

**Data Science Project Training Report**  
**on**  
**Fake News Detection using Machine Learning Techniques**

**BACHELOR OF TECHNOLOGY**

**Session 2021-22**  
**in**

**Computer Science**

**By -**

- 1) RAJ CHAUHAN (2000320120132)**
- 2) ZUBAIN AHMAD (200320120201)**
- 3) TANYA MEHTA (2000320120182)**

**MS SAPNA JAIN**  
**ASSISTANT PROFESSOR**

**DEPARTMENT OF COMPUTER SCIENCE**  
**ABES ENGINEERING COLLEGE, GHAZIABAD**



**AFFILIATED TO**  
**DR. A.P.J. ABDUL KALAM TECHNICAL UNIVERSITY, U.P., LUCKNOW**  
**(Formerly UPTU)**

## Student's Declaration

I / We hereby declare that the work being presented in this report entitled “**FAKE NEWS DETECTION USING MACHINE LEARNING**” is an authentic record of my / our own work carried out under the supervision of Ms. **SAPNA JAIN, Assistant Professor, Computer Science**

Date:

**Signature of student  
student**

**RAJ CHAUHAN  
(2000320120132 )  
Department: CS**

**Signature of student**

**TANYA MEHTA  
(2000320120182)  
Department: CS**

**Signature of**

**ZUBAIN AHMAD  
(2000320120201)  
Department: CS**

This is to certify that the above statement made by the candidate(s) is correct to the best of my knowledge.

**Signature of HOD**

**Prof.(Dr.) Pankaj Kumar Sharma  
Computer Science**

**Signature of Teacher**

**Sapna Jain  
Assistant Professor  
Computer Science**

Date:.....

## **TABLE OF CONTENTS-**

<b>S.NO</b>	<b>CONTENTS</b>	<b>PAGE NO.</b>
	<b>Students declaration</b>	
<b>1.</b>	<b>INTRODUCTION</b>	
1.1	Problem statement	
1.2	Objective	
1.3	Methodology	
1.4	Literature Survey	
<b>2.</b>	<b>DATA EXPLORATION</b>	
2.1	Dataset	
2.2	Exploratory Data Analysis	
<b>3.</b>	<b>PREPROCESSING</b>	
3.1	Vectorization of data - using count vectorization.	
3.2	Train test splitting of data	
<b>4.</b>	<b>MODELING</b>	
4.1	Using Native Bayes Classifier	
4.2	Using Logistic Regression	