Formerly UPTU)

May, 2023

DECLARATION

We hereby declare that this submission is our own work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Signature
Name:Mohd. Naseem
Roll No:19290100085
Date:27/05/2023
Signature
Name:Paras Jain
Roll No:19290100098
Date:27/05/2023
Signature
Name: Prashansa Rai
Roll No:1900290100105
Date:27/05/2023

CERTIFICATE

This is to certify that Project Report entitled “Fake News Detection Using Ensemble Learning” which is submitted by Mohd Naseem, Paras Jain, Prashansa Rai , in partial fulfillment of the requirement for the award of degree B. Tech. in Department of Computer Science & Engineering of Dr. A.P.J. Abdul Kalam Technical University, Lucknow is a record of the candidates own work carried out by them under my supervision. The matter embodied in this report is original and has not been submitted for the award of any other degree.

.

Date:27/05/2023 **Supervisor Name: Pushpendra Kumar**

Department of CSE

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the report of the B. Tech Project undertaken during B. Tech. Final Year. We owe special debt of gratitude to Prof. Pushpendra Kumar, Department of Computer Science & Engineering, KIET, Ghaziabad, for his constant support and guidance throughout the course of our work. His sincerity, thoroughness and perseverance have been a constant source of inspiration for us. It is only his cognizant efforts that our endeavors have seen light of the day.

We also take the opportunity to acknowledge the contribution of Dr. Vineet Sharma, Head of the Department of Computer Science & Engineering, KIET, Ghaziabad, for his full support and assistance during the development of the project. We also do not like to miss the opportunity to acknowledge the contribution of all the faculty members of the department for their kind assistance and cooperation during the development of our project.

We also do not like to miss the opportunity to acknowledge the contribution of all faculty members, especially faculty/industry person/any person, of the department for their kind assistance and cooperation during the development of our project. Last but not the least, we acknowledge our friends for their contribution in the completion of the project.

Signature:

Name : Mohd Naseem

Roll No:1900290100085

Date: 27/05/2023

Signature:

Name : Prashansa Rai

Roll No:1900290100105

Date: 27/05/2023

Signature:

Name : Paras Jain

Roll No:1900290100098

Date: 27/05/2023

ABSTRACT

The massive propagation of false information via numerous internet channels has made fake news a serious problem in today's digital era. For information to remain accurate and for decision-making to be made with knowledge, it is essential to be able to recognize and counteract false news. This abstract provides a summary of methods for identifying false news, with a particular emphasis on those that use machine learning and natural language processing (NLP).

In order to discern between trustworthy and false information, fake news detection requires analyzing the textual content and metadata linked to news stories. Identification of language patterns, sentiment analysis, and semantic interpretation of the text all depend critically on NLP approaches. Then, using labelled datasets to train models, machine learning methods are used to enable models to generalize and forecast the veracity of news stories.

Language traits, source reliability, social context, and user interaction are just a few of the features and indications used to identify false news. Linguistic characteristics, which may expose inconsistencies and abnormalities in false news pieces, include lexical and syntactic patterns, readability, and semantic consistency. Social context investigates the patterns of information transmission and user responses to the news piece, while source credibility looks at the track record and reputation of the publishing platform or author. Indicators for user involvement, such as likes, shares, and comments, are also utilized since bogus news often generates strong emotional reactions.

Over the recent years, the generation and exchange of news and information have rapidly risen with the emergence of technology. The major sources of news and information have been social media feeds, news websites, blogs, and articles. A significant part of this news is reportedly fake and unauthentic, which can be misleading and spread fake rumors in public. This fake news is intentionally propagated to damage the reputation of an individual or an organization. This paper tried to solve this issue by applying machine learning to detect and effectively dismantle disinformation.

Keywords: Fake news, Social media, Authenticity, Artificial Intelligence, Logistic regression, Support vector machine, Naïve Bayes algorithm, Random Forest algorithm, Ensemble Learning.

TABLE OF CONTENTS

Page No.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

TF-IDF	Term Frequency-Inverse Document Frequency
Sci-kit learn	sklearn
LR	Logistic Regression
MLP	Multilayer Perceptron
NLTK	Natural Language Toolkit
API	Application Programming Interface
gRPC	Google Remote Procedure Call
ML	Machine Learning
NLP	Natural Language Processing

CHAPTER 1
INTRODUCTION

1.1 INTRODUCTION

In today's culture, fake news, sometimes referred to as disinformation, has emerged as a serious problem since it has the potential to significantly affect how people perceive critical events and situations. The spread of false news on social media and other online platforms has made it harder for a person to tell the difference between reliable and unreliable information sources. This not only causes misunderstandings and confusion, but it may also have detrimental effects including the propagation of false information and public distrust of the media.

The spread of fake news has become a major issue in recent years, especially with the rise of social media and other internet platforms that allow for the speedy dissemination of information. Just a few of the main repercussions of fake news include the spread of incorrect information, the amplification of negative narratives, and the erosion in public faith in the media and political institutions. The ability to spot fake news is now highly valued by society, and machine learning (ML) techniques have emerged as a practical solution.

Researchers have suggested machine learning (ML) approaches for identifying bogus news as a solution to this issue. These algorithms can analyze a lot of data, which makes it possible to find patterns and traits that are suggestive of false news. To categorize articles as legitimate or fraudulent, natural language processing (NLP) methods, such as text classification, have been widely employed.

The purpose of this research is to examine how ML algorithms perform in relation to more established techniques for identifying false news in internet text. We will pay particular attention to text classification techniques and investigate several feature representations, such as bag in order to represent the text data, comprising words and word embeddings. We will also look for methods to include more data sources, such meta-data or user interaction data, to enhance the performance of the ML models.

Manually identifying false news may be difficult and often needs a high degree of subject-matter knowledge. It has been simpler to create and spread false information as a result of recent advances in computer science and technology, but it is much more difficult to tell the difference between fact and fiction. Fake news' spread and virality may have detrimental effects on people' political careers as well as on goods and companies.

Back in February 2020, the WHO reported that the coronavirus pandemic had caused a "infodemic" that was made up of a lot of information, some of which was accurate and some of which was false. People now find it exceedingly difficult to determine whether sources of data or information are true and trustworthy. Too much misleading information may lead to the widespread spread of prejudice, fear, and other negative emotions.[4]

This research proposes a novel method and tool for detecting fake news that makes use of artificial intelligence. The author contrasts the Naive Bayes Classifier, Logistic Regression, and Passive Aggressive Classifier, three supervised classification methods. The dataset, which contains both legitimate and false news, produces encouraging findings.

Large datasets of news stories, social media messages, and other types of online information are analysed for patterns and correlations using ML-based techniques, which make use of algorithms and statistical models. The network structure of social media sites may be analysed using ML-based methods to spot propaganda and disinformation trends.

In recent years, academics have created a variety of machine learning (ML)-based algorithms for detecting false news, including supervised and unsupervised learning techniques as well as deep learning ways to education that can evaluate both word and visual material. These methods have showed promise in identifying bogus news in a range of settings, including online forums, news websites, and social media platforms.

However, identifying fake news is a difficult operation that requires carefully taking into account a variety of elements, such as the veracity of the sources, the context of the material, and any possible biases in the data. Furthermore, because the propagation of false news is a dynamic phenomena, it need constant study and the creation of fresh methods and strategies in order to keep up with emerging strategies.

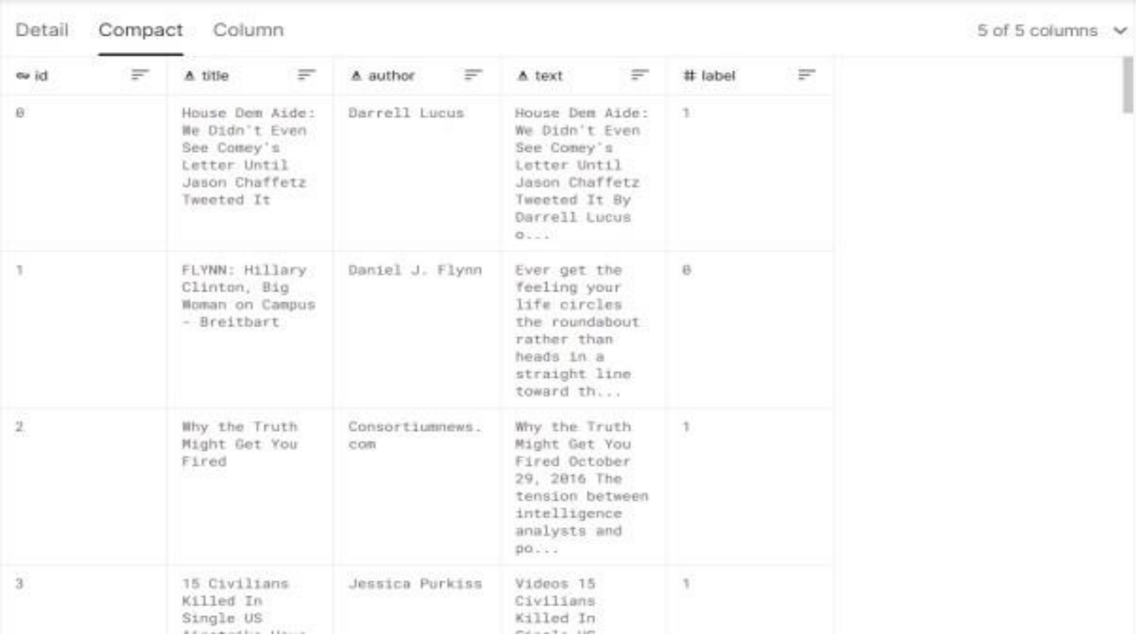
This essay is divided into eight separate sections: an abstract, introduction, thorough literature review, methods section, and conclusion. The technique portion is further separated into a number of subsections, which cover topics like Logistic Regression and its different forms, the Naive Bayes classifier, the Passive Aggressive classifier, Porterstemmer, Sci-kit Learn (sklearn), Tfidvectorizer, Stopwords, Tensorflow, and Android Studio. The dataset utilised is covered in the fifth portion of the article, which is followed by the implementation and outcome section, which contains comparison tables and graphs to help readers fully comprehend the extent of the study. A subsection on simulation is also included in this section. A full list of references is included in the eighth and last portion of the article, which follows the seventh section on the conclusion.

1.2 PROJECT DESCRIPTION

In our fake news detection ML project, we conducted several experiments to determine the best approach for detecting fake news. The following are the key experiments we performed:

1. Dataset of News Article:

To train our machine learning model, we collected a dataset of newsarticles, including real and fake news. We used 80% of the data for training and 20% for testingthe model's performance. Downloaded the dataset from Kaggle and further used that data to train our machine learning model with the help of Logistic Regression Model. After that data pre-processing has been done which is a technique using which the raw data is transformed intoa format which can be easily understood by the computers. The shape or size of the dataset is (20800,5). Dataset contains 20799 rows and 5 columns. Columns are id, title, author, text and label. Different types of news categories are present in the dataset like business, sports, entertainment, social etc. The authenticity of the news lies in the fact that the news has been checked by the journalist and also labelled as “fake” or “real”.



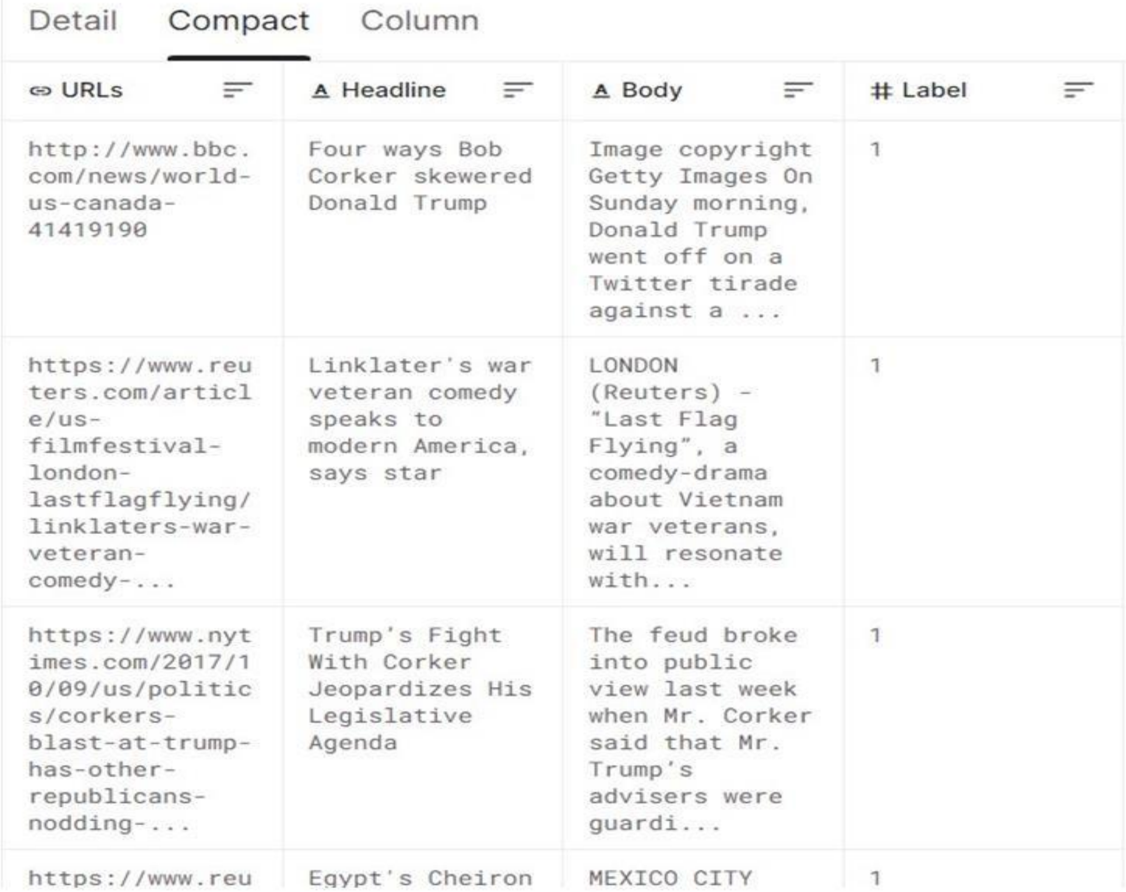
id	title	author	text	label
0	House Dem Aide: We Didn't Even See Comey's Letter Until Jason Chaffetz Tweeted It	Darrell Lucus	House Dem Aide: We Didn't Even See Comey's Letter Until Jason Chaffetz Tweeted It By Darrell Lucus o...	1
1	FLYNN: Hillary Clinton, Big Woman on Campus - Breitbart	Daniel J. Flynn	Ever get the feeling your life circles the roundabout rather than heads in a straight line toward th...	0
2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, 2016 The tension between intelligence analysts and po...	1
3	15 Civilians Killed In Single US	Jessica Purkiss	Videos 15 Civilians Killed In	1

Fig1. Dataset 1

The second dataset (containing 3,352 articles) is accessible on Kaggle. The authentic articles are sourced from The New York Times, CNN, Reuters, and other reputable online sources, whereas false news is sourced from news websites that have been known to publish unreliable information. This dataset covers a variety of topics including politics, entertainment, and sports.
Link to the dataset:: https://www.kaggle.com/datasets/jruvika/fake-news-detection
Description of the dataset:

train.csv: A full training dataset with the following attributes:
- **URL**: Link for the news article
- **Headline**: the title of a news article
- **Body**: the text of the article; could be incomplete
- **label**: a label that marks the article as potentially unreliable
- 1: unreliable
- 0: reliable

test.csv: A testing training dataset with all the same attributes at train.csv without the label.

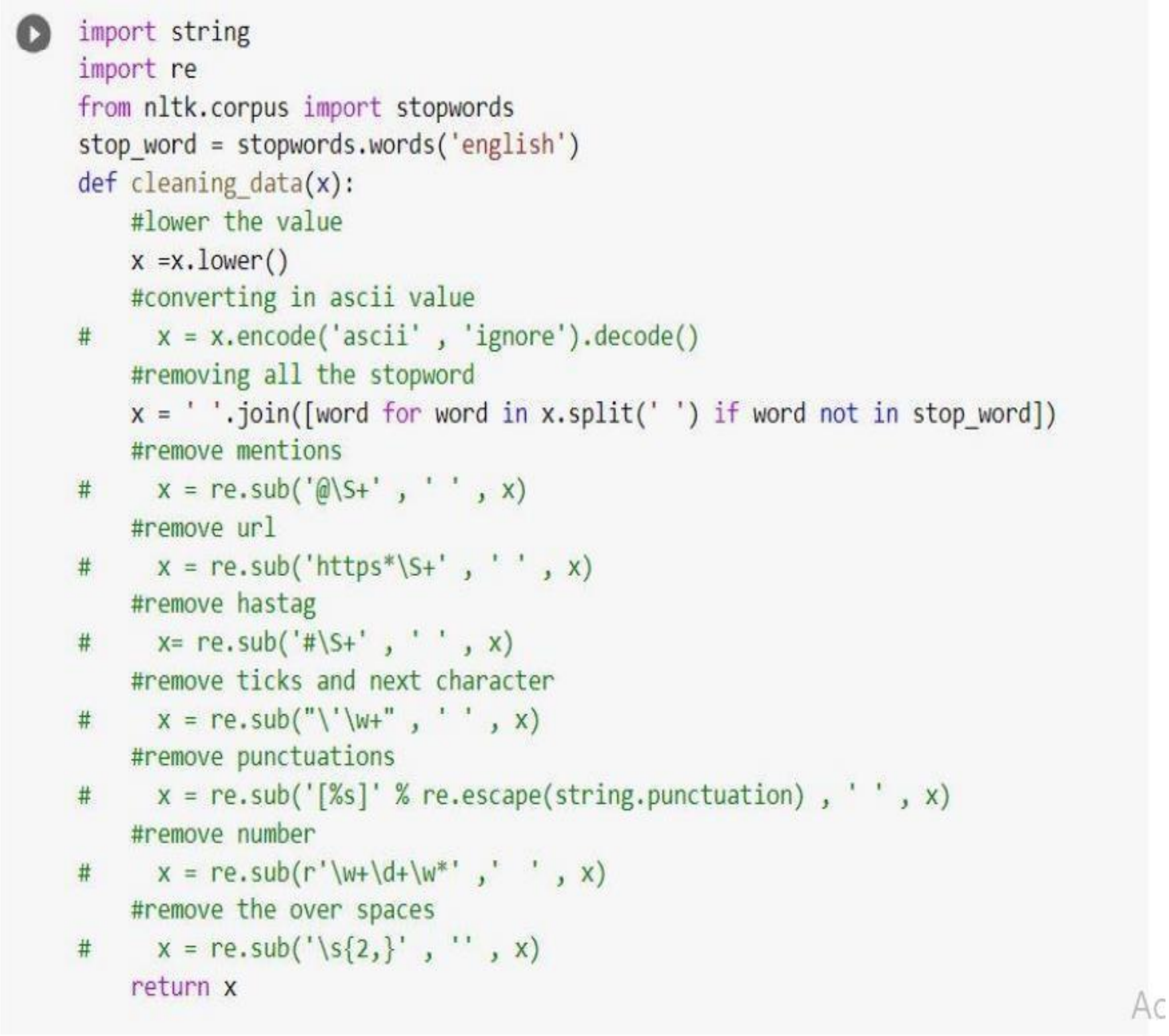


URLs	Headline	Body	Label
http://www.bbc.com/news/world-us-canada-41419190	Four ways Bob Corker skewered Donald Trump	Image copyright Getty Images On Sunday morning, Donald Trump went off on a Twitter tirade against a ...	1
https://www.reuters.com/article/us-filmfestival-london-lastflagflying/linklaters-war-veteran-comedy-...	Linklater's war veteran comedy speaks to modern America, says star	LONDON (Reuters) - "Last Flag Flying", a comedy-drama about Vietnam war veterans, will resonate with...	1
https://www.nytimes.com/2017/10/09/us/politics/corkers-blast-at-trump-has-other-republicans-nodding-...	Trump's Fight With Corker Jeopardizes His Legislative Agenda	The feud broke into public view last week when Mr. Corker said that Mr. Trump's advisers were guardi...	1
https://www.reu	Egypt's Cheiron	MEXICO CITY	1

Fig2. Dataset 2

2. Data Pre-Processing:

We performed several preprocessing steps on the dataset to prepare itfor training. In Data pre-processing tokenization has been done, which means breaking down of a sentence into words. Stop words removal has been performed. Stop words are those wordswhich are abundantly found in the text and are of no use. Lastly, Stemming has been done in which words are reduced by removing suffix part from them.

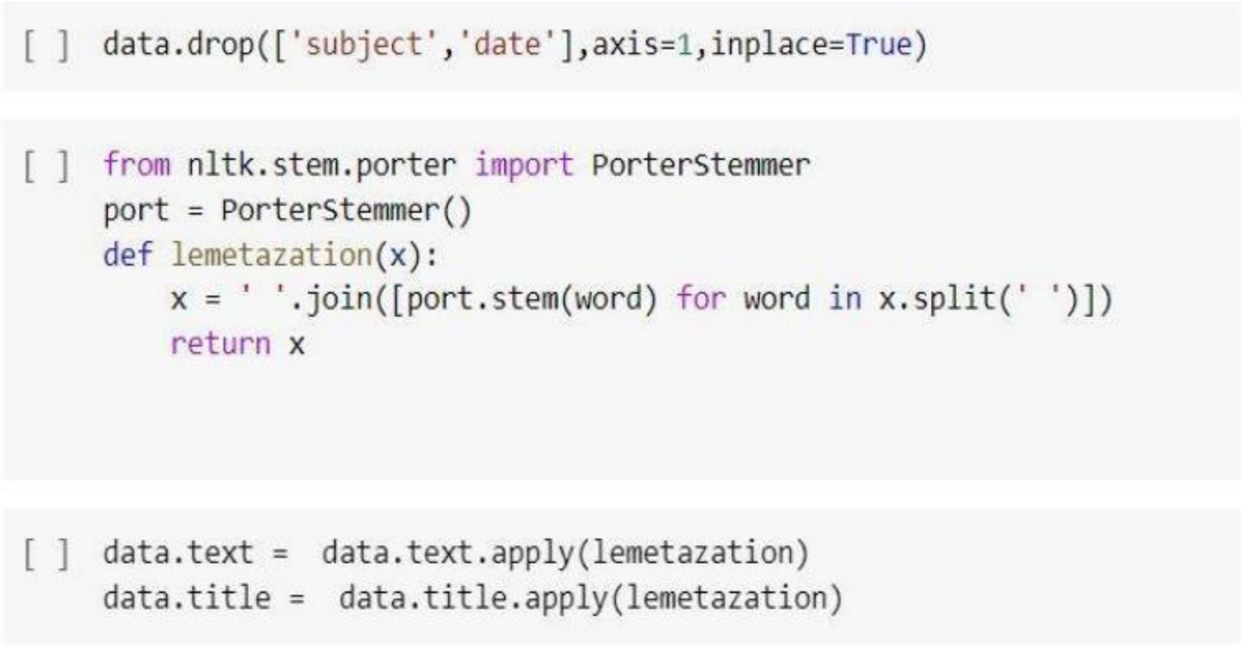


```
import string
import re
from nltk.corpus import stopwords
stop_word = stopwords.words('english')
def cleaning_data(x):
    #lower the value
    x =x.lower()
    #converting in ascii value
#     x = x.encode('ascii' , 'ignore').decode()
    #removing all the stopword
    x = ' '.join([word for word in x.split(' ') if word not in stop_word])
    #remove mentions
#     x = re.sub('@\S+' , ' ' , x)
    #remove url
#     x = re.sub('https*\S+' , ' ' , x)
    #remove hastag
#     x= re.sub('#\S+' , ' ' , x)
    #remove ticks and next character
#     x = re.sub("\'\w+" , ' ' , x)
    #remove punctuations
#     x = re.sub('[%s]' % re.escape(string.punctuation) , ' ' , x)
    #remove number
#     x = re.sub(r'\w+\d+\w*' ,'  ' , x)
    #remove the over spaces
#     x = re.sub('\s{2,}' , '' , x)
    return x
```

Fig 3. Data Preprocessing

3. Split Dataset:

Dataset has been split to training and testing data. For training the model 80%dataset is used and for testing the rest 20% dataset is used.



```
[ ]  data.drop(['subject','date'],axis=1,inplace=True)
```

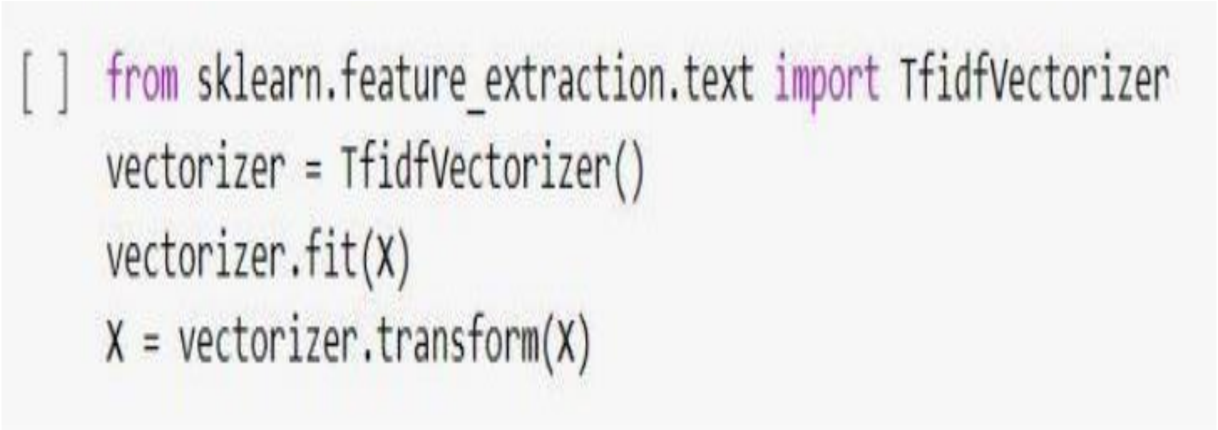
```
[ ]  from nltk.stem.porter import PorterStemmer
     port = PorterStemmer()
     def lemetazation(x):
         x = ' '.join([port.stem(word) for word in x.split(' ')])
         return x
```

```
[ ]  data.text =  data.text.apply(lemetazation)
     data.title =  data.title.apply(lemetazation)
```

Fig 4. Data Splitting

4. Feature Extraction:

Then the encoding of the text data in numerical data is done as wecannot train machine from the text. So, to achieve this, Sklearn library is used to import TfidVectorizer.

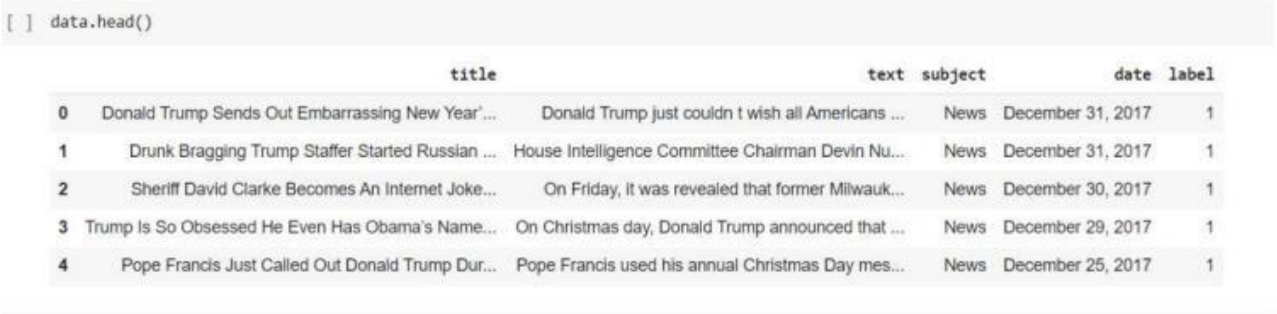


```
[ ]  from sklearn.feature_extraction.text import TfidfVectorizer
     vectorizer = TfidfVectorizer()
     vectorizer.fit(X)
     X = vectorizer.transform(X)
```

Fig 5. Tf-IDF

5. Training the Classifier:

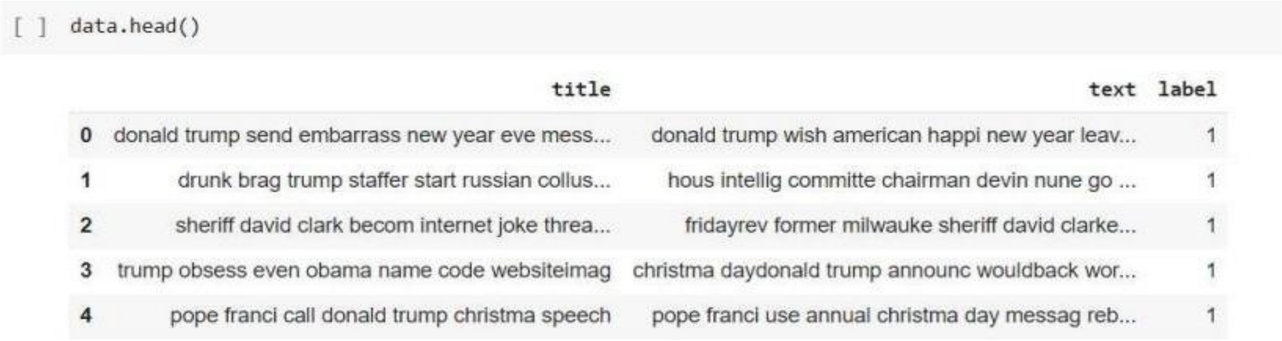
For the purpose of classification label column has been used i.e If the output is ‘0’ hence the news is Real and if it is ‘1’ then it is Fake. To train the model logistic regression classifier had been used. Got an accuracy score of 97.9%.



```
[ ] data.head()
```

	title	text	subject	date	label
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	1
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	1
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	1
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	1
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	1

Fig 6. Data before pre-processing



```
[ ] data.head()
```

	title	text	label
0	donald trump send embarrass new year eve mess...	donald trump wish american happi new year leav...	1
1	drunk brag trump staffer start russian collus...	hous intellig committe chairman devin nune go ...	1
2	sheriff david clark becom internet joke threa...	fridayrev former milwauke sheriff david clarke...	1
3	trump obsess even obama name code websiteimag	christma daydonald trump announc wouldback wor...	1
4	pope franci call donald trump christma speech	pope franci use annual christma day messag reb...	1

Fig 7. Data after pre-processing

6. Parametric Validation: To assess the effectiveness of the trained model, parametric validation is done. Accuracy, precision, recall, and f-measure have been selected as appropriate performance indicators that aid in assessing how well the model performs on test data.Naive Bayes, Logistic Regression, Passive Aggressive, and Multilayer Perceptron were some of the classifiers we worked with. We observed that the Logistic Regression performed the best and had an accuracy of 97.09% after comparing their performances.

CHAPTER 2

LITERATURE REVIEW

The [1] author looks at how people's news consumption has changed as a result of the introduction of new electronic media, such as the internet and digital platforms. The purpose of the research is to provide light on the trends, tastes, and practices of news consumption in the changing media landscape.

Douglas combines qualitative and quantitative research techniques to do this. The study makes use of previously published works and ideas on media use, journalism, and technology development. In order to obtain empirical data and develop a thorough grasp of the topic, the author also undertakes surveys, interviews, and content analysis.

The study's results point to important changes in news consuming habits brought on by the development of electronic media. The growing accessibility of news material through online platforms, the effects of personalized news services, and the function of social media as a news source are all covered in the article. It also discusses how these modifications may affect established news organizations, journalistic standards, and reader engagement.

The study also considers how the change to electronic media may affect democratic procedures and public dialogue. The fragmentation of news consumption, the emergence of echo chambers and filter bubbles, and the difficulties of information verification in the digital era are some of the topics it examines.

The 2016 study report by Jonathan Wong [2] titled "Almost All the Traffic to Fake News Sites is from Facebook: New Data Show" focuses on Facebook's contribution to the growth of fake news websites.

New results from the research show a disturbing pattern: a disproportionate amount of traffic to fake news websites comes from Facebook. Wong investigates the origins of visitors to these websites and evaluates how social media platforms affect the dissemination of false information.

The author emphasizes that Facebook, one of the most widely used social media sites at the time, contributed significantly to the spread of false information. Wong proves that a disproportionately large quantity of traffic to fake news websites was produced via Facebook referrals by examining user behavior and website data.

The results imply that Facebook's news feed optimization and algorithm may have helped false news material become more visible and spread. The article expresses concerns about the possible negative effects of this phenomena, including the dissemination of false information, the decline in faith in reputable news outlets, and its effects on democratic processes and public opinion.

The study article emphasizes the necessity for platforms and consumers to be more discriminating in verifying information and critically analyzing the sources of news material by noting the substantial role that Facebook plays in directing traffic to fake news websites. In order to combat the spread of false news and its possible social effects, the research emphasizes the significance of media literacy and responsible sharing practices.

In general, this study advances knowledge on how social media sites, notably Facebook, are used to spread false information. It highlights how urgent it is to solve this problem and urges further study and projects to lessen how damaging false information is to democratic society and public conversation.

The study [3] covers important facets of false news, such as its definition, traits, and psychological elements that contribute to its dissemination. It covers the topic of how false information may be consciously spread, amplified, and targeted in order to sway public opinion. The authors also look at how social media sites' algorithms contribute to the spread of false information.

The writers provide insights into the effects of false news on society by drawing on empirical research and data analysis. They talk about research showing how people may be duped by misleading information, how disinformation can endure even after being corrected, and how fake news can cause polarization and societal divides.

The study [4] examines several methods for gathering labelled data to train fake news detection algorithms and emphasizes the importance of training datasets. In order to accurately detect false news stories, it emphasizes the significance of combining a variety of lexical, structural, and contextual factors. The report also examines several machine learning models and algorithms that have been used, ranging from conventional methods to more sophisticated deep learning approaches.

The performance assessment of these models is described together with the evaluation criteria for false news detection systems. Given the dynamic nature of false news and the adversarial strategies used by authors, the difficulties of developing reliable and real-time detection algorithms are also discussed.

The revolutionary effect of new electronic media, such as the internet and digital platforms, on how people consume news is highlighted in Andrew Douglas' article [6] on news consumption and the new media. The results show changes in behavior and preferences, highlighting the easier access to news material on online platforms and the significance of social media as a news source. The research highlights crucial questions for conventional media outlets, journalistic standards, and the effects on democratic procedures and public dialogue.

The study also discusses how crucial data sharing, openness, and analysis are in halting the infodemic. In order to give timely and correct information to the public and combat disinformation, it emphasizes the necessity for dependable data sources, sophisticated data analytics, and effective communication techniques.

In order to understand and mitigate the negative consequences of the infodemic during public health emergencies, the study article highlights the crucial role that data-driven initiatives play. It emphasizes how important it is for scientists, public health professionals, and the media to work together to battle false information and advance evidence-based decision-making.

The knowledge acquired from this research may guide future public health emergencies as well as COVID-19 in terms of developing strategies and policies to deal with the problems brought on by the infodemic phenomena. In light of quickly changing health situations, the study advances knowledge of information transmission, data analysis, and communication tactics.

The authors of [7] discuss the rise of false news in the digital age and the necessity for reliable techniques of identification. To capture the semantic content and context of news stories, they suggest the Ex Bake model, which makes use of the capabilities of BERT, a cutting-edge language representation model.

The authors provide evidence for the usefulness of the Ex BAKE model in precisely identifying false news via rigorous testing and review. They demonstrate that Ex BAKE beats them in terms of accuracy, precision, recall, and F1 score by comparing its performance to a number of baseline models.

The study's findings show that BERT-based models, like Ex BAKE, are capable of accurately capturing the subtle language signals and semantic cues found in false news items. The approach may find errors, bias, and misleading information in the text by examining contextual information and word associations.

The research also addresses the ramifications of the Ex BAKE model in real-world settings like social media networks and online news sites. It emphasizes how automated fake news detection systems may help prevent the spread of false information and advance information integrity. One of the well-liked machine learning methods for classification and regression applications is the support vector machine (SVM). This method determines the ideal decision boundary for dividing or categorizing the data into several groups. It is selected to have the greatest possible distance between it and the data points that are furthest away from it. New data may be categorized by which side of the border it is on after the boundary has been established. SVM has a broad variety of applications, including text and picture classification, bioinformatics, and natural language processing[8].

Machine learning models called ensemble learners aggregate the predictions of several smaller models to enhance performance as a whole. Averaging the forecasts, weighting the predictions depending on the performance of each individual model, or training a meta-model to make a final prediction based on the results of the smaller models are just a few of the methods available to do this. Because they may result in large performance increases over a single model, ensemble approaches are often utilized in practice, particularly when the individual models have high bias or high variance[9].

The report also cites four significant research problems that may help future studies focus on the identification of misleading news. First and foremost, there is a big problem with the absence of comprehensive multi-modal datasets for false news identification. Second, multi-modal verification techniques must be developed because, although language approaches have been effective in identifying bogus news, visual presentation also plays a significant role.

Thirdly, the report notes that source verification is a significant research problem since current approaches do not take the source of the news item into account. As a future research challenge, the authors also propose author credibility check, which entails determining the order of news articles published by the same author or group of writers in order to identify false news.

The article [10] clarified the problem of false news on social media and the dangers it presents to the general population and to current events. The authors present an overview of current and previous research in the area and analyze several machine learning methods used for false news identification. They also provide instances of how false information might affect society relationships and views. The article provides insightful information on the issue of false news as well as solutions.

The publication [11] offers a perceptive summary of several methods for identifying false news. The writers describe the many forms of false news and identify four key ways to spot it. The relevant terminologies and prejudices connected to false news are also covered in the study. The information is well researched and presented, making it an important tool for policymakers, practitioners, and scholars. This well-organized and educational work presents hitherto unexplored methods for qualitative and quantitative research on a sizable quantity of false news data.

The LIAR dataset, which is presented by the author [15], is a complete dataset that includes statements that are labelled with associated truth values and specific information on the statement's topic, context, and source. The dataset includes both true and incorrect claims, giving a fair and accurate depiction of actual circumstances.

The technique for gathering and annotating the dataset is covered in the article, including the use of fact-checking agencies and several annotators to guarantee dependability and correctness. Numerous statistical studies are performed to show the dataset's richness and variety, underscoring its potential for assessing the effectiveness of algorithms for detecting false news.

The authors [17] provide a summary of deep learning's core ideas and methods while showing the technology's potential for use in a range of industries, including robotics, computer vision, voice recognition, and natural language processing. They talk about the problems with conventional machine learning methods and stress the value of deep neural networks in overcoming these constraints.

The study explores the architecture of deep neural networks, especially convolutional and recurrent neural networks (CNNs and RNNs), which have shown outstanding performance in applications including image classification, object identification, and sequence modelling. The ideas behind these designs are described by the authors, including the use of several layers, non-linear activation functions, and methods like backpropagation for network training.

The authors also go into detail on the value of huge datasets and computing resources for efficiently training deep neural networks. They also mention the idea of unsupervised learning, where the network picks up knowledge from unlabeled input to find underlying structures and patterns.

The study report acknowledges that deep learning has considerably enhanced artificial intelligence and machine learning, resulting in advances in a variety of disciplines, in its conclusion. The authors stress how the ability for computers to learn and extract useful information from complicated data will enable deep learning approaches to continue advancing technology and revolutionize whole sectors.

The difficulty of efficiently training deep neural networks, which are made up of numerous layers of linked units, is addressed in the study [25]. A particular kind of deep neural network called a "deep belief net" is made up of layers of hidden units and a visible layer that displays the input data. These networks have shown potential in a number of applications, including voice processing and picture identification.

The authors propose a learning algorithm called the "contrastive divergence" algorithm that enables efficient training of deep belief nets. This algorithm combines a form of unsupervised learning known as "restricted Boltzmann machines" (RBMs) with a gradient-based fine-tuning procedure.

The paper provides a detailed explanation of the contrastive divergence algorithm, highlighting its advantages over other existing learning algorithms. The authors demonstrate the algorithm's effectiveness through experimental results on various datasets, showing improved performance in terms of training speed and accuracy compared to previous methods.

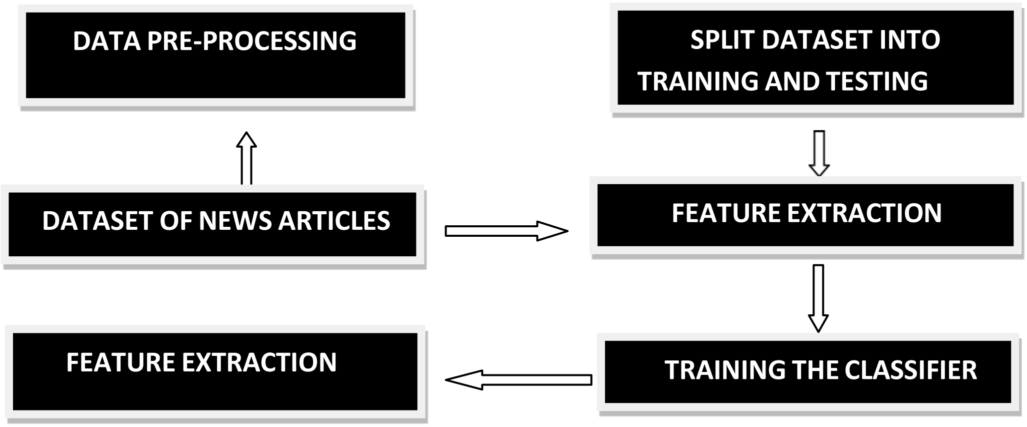
The research presented in this paper contributes to the field of deep learning by introducing a fast and effective learning algorithm for deep belief nets. The findings have implications for the development of more efficient and powerful neural network architectures, paving the way for advancements in areas such as artificial intelligence, pattern recognition, and machine learning.

CHAPTER 3

PROPOSED METHODOLOGY

First convert the textual data into a suitable form for data modeling, there are various techniques that can be employed for converting textual data to numeric data and this is known as data preprocessing which is the first step of the ML model.

Fig 8**.** Flowchart for fake news detection



Downloaded the dataset from Kaggle and further used that data to train our machine learning model with the help of Logistic Regression Model. After that data pre-processing has been done which is a technique using which the raw data is transformed into a format which can be easily understood by the computers.

In Data pre-processing tokenization has been done, which means breaking down of a sentence into words. Stop words removal has been performed. Stop words are those words which are abundantly found in the text and are of no use. Lastly, Stemming has been done in which words are reduced by removing suffix part from them. Dataset has been split to training and testing data. For training the model 80% dataset is used and for testing the rest 20% dataset is used.

Then the encoding of the text data in numerical data is done as we cannot train machine from the text. So, to achieve this, Sklearn library is used to import TfidVectorizer. For the purpose of classification label column has been used i.e If the output is ‘0’ hence the news is Real and if it is ‘1’ then it is Fake. To train the model logistic regression classifier had been used. Got an accuracy score of 97.9%.

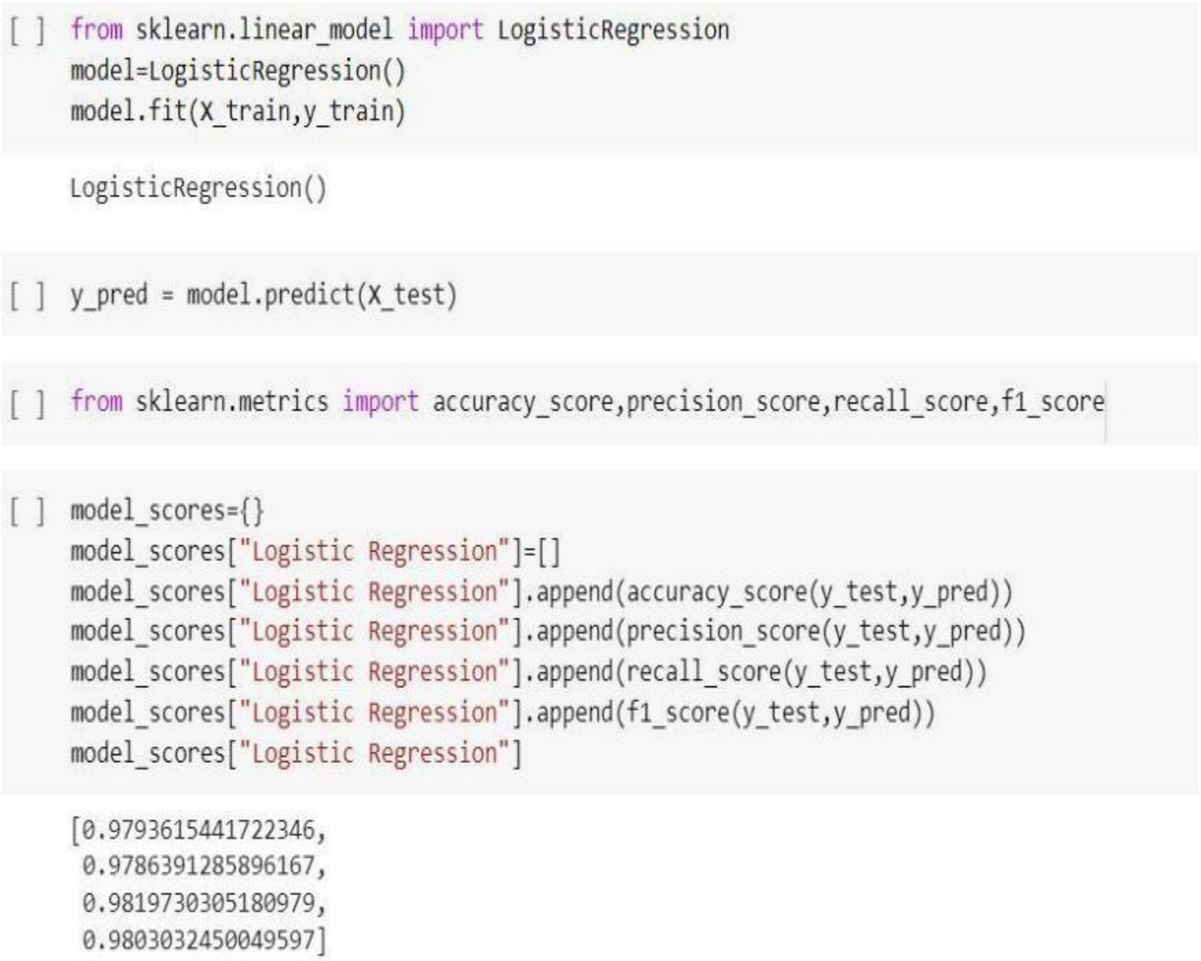
Parametric validation is performed for evaluating the performance of the trained model. Appropriate performance metrics has been chosen that are accuracy, precision, recall and f-measure that helps in evaluating the performance of the model on the testing data.

Lastly, done the integration of ML model with android app. Made a News app which will work as user interface. This app will show news of different categories through API calls. Used a third-party API called News API. In every news article enabled a button to check or verify the news and deployed the ML model on TensorFlow. After deployment used the gRPC (Google Remote Procedure Call) it is a remote procedure call (RPC) framework that brings performance benefits and modern facilities to client-server applications.

3.1 LOGISTIC REGRESSION

Supervised learning technique used for binary classification of data. Fig 2 shows us the Graphical representation of Logistic Regression Model.

In the graph shown below (in Fig 2) the sigmoid logistic function is used to convert the probabilities into binary values which could be further used for making predictions. According to this graph if the value comes less than 0.5 than it comes under the class 0 and if it comes more than 0.5 then it comes under class 1. The value should be between 0 and 1. The concept of threshold value is been used by the Logistic Regression Model, which either defines the probability as 0 or 1.



```
[ ] from sklearn.linear_model import LogisticRegression
    model=LogisticRegression()
    model.fit(X_train,y_train)

    LogisticRegression()

[ ] y_pred = model.predict(X_test)

[ ] from sklearn.metrics import accuracy_score,precision_score,recall_score,f1_score

[ ] model_scores={}
    model_scores["Logistic Regression"]=[]
    model_scores["Logistic Regression"].append(accuracy_score(y_test,y_pred))
    model_scores["Logistic Regression"].append(precision_score(y_test,y_pred))
    model_scores["Logistic Regression"].append(recall_score(y_test,y_pred))
    model_scores["Logistic Regression"].append(f1_score(y_test,y_pred))
    model_scores["Logistic Regression"]

    [0.9793615441722346,
     0.9786391285896167,
     0.9819730305180979,
     0.9803032450049597]
```

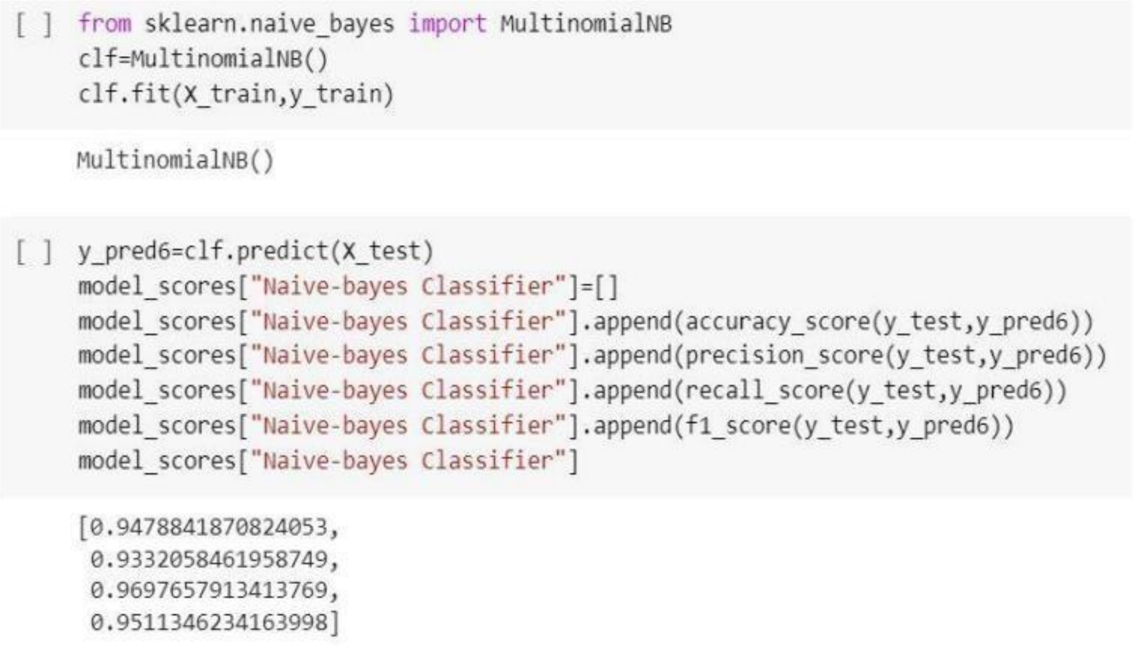
Fig 9. Logistic Regression Implementation

3.2 NAÏVE BAYES CLASSFIER

Naïve Bayes Classifier is similar to the bayes theorem that gives probability of happening of an event based on some given prior condition. Naïve Base Classifier comes under the category of Supervised Learning Algorithm.

Types of Naïve Base Classifier

1. Gaussian
2. Multinomial
3. Binomial



```
[ ] from sklearn.naive_bayes import MultinomialNB
    clf=MultinomialNB()
    clf.fit(X_train,y_train)

    MultinomialNB()
```

```
[ ] y_pred6=clf.predict(X_test)
    model_scores["Naive-bayes Classifier"]=[]
    model_scores["Naive-bayes Classifier"].append(accuracy_score(y_test,y_pred6))
    model_scores["Naive-bayes Classifier"].append(precision_score(y_test,y_pred6))
    model_scores["Naive-bayes Classifier"].append(recall_score(y_test,y_pred6))
    model_scores["Naive-bayes Classifier"].append(f1_score(y_test,y_pred6))
    model_scores["Naive-bayes Classifier"]

    [0.9478841870824053,
     0.9332058461958749,
     0.9697657913413769,
     0.9511346234163998]
```

Fig 10.Naïve Bayes Classifier Implementation

The formula that is used in bayes theorem is given below

$$P(A/B) = P(B/A) * P(A) / P(B) \quad (1)$$

P(A/B): Probability of A when B is already happened

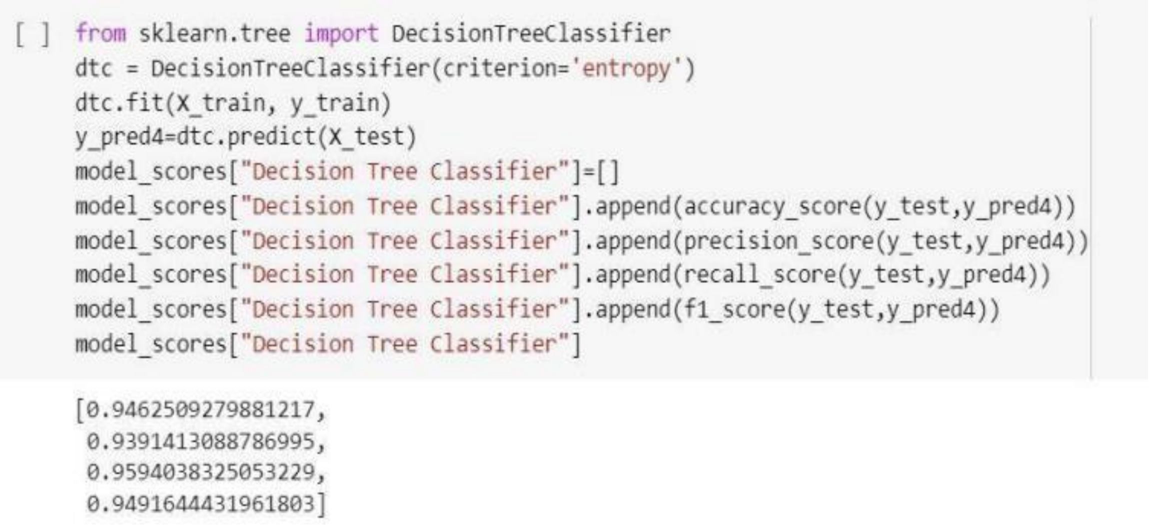
P(B/A): Probability of B when A is already happened

3.3 PASSIVE AGGRESSIVE CLASSIFIER

It is used for both regression as well as classification. For correct classification it remains passive and becomes aggressive for incorrect classification

3.4 DECISION TREE

Each internal node in a decision tree represents a feature or attribute, and each branch represents one or more potential values for that property. The projected class or value is represented by the tree's leaves or terminal nodes. At each stage of the building process, the best attribute is chosen to divide the data, with the goal of maximising the homogeneity or purity of the resultant subsets.



```
[ ]  from sklearn.tree import DecisionTreeClassifier
     dtc = DecisionTreeClassifier(criterion='entropy')
     dtc.fit(X_train, y_train)
     y_pred4=dtc.predict(X_test)
     model_scores["Decision Tree Classifier"]=[]
     model_scores["Decision Tree Classifier"].append(accuracy_score(y_test,y_pred4))
     model_scores["Decision Tree Classifier"].append(precision_score(y_test,y_pred4))
     model_scores["Decision Tree Classifier"].append(recall_score(y_test,y_pred4))
     model_scores["Decision Tree Classifier"].append(f1_score(y_test,y_pred4))
     model_scores["Decision Tree Classifier"]
```

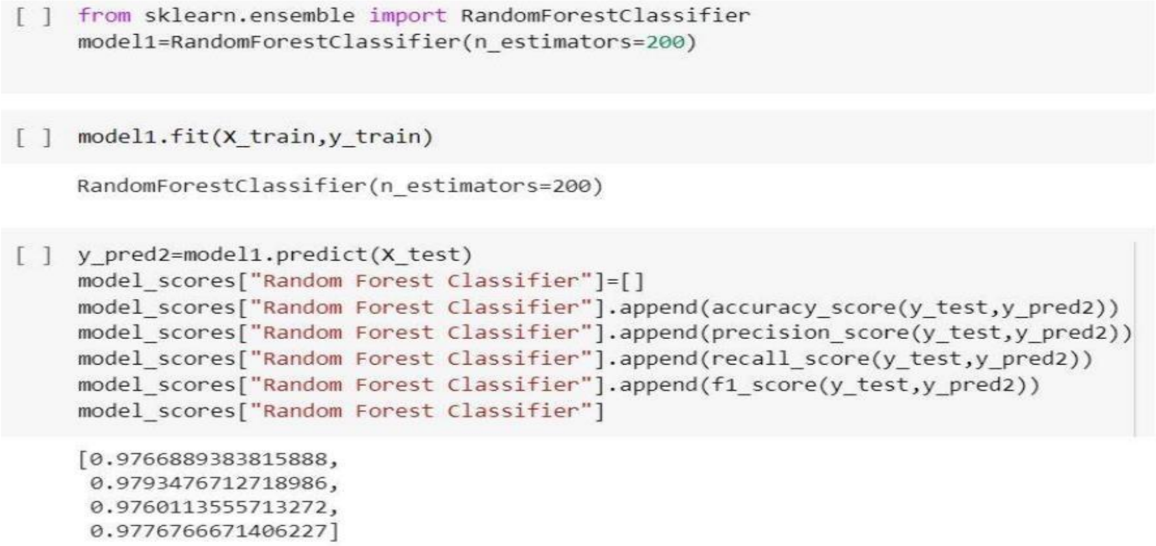
```
[0.9462509279881217,
 0.9391413088786995,
 0.9594038325053229,
 0.9491644431961803]
```

Fig 11.Decision Tree Implementation

3.5 RANDOM FOREST

Each decision tree in a Random Forest ensemble is trained using a different subset of the training data and features. This is how Random Forest works. Each decision tree separately learns to predict the target variable throughout the training phase using various combinations of the input characteristics. By combining all of the individual trees' predictions, either by voting (for classification) or averaging (for regression), the Random Forest's final forecast is established.

Random Forest's capacity to handle complicated datasets with high dimensionality and plenty of characteristics is its main benefit. It is resistant to overfitting because it can accurately represent non-linear correlations and interactions between variables. Furthermore, Random Forest offers estimates of feature relevance, which enables feature selection and interpretation, and it can deal with missing data and outliers.



```
[ ]  from sklearn.ensemble import RandomForestClassifier
     model1=RandomForestClassifier(n_estimators=200)
```

```
[ ]  model1.fit(X_train,y_train)

     RandomForestClassifier(n_estimators=200)
```

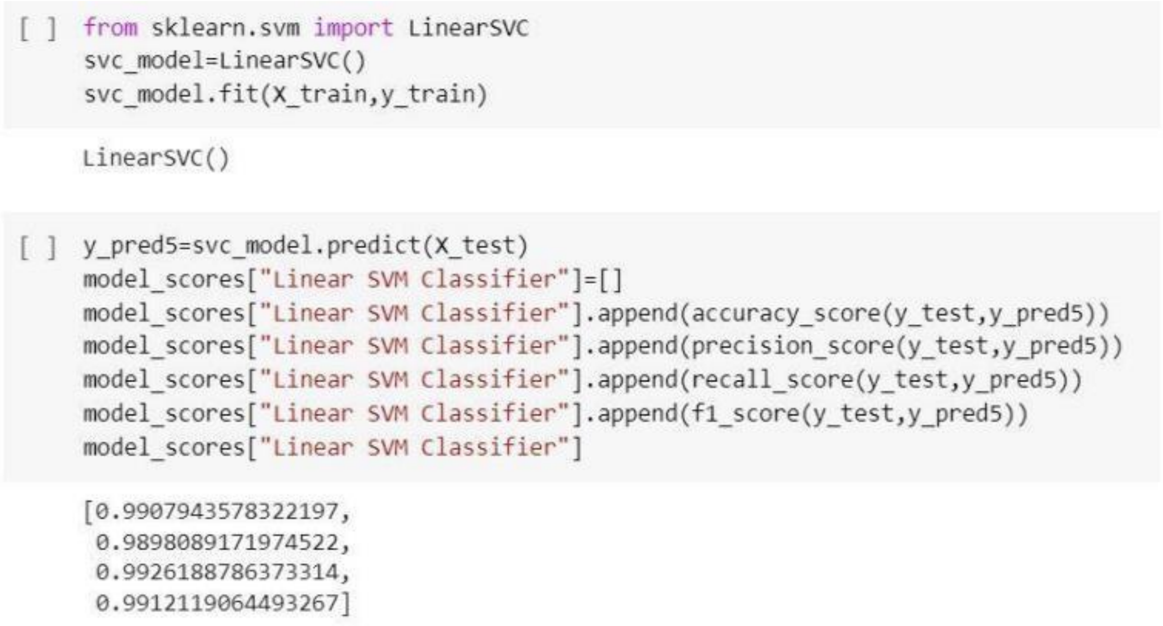
```
[ ]  y_pred2=model1.predict(X_test)
     model_scores["Random Forest Classifier"]=[]
     model_scores["Random Forest Classifier"].append(accuracy_score(y_test,y_pred2))
     model_scores["Random Forest Classifier"].append(precision_score(y_test,y_pred2))
     model_scores["Random Forest Classifier"].append(recall_score(y_test,y_pred2))
     model_scores["Random Forest Classifier"].append(f1_score(y_test,y_pred2))
     model_scores["Random Forest Classifier"]

     [0.9766889383815888,
      0.9793476712718986,
      0.9760113555713272,
      0.9776766671406227]
```

fig 12. Random Forest Implementation

3.6 SUPPORT VECTOR MACHINE(SVM)

A strong and popular machine learning approach for classification and regression problems is called Support Vector Machine (SVM). It works best in situations when the data cannot be separated linearly or if the patterns are intricate. In order for SVM to function, an ideal hyperplane must be found that optimally divides distinct classes in the feature space.

The fundamental principle underlying SVM is to use a kernel function to transfer the input data into a higher-dimensional space where the classes may be distinguished by a hyperplane. The greatest margin hyperplane, or the distance between the hyperplane and the nearest data points from each class, is what SVM seeks to identify. By maximizing the margin, the model's robustness and generalizability are improved.



```
[ ]  from sklearn.svm import LinearSVC
     svc_model=LinearSVC()
     svc_model.fit(X_train,y_train)

     LinearSVC()
```

```
[ ]  y_pred5=svc_model.predict(X_test)
     model_scores["Linear SVM Classifier"]=[]
     model_scores["Linear SVM Classifier"].append(accuracy_score(y_test,y_pred5))
     model_scores["Linear SVM Classifier"].append(precision_score(y_test,y_pred5))
     model_scores["Linear SVM Classifier"].append(recall_score(y_test,y_pred5))
     model_scores["Linear SVM Classifier"].append(f1_score(y_test,y_pred5))
     model_scores["Linear SVM Classifier"]

     [0.9907943578322197,
      0.9898089171974522,
      0.9926188786373314,
      0.9912119064493267]
```

Fig 13. Support Vector Machine Implementation

3.7 ENSEMBLE LEARNERS

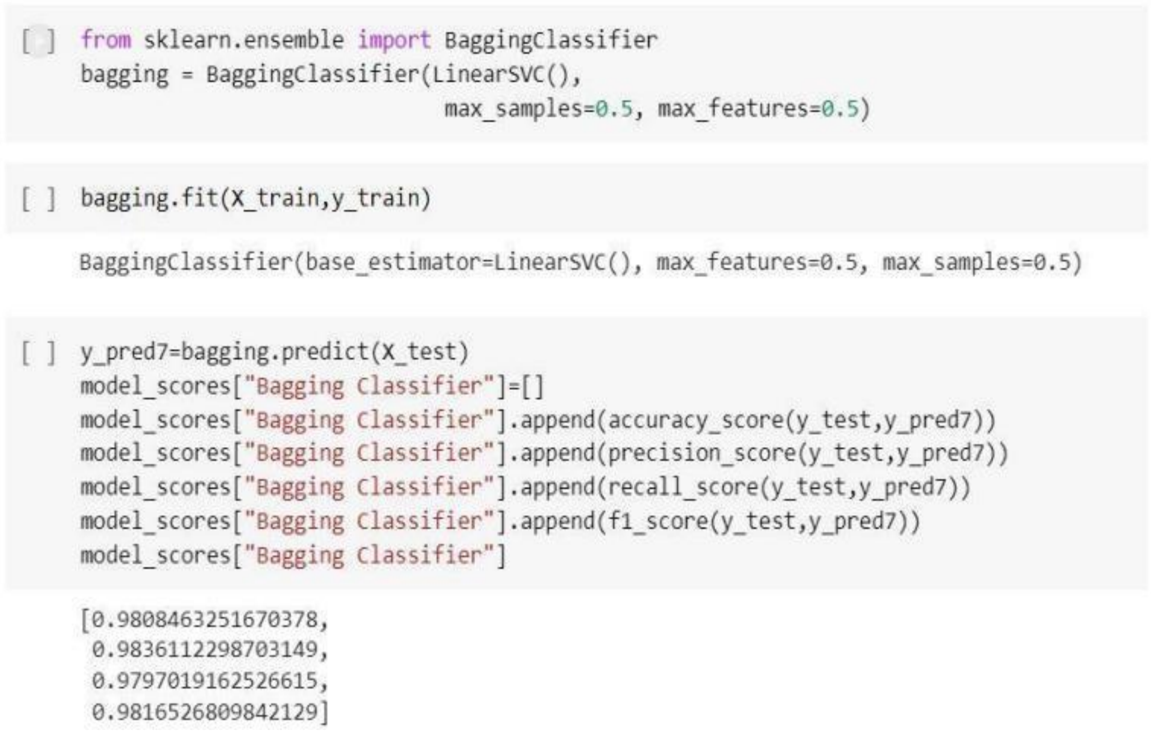
individual models to improve overall predictive performance. Instead of relying on a single model, ensemble learners leverage the diversity and collective intelligence of multiple models to make more accurate predictions.

The idea behind ensemble learning is that different models may have varying strengths and weaknesses, and by combining them, the weaknesses of one model can be compensated by the strengths of others. Ensemble methods can be used for both classification and regression tasks and have been proven to be effective in improving prediction accuracy and generalization.

3.8 Bagging

In bagging, several models, also known as base models or weak learners, are separately trained on various subsets of the training data. The original training data is randomly sampled with replacement to construct each subset, enabling certain occurrences to appear more than once and others to be deleted. The method is referred to as bootstrap sampling.

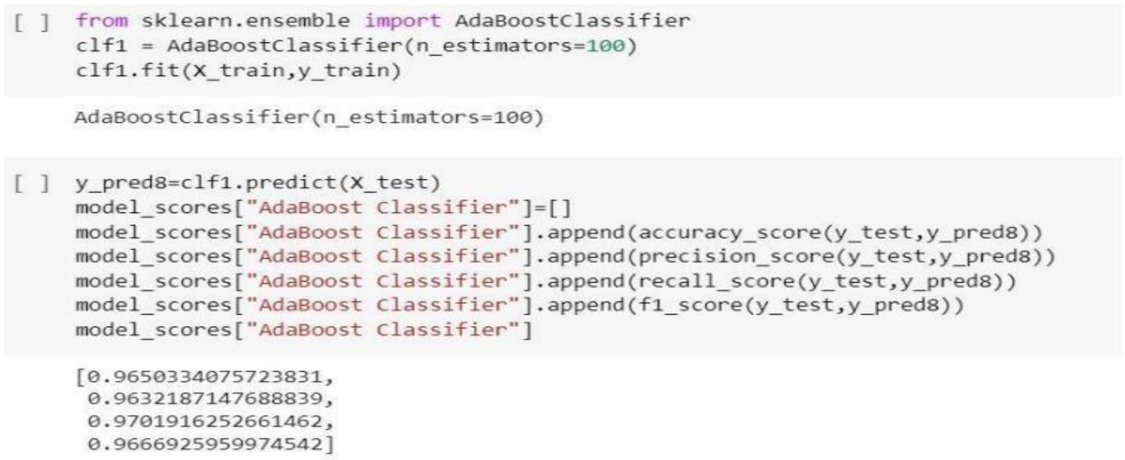
Each model gains the ability to predict outcomes based on its unique bootstrap sampling of the data during training. Bagging aggregates all of the models' predictions for unknown occurrences by voting (classification) or averaging (regression) to arrive at the final forecast. The total variance is decreased, and the effects of individual model flaws are lessened thanks to this ensemble technique.



```
[ ] from sklearn.ensemble import BaggingClassifier
    bagging = BaggingClassifier(LinearSVC(),
                                max_samples=0.5, max_features=0.5)

[ ] bagging.fit(X_train,y_train)

    BaggingClassifier(base_estimator=LinearSVC(), max_features=0.5, max_samples=0.5)

[ ] y_pred7=bagging.predict(X_test)
    model_scores["Bagging Classifier"]=[]
    model_scores["Bagging Classifier"].append(accuracy_score(y_test,y_pred7))
    model_scores["Bagging Classifier"].append(precision_score(y_test,y_pred7))
    model_scores["Bagging Classifier"].append(recall_score(y_test,y_pred7))
    model_scores["Bagging Classifier"].append(f1_score(y_test,y_pred7))
    model_scores["Bagging Classifier"]

    [0.9808463251670378,
     0.9836112298703149,
     0.9797019162526615,
     0.9816526809842129]
```

Fig 14.Bagging Impelementation

3.9 Boosting



```
[ ] from sklearn.ensemble import AdaBoostClassifier
    clf1 = AdaBoostClassifier(n_estimators=100)
    clf1.fit(X_train,y_train)

    AdaBoostClassifier(n_estimators=100)

[ ] y_pred8=clf1.predict(X_test)
    model_scores["AdaBoost Classifier"]=[]
    model_scores["AdaBoost Classifier"].append(accuracy_score(y_test,y_pred8))
    model_scores["AdaBoost Classifier"].append(precision_score(y_test,y_pred8))
    model_scores["AdaBoost Classifier"].append(recall_score(y_test,y_pred8))
    model_scores["AdaBoost Classifier"].append(f1_score(y_test,y_pred8))
    model_scores["AdaBoost Classifier"]

    [0.9650334075723831,
     0.9632187147688839,
     0.9701916252661462,
     0.9666925959974542]
```

Fig 15. Boosting Implementation

3.10 TF-IDF

TfidVectorizer from the Sklearn package has been loaded in order to convert the textual input to numerical data. TF stands for "Term Frequency" and IDF for "Inverse Document Frequency." aids in determining the importance of a certain word in a text.

Term Frequency - It determines how often a certain word appears in a text. The inverse document frequency gauges how often a word appears in a document and how common it is.

3.11 STOPWORDS

The Words which do not add any significance to the text document are called the Stopwords. Examples of few stopwords are “a”, “an”, “the”, “so”, “what”. For this purpose, Stopwords have been imported from NLTK (Natural Language Toolkit).

Why do we remove stop words?

Reducing noise:

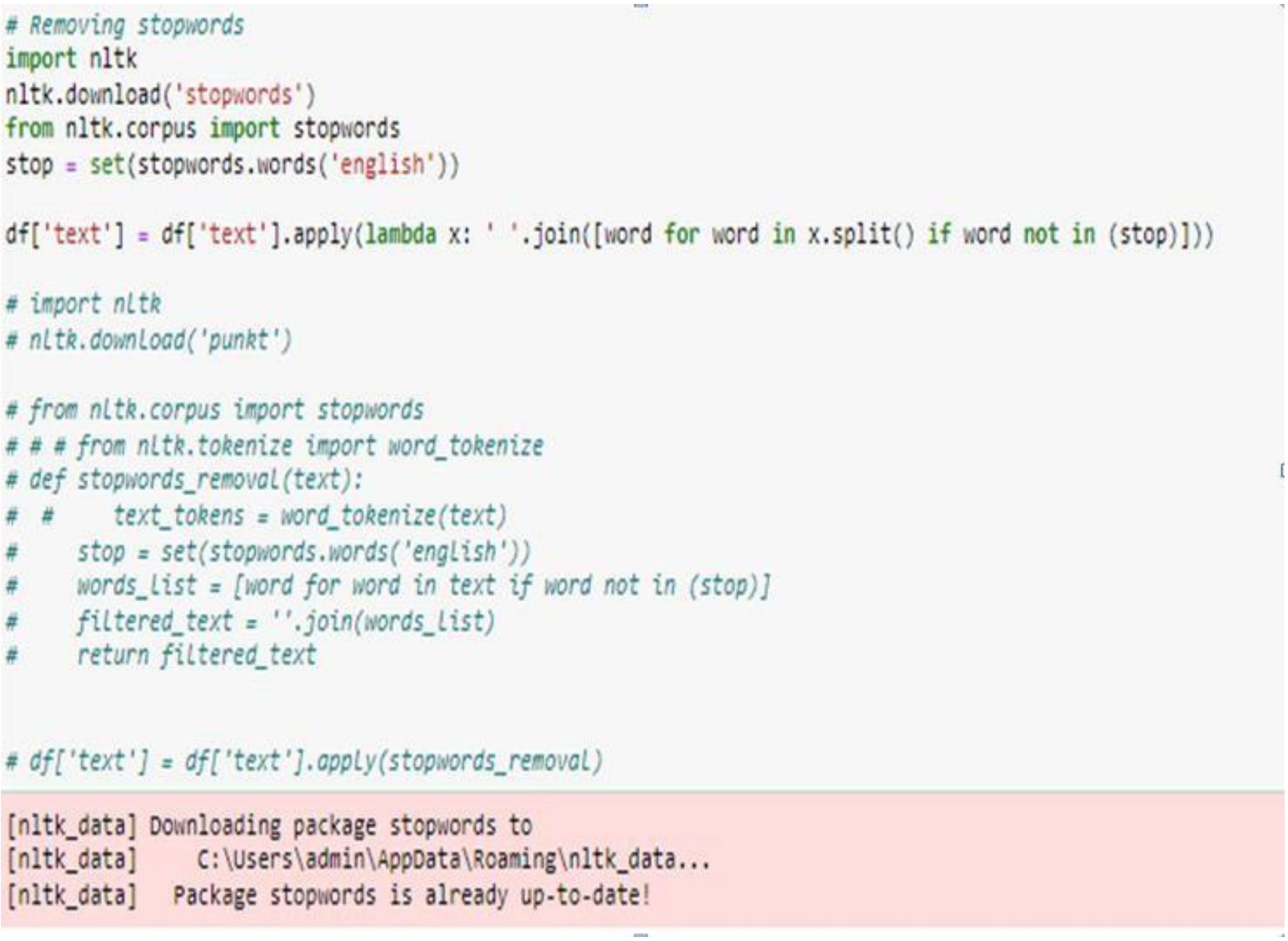
Stop words are frequently occurring words that add little value to the overall meaning of a sentence. By removing them, we can reduce the noise in the data and focus on the more important and meaningful words.

Improving efficiency:

In many natural language processing tasks, such as text classification or information retrieval, the size of the text corpus can be significant. Removing stop words helps reduce the total number of words in the dataset, which can lead to more efficient processing and analysis. It reduces the computational overhead and can speed up algorithms and models.

Reducing dimensionality:

Stop words are typically very common words that appear in almost every document. When performing tasks like text clustering or topic modeling, the presence of these common words may not contribute much to distinguishing between documents or identifying meaningful topics. By removing stop words, we can focus on the words that are more informative and help capture the essence of the text.



```
# Removing stopwords
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
stop = set(stopwords.words('english'))

df['text'] = df['text'].apply(lambda x: ' '.join([word for word in x.split() if word not in (stop)]))

# import nltk
# nltk.download('punkt')

# from nltk.corpus import stopwords
# # # from nltk.tokenize import word_tokenize
# def stopwords_removal(text):
#  #     text_tokens = word_tokenize(text)
#     stop = set(stopwords.words('english'))
#     words_list = [word for word in text if word not in (stop)]
#     filtered_text = ''.join(words_list)
#     return filtered_text


# df['text'] = df['text'].apply(stopwords_removal)
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\admin\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

Fig 16. Stopwords Implementation

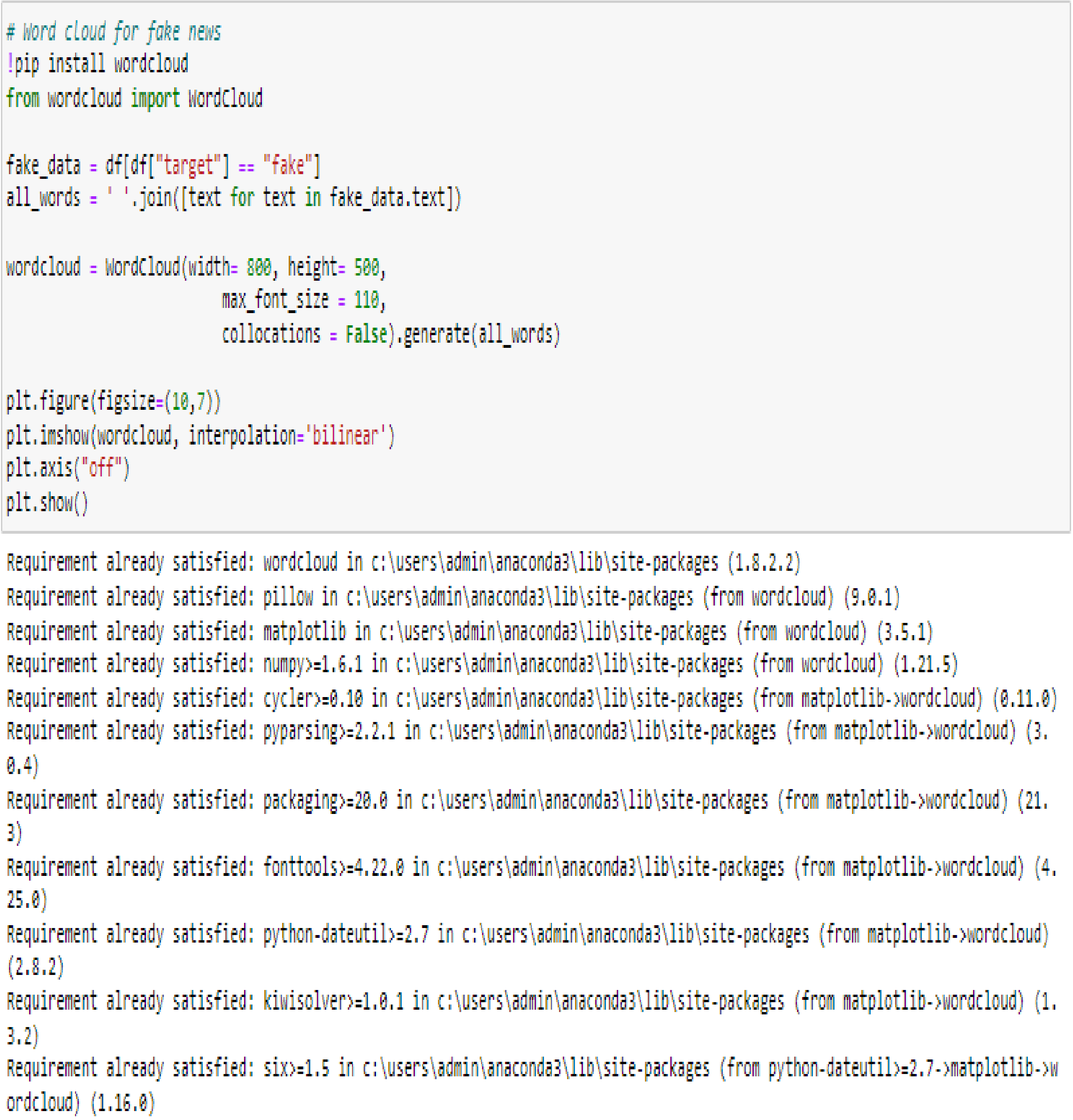
3.12 WORDCLOUD

The most common term is shown in the largest font size in a word cloud, which is a method of visualizing text data. We will discover how to make a unique word cloud in Python in this article. The ideas of creating a word cloud are straightforward and can be loosely broken down into the following steps:

Segment text data first. Additionally, this is the initial stage of NLP text processing. Processing stop words is the major focus of the word cloud process test() function.

Next, create a hash table based on the frequency of each term in the text. Word count is the initial use case for many distributed computing systems, and word frequency calculation has the same significance as hello world programmers in different languages.

The third step is to build a visual arrangement proportionately depending on the word frequency value. The word cloud algorithm and the fundamental component of the word cloud data visualization technique are represented by the class IntegralOccupancyMap.



```
# Word cloud for fake news
!pip install wordcloud
from wordcloud import WordCloud

fake_data = df[df["target"] == "fake"]
all_words = ' '.join([text for text in fake_data.text])

wordcloud = WordCloud(width= 800, height= 500,
                      max_font_size = 110,
                      collocations = False).generate(all_words)

plt.figure(figsize=(10,7))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
```

```
Requirement already satisfied: wordcloud in c:\users\admin\anaconda3\lib\site-packages (1.8.2.2)
Requirement already satisfied: pillow in c:\users\admin\anaconda3\lib\site-packages (from wordcloud) (9.0.1)
Requirement already satisfied: matplotlib in c:\users\admin\anaconda3\lib\site-packages (from wordcloud) (3.5.1)
Requirement already satisfied: numpy>=1.6.1 in c:\users\admin\anaconda3\lib\site-packages (from wordcloud) (1.21.5)
Requirement already satisfied: cycler>=0.10 in c:\users\admin\anaconda3\lib\site-packages (from matplotlib->wordcloud) (0.11.0)
Requirement already satisfied: pyparsing>=2.2.1 in c:\users\admin\anaconda3\lib\site-packages (from matplotlib->wordcloud) (3.
0.4)
Requirement already satisfied: packaging>=20.0 in c:\users\admin\anaconda3\lib\site-packages (from matplotlib->wordcloud) (21.
3)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\admin\anaconda3\lib\site-packages (from matplotlib->wordcloud) (4.
25.0)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\admin\anaconda3\lib\site-packages (from matplotlib->wordcloud)
(2.8.2)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\admin\anaconda3\lib\site-packages (from matplotlib->wordcloud) (1.
3.2)
Requirement already satisfied: six>=1.5 in c:\users\admin\anaconda3\lib\site-packages (from python-dateutil>=2.7->matplotlib->w
ordcloud) (1.16.0)
```

Fig 18. WordCloud Implementation

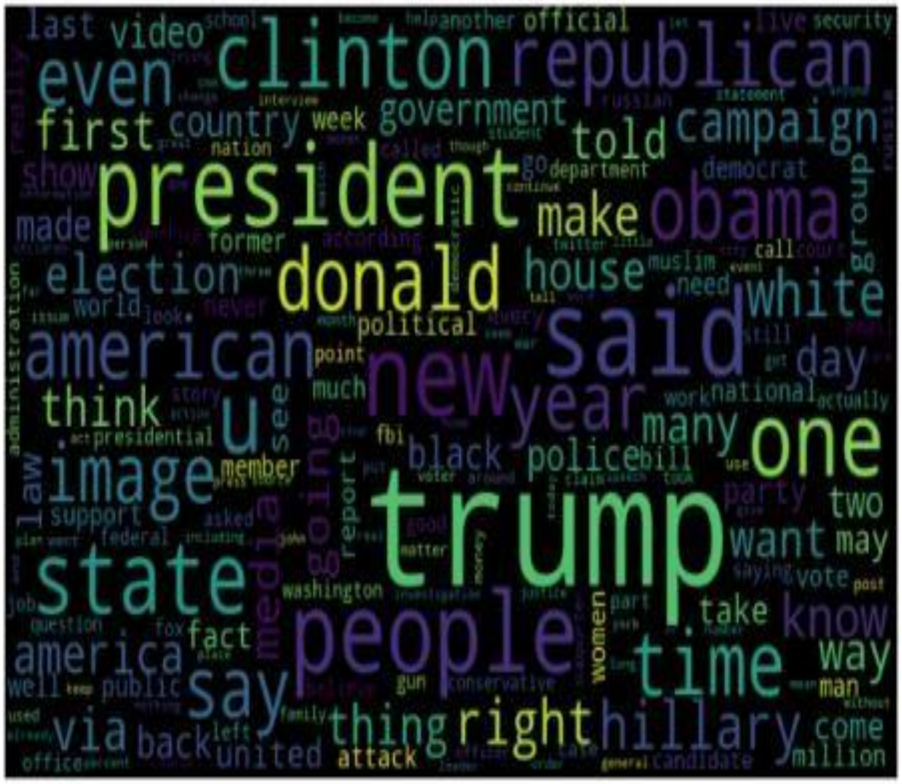


Fig 19. WordCloud

CHAPTER 4

RESULTS AND DISCUSSION

The outcomes of our experiment show that machine learning may be an effective technique for identifying false information. We were able to develop a model that can precisely differentiate between authentic and false news items by combining preprocessing approaches, feature extraction techniques, and machine learning algorithms. This strategy may help stop the spread of false information and lessen its detrimental consequences on society.

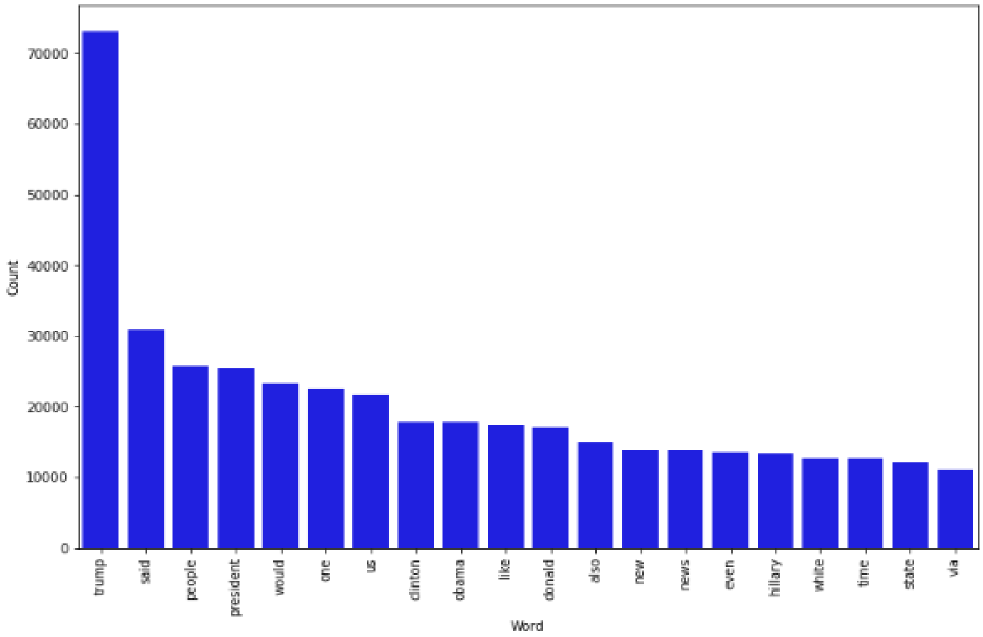


Fig 20. Graph of count of words in news

Table 1. Naïve Base Classifier Confusion Matrix

n=8984	Predicted Yes	Predicted No
Actual Yes (4290)	True Positive (3973)	False Negative (317)
Actual No (4694)	False Positive (218)	True Negative (4476)

Table -1 shows that the Actual Real News column preset in the dataset are “4290”, out of which the Naïve Base Classifier predicted yes are “3973” and predicted no are “317”. Actual Fake News present are “469”, out of which it predicted yes are “18” and predicted no are “4476”.

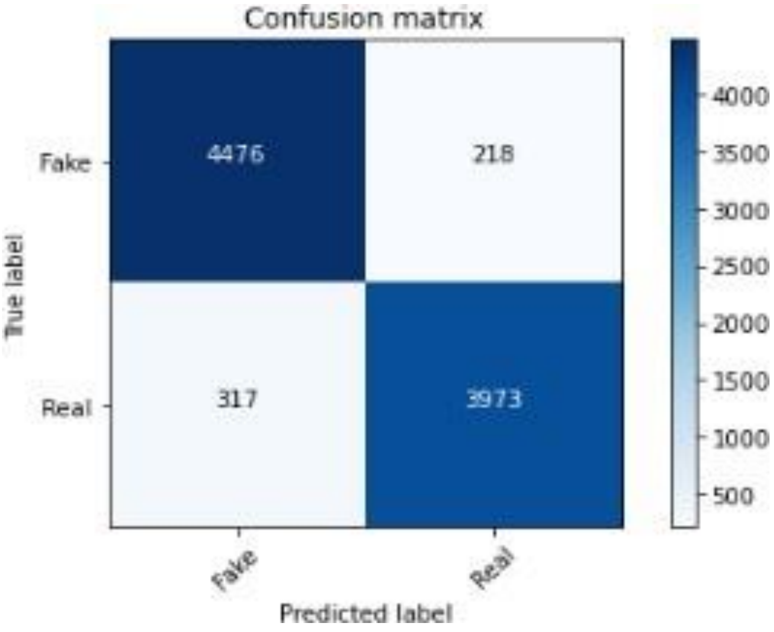


Fig 21. Naïve Bayes Confusion Matrix

Table 2. Decision Tree Classifier Confusion Matrix

n=8984	Predicted Yes	Predicted No
Actual Yes (4290)	True Positive (4267)	False Negative (23)
Actual No (4694)	False Positive (21)	True Negative (4673)

Table - 2 shows that the Actual Real News columns present in the dataset are "4290", out of which the Decision Tree Classifier predicted yes are "4267" and predicted no are "23". Actual Fake News present are "4694", out of which it predicted yes are "21" and predicted no are "4673".

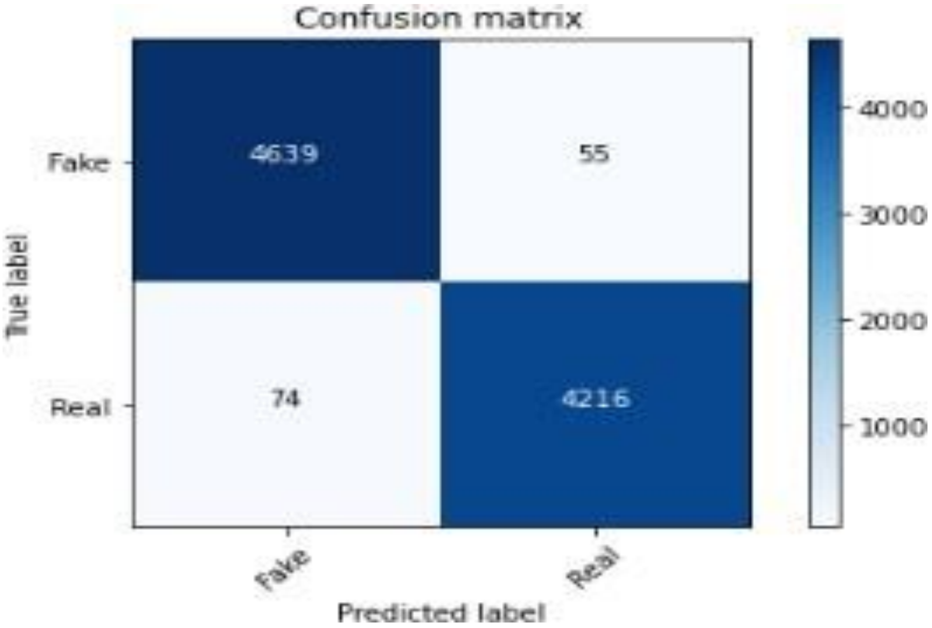


Fig 22. Decision Tree Confusion Matrix

Table 3. Logistic Regression Classifier Confusion Matrix

n=8984	Predicted Yes	Predicted No
Actual Yes (4290)	True Positive (4239)	False Negative (51)
Actual No (4694)	False Positive (56)	True Negative (4638)

Table - 3 shows that the Actual Real News columns present in the dataset are “4290”, out of which the Logistic Regression Classifier predicted yes are “4239” and predicted no are “51”. Actual Fake News present are “4694”, out of which it predicted yes are “56” and predicted no are“4638”.

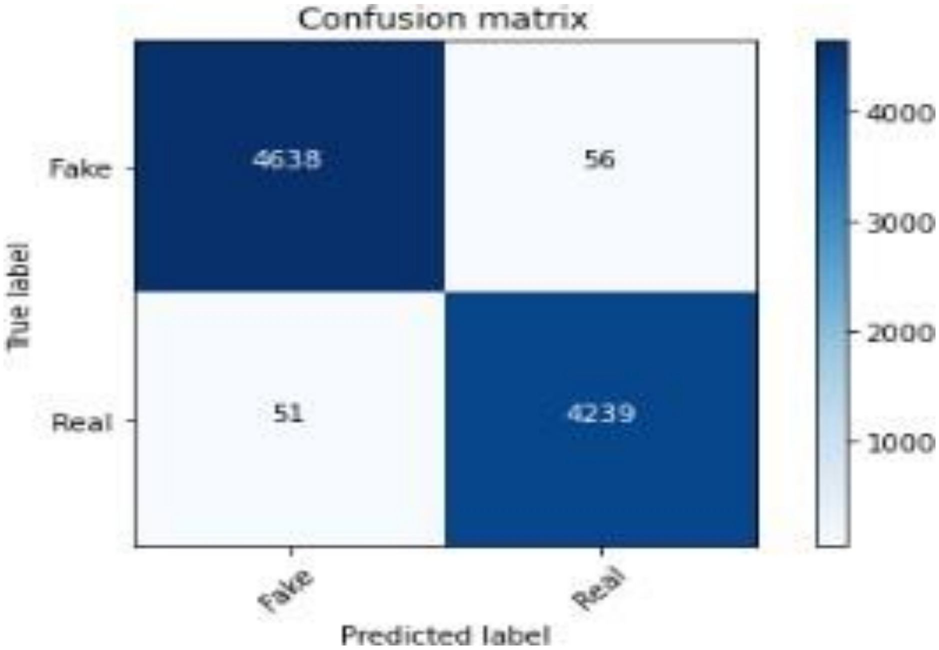


Fig 23. Logistic Regression Confusion Matrix

Table 4. Random Forest Classifier Confusion Matrix

n=8984	Predicted Yes	Predicted No
Actual Yes (4290)	True Positive (4216)	False Negative (74)
Actual No (4694)	False Positive (55)	True Negative (4639)

Table - 4 shows that the Actual Real News present in the dataset are “4290”, out of which the Random Forest Classifierpredicted yes are “421” and predicted no are “74”. Actual Fake News present are “4694”, out of which it predicted yes are “55” and predicted no are “463”

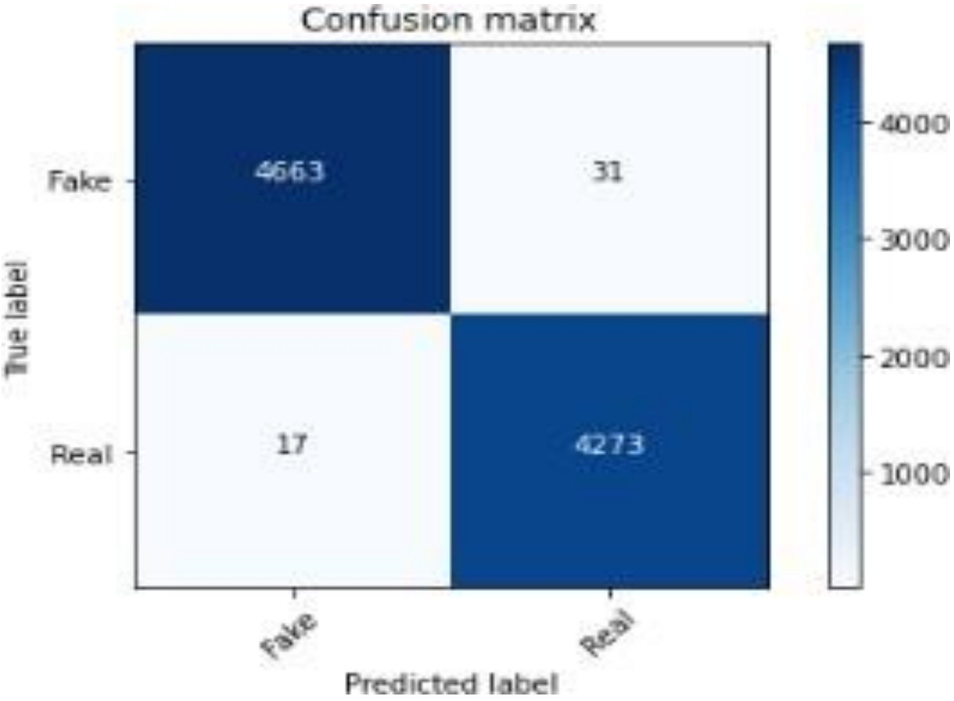


Fig 24. Random Forest Confusion Matrix

Table 5. SVM Classifier Confusion Matrix

n=8984	Predicted Yes	Predicted No
Actual Yes (4290)	True Positive (4273)	False Negative (17)
Actual No (4694)	False Positive (31)	True Negative (4663)

Table - 5 shows that the Actual Real News columns present in the dataset are “4290”, out of which the SVM Classifier predicted yes are “4273” and predicted no are “17”. Actual Fake News present are “4694”, out of which it predicted yes are “31” and predicted noare “4663”

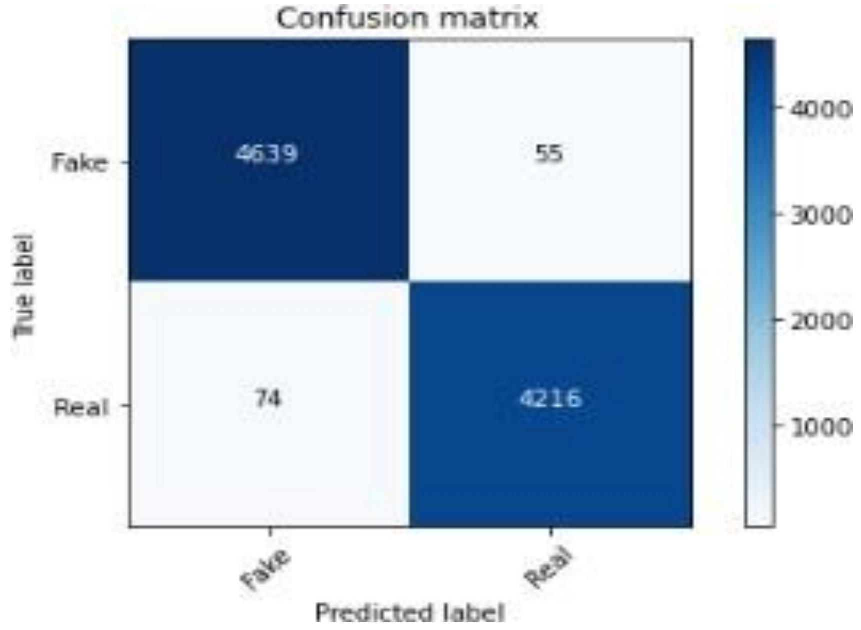


Fig 25. SVM Classifier Confusion Matrix

Table 7. Comparison with other existing approaches

Classifier	Accuracy
Naïve Base Classifier	94.04%
SVM Classifier	99.4%
Logistic Regression Classifier	98.18%
Decision Tree Classifier	99.5%
Random Forest Classifier	98.51%

The comparison of the accuracy of the four classifiers is shown in Table 7. The accuracy of the Nave Base Classifier is "94.04"%, that of the SVM Classifier is "99.4%", that of the Logistic Regression Classifier is "98.18%", that of the Decision Tree Classifier is "98.18%", and that of the Random Forest Classifier is "98.51%," according to the table. The Decision Tree Classifier has the maximum accuracy of 99.5%, as seen in the table. There are 99.6% possibilities that a user-submitted news item will be categorized according to its real nature.

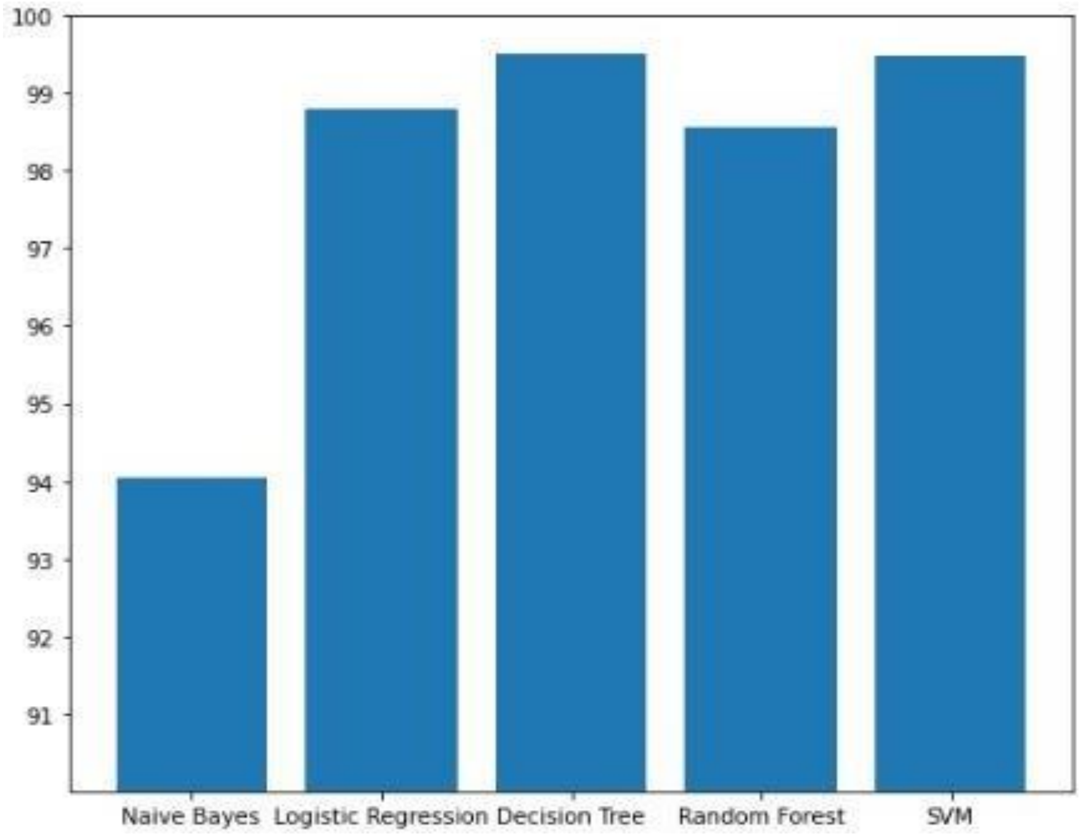


Fig 26. Graph showing accuracy of different classifier

Fig. 4 represents the graph that shows the accuracy of different classifier. Out of which the Decision Tre Classifier gives the maximum accuracy of “99.5%”. Logistic regression can output a probability of the input text being fake news, which allows for easy thresholding to make a binary decision. Naive Bayes also provides probability outputs, but it's not as reliable as logistic regression. Secondly Logistic regression is more robust to correlated features, which is important in text classification problems, where words are often correlated. Naive Bayes is based on the assumption of independence between features, which may not hold in text classification problems and in case of complex and highly non-linear data, logistic regression is more robust.

The obtained results can be discussed as Logistic regression has been one of the most common classification methods in machine learning due to its simplicity, interpretability, and high performance on certain types of datasets. In this study, the performance of different classifier was compared to determine which algorithm would provide the highest accuracy on a specific dataset.

The results shows that logistic regression outperformed other classifiers, such as naiveBayes, passive aggressive, and multilayer perceptron. The reason that Logistic regression out performed better than other classifier is Logistic regression is better suited to a specific characteristic of the chosen dataset which is linear separability of the data and due its relatively simple structure in this case.

The performance of machine learning algorithms can be affected by random variation, and the results of a single project do not necessarily generalize to other datasets or problems. To overcome this issue, researchers should consider cross-validation techniques, which help to assess the generalizability of a model by training and testing it on multiple datasets.

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

1.1 CONCLUSION

The approach for classifying false news is described in this essay, which also looks at the best qualities and methods for doing so. The authors research the topic of false news, current detection techniques, and its effects before proposing a novel model that makes use of feature extraction and three distinct AI techniques. Using a Logistic Regression classifier, the suggested model's accuracy score of 97.09% was the highest. The fact that Logistic regression can generate a chance that the input text is false news enables for simple thresholding to create a binary judgement, which is why it produced better results than other classifiers. However, logistic regression is more resistant to correlated features, which is crucial in text classification issues where words are often associated. Naive Bayes likewise produces probability outputs, but it is less trustworthy than logistic regression. Naive Bayes is built on the idea of feature independence, which could not hold true in text categorization issues. Additionally, Passive Aggressive Classifier is helpful when the data can be separated linearly, while logistic regression is more reliable when the data is complicated and very non-linear. When dealing with text classification issues, logistic regression also outperforms Naive Bayes and Passive Aggressive classifier in terms of accuracy and generalizability. Although each of the three classifiers has pros and limitations of their own, when it comes to fake news detection, Logistic Regression is seen to be the best option because of its simplicity, resilience, and capacity to output the likelihood that the input text is false.

1.2 FUTURE SCOPE

Fake news detection is a rapidly evolving field, and there are several areas where future work can be done to improve the performance and effectiveness of MLbased fake news detection projects. Some potential areas of future work in this field include:

1. Developing more sophisticated models:

There is always a scope of improving the model's performance by developing more advanced algorithms that can better distinguish between fakeand real news.

2. Incorporating multi-modal data:

Multi-modal data, which includes both text and visual information, is becoming increasingly popular on social media platforms. Future work can focus on developing models that can effectively incorporate both text and visual data to detectfake news.

3. Addressing bias:

Fake news detection models can be biased towards certain groups of people or certain types of news. Future work can focus on developing models that can address bias and ensure fairness in the fake news detection process.

4. Enhancing explainability:

Explainability is an essential aspect of ML-based systems, andfake news detection systems are no exception. Future work can focus on developing models that provide clear explanations for why certain news items are classified as fake.

5. Incorporating context:

Fake news detection models can be improved by considering the context in which the news was published. For example, understanding the political landscape, cultural differences, and geographical locations can help detect fake news better.

6. Developing better datasets:

The effectiveness of false news detection methods depends critically on the quality of the training data. Future research may concentrate on creating datasets that are bigger and more varied in order to enhance the precision and efficacy of false news detection methods.

7. Integrating with fact-checking tools:

Future work can focus on integrating fake news detection models with fact-checking tools to provide users with a comprehensive overview of the news item. This can help users make more informed decisions about the authenticity of news items they come across on social media platforms.

Overall, the future of fake news detection involves developing more sophisticated models that can address the challenges posed by multi-modal data, bias, and context while enhancing explain ability and integrating with fact-checking tools.

REFERENCES

1. A. Douglas, “News consumption and the new electronic media,” *The International Journal of Press/Politics*, vol. 11, no. 1, pp. 29–52, 2006.

2. J. Wong, “Almost all the traffic to fake news sites is from facebook, new data show,” 2016.

3. D. M. J. Lazer, M. A. Baum, Y. Benkler et al., “The science of fake news,” *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.

4. Can Machines Learn to Detect Fake News? A Survey Focused on Social Media Available at: https://scholarspace.manoa.hawaii.edu/handle/10125/59713.

5. Johan Hovold. ”Naive bayes spam filtering using word-position-based attributes.” In CEAS, 2005.R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

6. J. Hua and R. Shaw, “Corona virus (covid-19) “infodemic” and emerging issues through a data lens: the case of China,” *International Journal of Environmental Research and Public Health*, vol. 17, no. 7, p. 2309, 2020.

7. H. Jwa, D. Oh, K. Park, J. M. Kang, and H. Lim, “exBAKE: automatic fake news detection model based on bidirectional encoder representations from transformers (bert),” *Applied Sciences*, vol. 9, no. 19, 2019.

8. Bharath, G., Manikanta, K. J., Prakash, G. B., Sumathi, R., & Chinnasamy, P. (2021). *Detecting Fake News Using Machine Learning Algorithms. 2021 International Conference on Computer Communication and Informatics (ICCCI).*

9. Kumar, S., Kumar, S., Yadav, P., & Bagri, M. (2021). *A Survey on Analysis of Fake News Detection Techniques. 2021 International Conference on Artificial Intelligence and Smart Systems(ICAIS).*

10. D. S. K. R. Vivek Singh, Rupanjal Dasgupta and I. Ghosh, “Automated fake news detection using linguistic analysis and machine learning,” in International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction and Behavior Representation in Modeling and Simulation (SBPBRiMS), 2017

11. Baarir, N. F., & Djeffal, A. (2021). *Fake News detection Using Machine Learning. 2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-Being (IHSH)*

12. Parikh, S. B., & Atrey, P. K. (2018, April). Media-Rich Fake News Detection: A Survey. In 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR) (pp. 436-441). IEEE.

13. Conroy, N. J., Rubin, V. L., & Chen, Y. (2015, November). Automatic deception detection: Methods for finding fake news. In Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community (p. 82). American Society for Information Science.

14. Helmstetter, S., & Paulheim, H. (2018, August). Weakly supervised learning for fake news detection on Twitter. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 274-277). IEEE.

15. Wang, W. Y. (2017). " liar, liar pants on fire": A new benchmark dataset for fake news detection. arXiv preprint arXiv:1705.00648.

16. Stahl, K. (2018). Fake News Detection in Social Media.

17. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436.

18. Della Vedova, M. L., Tacchini, E., Moret, S., Ballarin, G., DiPierro, M., & de Alfaro, L. (2018, May). Automatic Online Fake News Detection Combining Content and Social Signals. In 2018 22nd Conference of Open Innovations Association (FRUCT) (pp. 272-279). IEEE.

19. Tacchini, E., Ballarin, G., Della Vedova, M. L., Moret, S., & de Alfaro, L. (2017). Some like it hoax: Automated fake news detection in social networks. arXiv preprint arXiv:1704.07506.

20. Shao, C., Ciampaglia, G. L., Varol, O., Flammini, A., & Menczer, F. (2017). The spread of fake news by social bots. arXiv preprint arXiv:1707.07592, 96-104.

21. Chen, Y., Conroy, N. J., & Rubin, V. L. (2015, November). Misleading online content: Recognizing clickbait as false news. In Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection (pp. 15-19). ACM.

22. Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. Journal of Big Data, 2(1), 1.

23. Haiden, L., & Althuis, J. (2018). The Definitional Challenges of Fake News.

24. Zhang, J., Cui, L., Fu, Y., & Gouza, F. B. (2018). Fake news detection with deep diffusive network model. arXiv preprint arXiv:1805.08751.

25. G. Hinton, S. Osindero, and Y. Teh. A fast learning algorithm for deep belief nets. Neural Comput., 2006.

26. R. salakhutdinov and G. Hinton. Semantic hashing. International Journal of Approximate Reasoning, 2009.

27. H. Jaeger. Tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the "echo state network" approach. Technical report, Fraunhofer Institute for Autonomous Intelligent Systems (AIS), 2002

RESEARCH PAPER

FAKE NEWS DETECTION USING ENSEMBLE LEARNING

Prof. Pushpendra Kumar[1], Mohd Naseem[2], Paras Jain[3], Prashansa Rai[4]

[1,2,3,4] Department of Computer Science and Engineering ,KIET Group Of Institutions,Ghaziabad,UP,India

Abstract

Over the recent years, the generation and exchange of news and information have rapidly risenwith the emergence of technology. The major sources of news and information have been social media feeds, news websites, blogs, and articles. A significant part of this news is reportedly fake and unauthentic, which can be misleading and spread fake rumours in public. This fake news is intentionally propagated to damage the reputation of an individual or an organization. This paper tried to solve this issue by applying machine learning to detect and effectively dismantle disinformation.

I . INTRODUCTION

The arrival of the world wide web and the vast emergence of social media platforms has increased data availability and provided real-time news to people.

The news evolution from newspapers to the internet has been rapidly speedy in recent years [1]. However, sometimes disinformation is propagated with a hidden agenda issued by the political parties and wealthy and influential people for their harmful intention causing riots and communal disputes between people.

Consumers always have the latest news at their fingertips. A large percentage of news site traffic is recommended by Facebook [2]. These social media provide a platform for the user where they can share their ideas and do a brief discussion about various topics, such as healthcare, democracy, and education but some people use these opportunities in a negative perspective to spread hate in our society [3].

Various types of data available on the internet, like text, multimedia and hyperlinks, serve in different ways. These data might necessarily not be authentic as they are collected various unknown resources.

Our paper proposes a novel method and tool for detecting fake news that utilizes machine learning. We have used ensemble learning models, including multiple commonly used classifiers, and have shown promising results in the project.

A total of five classification models have been utilized to detect fake news: Naive Bayes [4], Logistic Regression, Random Forest, Support Vector Machine and Decision Tree. The primary aim of this research paper is to show the benefits of machine learning through different classifiers. These models have shown great accuracy (above 90%) in the results. classifiers to achieve the best accuracy and precision [5].

Manually identifying false news may be difficult and often needs a high degree of subject-matter knowledge. It has been simpler to create and spread false information as a result of recent advances in computer science and technology, but it is much more difficult to tell the difference between fact and fiction.

Fake news spread and virality may have detrimental effects on people political careers as well as on goods and companies.

Back in February 2020, the WHO reported that the coronavirus pandemic had caused a "infodemic" that was made up of a lot of information, some of which was accurate and some of which was false. People now find it exceedingly difficult to determine whether sources of data or information are true and trustworthy. Too much misleading information may lead to the widespread spread of prejudice, fear, and other negative emotions.

This research proposes a novel method and tool for detecting fake news that makes use of artificial intelligence. The author contrasts the Naive Bayes Classifier, Logistic Regression, and Passive Aggressive Classifier, three supervised classification methods. The dataset, which contains both legitimate and false news, produces encouraging finding

II. RELATED WORK

In [10], the authors have proposed machine learning methods based on text processing. They have used Linguistic Inquiry and Word Count package to acquire features on provided texts on the dataset taken from Kaggle and from various prominent news agencies. Multiple Models have been used among which support vector machine gives the best result with 87% of accuracy.

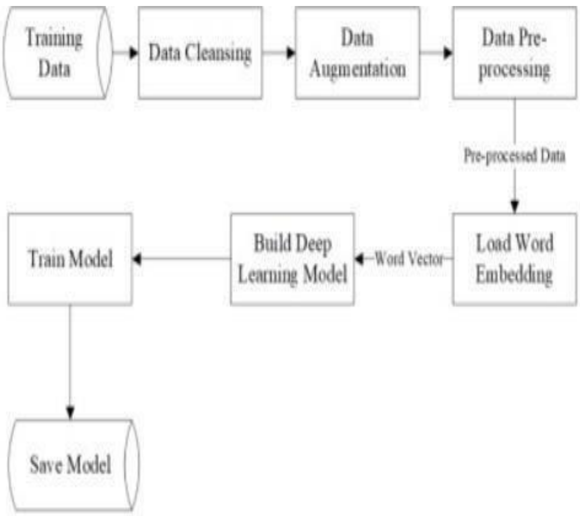
The authors of [11] have done a survey about all kinds of fake news that propagate through social media and how they impact social media user in different ways. Secondly the authors have explored various popular methods such as Rumour Classification, Truth Discovery, Click Bait Detection and Spammer Classification.

In [7], the authors have proposed an enhanced exBAKE model which is based on BERT model further which uses pre-training model. In this research paper the authors have given a solution for identifying fake news by comparing the heading of news and its corresponding content.

In [6], the research paper is based on the pandemic which occurs at the end of 2019 called COVID-19 which creates the emergency in the whole country. Many measures were taken to control this disease, but it spread like a cough and cold and stops the normal working of the country. After 6 months the vaccines like COVID SHIELD, COVAXIN comes in the market which are in two doses given to the people after a certain period of time.

III. METHODOLOGY

The fake news model detection is built using steps like Text Collection, Text Pre-processing, Feature Extraction and then classification using different classifiers. The experiments have been conducted on five different The definition of fake news is information that pushes people down the wrong road. Fake news spreads like wildfire, and people share it without confirming it. This is frequently done to promote or import specific views and is frequently accomplished through political agendas. Fake information is purposely or unintentionally spread throughout the internet. The massive dissemination of fake news has left an indelible mark on people and culture.



IV. DATASET USED

• We have downloaded two datasets named Fake, which consists of a list of articles considered fake news, and True, which consists of a list of articles considered real news. These news articles have been taken from Kaggle and contain news reports on diverse subjects. Further, we have used the data to train our machine-learning model. After that, we have done data pre-processing which is used before applying a machine learning algorithm to the model to make the raw data in a format suitable for input data for training. It is composed of data cleaning, transformation and reduction.

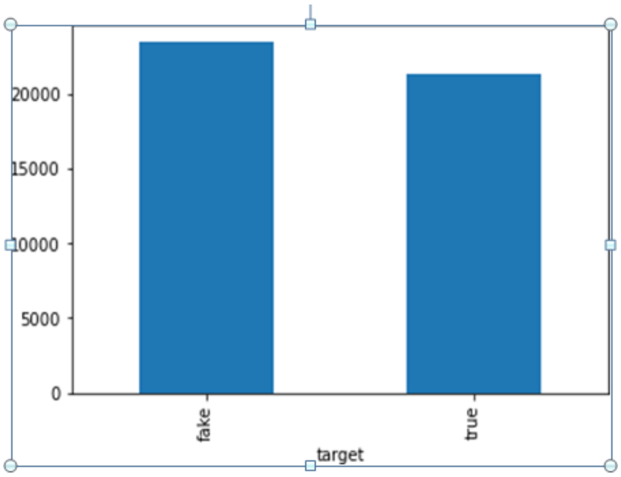


Fig.2 Distribution of Fake and Real News

- In Data Pre-processing, firstly, we cleaned the data by removing all irrelevant and null values. We added labels as fake and true for the training of data. Then we removed punctuation from these news reports. We have done tokenization which means breaking down sentences into words. We tokenized the data and removed unnecessary, most frequently occurring words. Lastly, we removed stop words from filtering out unnecessary data. Stop words are those words which are abundantly found in the text and are of no use. For the above processes, we imported Python libraries such as NumPy, Pandas and Scikit-Learn.

- After that, we split the dataset into training and testing data. Here we divide the dataset as 80% for training and the rest 20% to test the model.
- Then Different classification models were trained to find the best accuracy in detecting fake news from these articles.

V. Classification Models

- **Logistic Regression**

Logistic Regression is one of the supervised learning algorithms, i.e., a teacher or trainer is present to train the model.

In Logistic Regression, we classify the data according to our requirement and present the output in binary values such as true or false, which is transformed using a sigmoid function.

There are also various types of logistic regression functions such as Binomial, Multinomial and ordinal.

- **Decision Tree**

Decision tree has a tree like structure where every internal node represents a decision to be taken by this algorithm, each branch determines the result of that decision and the final classes are represented by leaf nodes. The path from root to leaf shows classification rule. It is used in ensemble learning and data mining for both regression and classification.

- **Random Forest**

It is a kind of ensemble learning algorithm that combines multiple decision trees. The idea behind using of random forest is to take combination of multiple decision trees, which minimize the chances of overfitting which is commonly found in decision trees. The Random forest algorithm makes multiple decision trees using random subset of training data and take the mean or majority vote of the prediction made by each individual trees for making last prediction. Random forest is usually used in regression and classification jobs. The Random forest is considered for producing highly accurate and robust results.

- **Naïve Bayes Classifier**

Naïve Bayes Classifier is a collection of classification algorithms based on Bayes Theorem which is used to determine the probability of an event with prior knowledge. The popular version of the algorithm is Gaussian Naïve Bayes used in classification task. It makes the prediction by scheming he expectations of the input data which is belonging to every class and then taking the class with maximum probability.in spite of this premise, the Naïve Bayes still perform good in exercising and used for spam filtering and classification.

- **Support Vector Machine (SVM)**

A support vector machine (SVM) is one of the popular machine learning algorithms used for classification and regression tasks. This algorithm finds the best decision boundary that separates or classify the data into different classes. It is chosen so that the distance between this hyperplane and the data points which are closest from this hyperplane is maximum.

Once the boundary is determined, new data can be classified by which side of the boundary it falls on. The applications of SVM have wide range such as image and text classification, bioinformatics, and natural language processing[8].

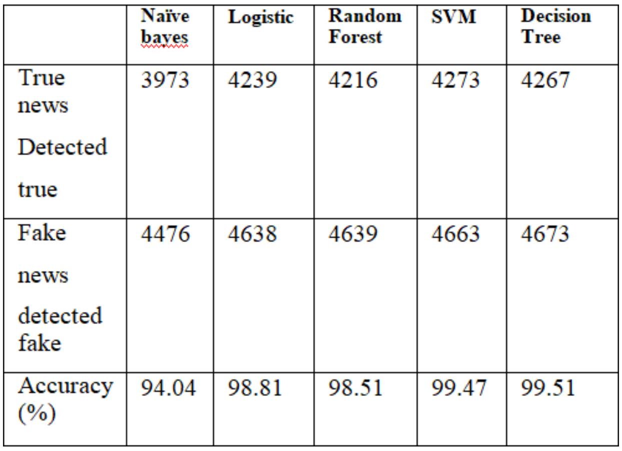
- **Ensemble Learners**

Ensemble learners are machine learning models that combine the predictions of multiple smaller models in order to improve the overall performance. There are variety of ways to do this, such as averaging the predictions, weighting the predictions based on the performance of each individual model, or training a meta-model to make a final prediction based on the outputs of the smaller models. Ensemble methods are often used in practice because they can lead to significant improvements in performance over a single model, especially in cases where the individual models have high bias or high variance[9].

VI. RESULTS

Initially we tested our dataset on Naïve Bayes classifier which gave an accuracy of 94.04%. The Logistic Regression Model performed well by giving an accuracy of 98.81%.

Random Forest Classifier performed closer to logistic regression model with an accuracy result of 98.51%. SVM and Decision Tree classifiers resulted in high accuracy with scores of 99.47% and 99.51%. These two models performed best among all and gave best results in accuracy.



	Naïve bayes	Logistic	Random Forest	SVM	Decision Tree
True news Detected true	3973	4239	4216	4273	4267
Fake news detected fake	4476	4638	4639	4663	4673
Accuracy (%)	94.04	98.81	98.51	99.47	99.51

VII. CONCLUSION

In our research paper, we tried to solve the fake news problem using machine learning with ensemble-based approach to experiment with various algorithms instead of one so that we can achieve best accuracy. The dataset we used were news articles which belonged to a variety of subjects. Data pre-processing removed a significant number of characters and words from these articles resulting in filtered useful data which trained our machine learning models

The overall average accuracy performs comes out as 98% which is a good score. Among all the algorithms Decision Tree and SVM gave highest accuracies with more than 99% followed by Random Forest and Logistic Regression algorithms scoring more than 98% accuracy and the least performance was given by Naïve Bayes classifier with 94%. By this experiment it is clear that using machine learning as a solution to recognize fake articles is a success.

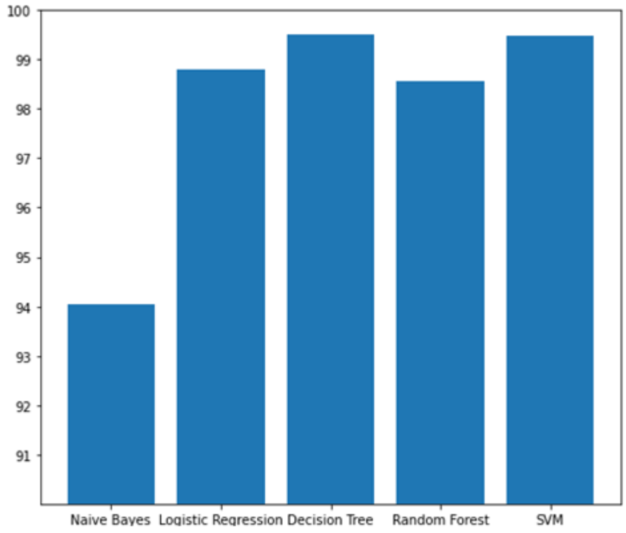


Fig.3 Accuracy Results of all the Algorithms

REFRENCES

[1] A. Douglas, “News consumption and the new electronic media,” *The International Journal of Press/Politics*, vol. 11, no. 1, pp. 29–52,2006.

[2] J. Wong, “Almost all the traffic to fake news sites is from facebook, new data show,” 2016.

[3] D. M. J. Lazer, M. A. Baum, Y. Benkler et al., “The science of fake news,” *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.

[4] Can Machines Learn to Detect Fake News? A Survey Focused onSocial Media Available at: https://scholarspace.manoa.hawaii.edu/handle/10125/59713.

[5] Johan Hovold. ”Naive bayes spam filtering using word-position-based attributes.” In CEAS, 2005.R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[6] J. Hua and R. Shaw, “Corona virus (covid-19) “infodemic” and emerging issues through a data lens: the case of China,” *International Journal of Environmental Research and Public Health*, vol. 17, no. 7, p. 2309, 2020.

[7] H. Jwa, D. Oh, K. Park, J. M. Kang, and H. Lim, “exBAKE: automatic fake news detection model based on bidirectional encoder representations from transformers (bert),” *Applied Sciences*, vol. 9, no. 19, 2019.

[8] Bharath, G., Manikanta, K. J., Prakash, G. B., Sumathi, R., & Chinnasamy, P. (2021). *Detecting Fake News Using Machine Learning Algorithms. 2021 International Conference on Computer Communication and Informatics (ICCCI).*

[9] Kumar, S., Kumar, S., Yadav, P., & Bagri, M. (2021). *A Survey on Analysis of Fake News Detection Techniques. 2021 International Conference on Artificial Intelligence and Smart Systems(ICAIS).*

[10] D. S. K. R. Vivek Singh, Rupanjal Dasgupta and I. Ghosh, “Automated fake news detection using linguistic analysis and machine learning,” in International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction and Behavior Representation in Modeling and Simulation (SBPBRiMS), 2017

[11] Baarir, N. F., & Djeffal, A. (2021). *Fake News detection Using Machine Learning. 2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-Being(IHSH).*

● 19% Overall Similarity

Top sources found in the following databases:

- 10% Internet database
- 6% Publications database
- Crossref database
- Crossref Posted Content database
- 16% Submitted Works database

TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	ABES Engineering College on 2019-04-22 Submitted works	2%
2	Middle East College of Information Technology on 2023-03-22 Submitted works	1%
3	ijraset.com Internet	<1%
4	github.com Internet	<1%
5	idoc.pub Internet	<1%
6	University of Bradford on 2023-03-29 Submitted works	<1%
7	researchgate.net Internet	<1%
8	Nottingham Trent University on 2022-07-01 Submitted works	<1%

9	University of Bradford on 2023-03-28	Submitted works	<1%
10	ABES Engineering College on 2020-05-20	Submitted works	<1%
11	daytrading.com	Internet	<1%
12	Liverpool John Moores University on 2022-11-22	Submitted works	<1%
13	Jasmine Shaikh, Rupali Patil. "Fake News Detection using Machine Lea...	Crossref	<1%
14	Meerut Institute of Engineering & Technology on 2022-05-31	Submitted works	<1%
15	core.ac.uk	Internet	<1%
16	Liverpool John Moores University on 2023-03-05	Submitted works	<1%
17	Liverpool John Moores University on 2022-11-22	Submitted works	<1%
18	University of Surrey on 2022-09-19	Submitted works	<1%
19	Texas A&M University, San Antonio on 2022-05-10	Submitted works	<1%
20	Liverpool John Moores University on 2023-02-28	Submitted works	<1%

21	University of Hertfordshire on 2022-04-29	Submitted works	<1%
22	University of Northumbria at Newcastle on 2022-01-22	Submitted works	<1%
23	Universiteit van Amsterdam on 2022-10-19	Submitted works	<1%
24	backend.orbit.dtu.dk	Internet	<1%
25	University of Technology, Sydney on 2022-10-28	Submitted works	<1%
26	eurchembull.com	Internet	<1%
27	University of East London on 2020-09-08	Submitted works	<1%
28	University of Hertfordshire on 2023-05-01	Submitted works	<1%
29	journalofcommunication.ro	Internet	<1%
30	"Advances in Computing and Data Sciences", Springer Science and Bus...	Crossref	<1%
31	Geoffrey E. Hinton, Simon Osindero, Yee-Whye Teh. "A Fast Learning A...	Crossref	<1%
32	Liberty University on 2023-04-17	Submitted works	<1%

33	Md. Nuruddin Qaisar Bhuiyan, Shantanu Kumar Rahut, Razwan Ahmed ... Crossref	<1%
34	Khyati Kapadiya, Usha Patel, Rajesh Gupta, Mohammad Dahman Alshe... Crossref	<1%
35	Liverpool John Moores University on 2023-02-13 Submitted works	<1%
36	Coventry University on 2023-04-03 Submitted works	<1%
37	cl.naist.jp Internet	<1%
38	export.arxiv.org Internet	<1%
39	ieeexplore.ieee.org Internet	<1%
40	yzf.co-aol.com Internet	<1%
41	"Recent Innovations in Computing", Springer Science and Business Me... Crossref	<1%
42	Ebtihal A. Hassan, Farid Meziane. "A Survey on Automatic Fake News I... Crossref	<1%
43	Nihel Fatima Baarir, Abdelhamid Djeffal. "Fake News detection Using ... Crossref	<1%
44	University of Hertfordshire on 2021-12-03 Submitted works	<1%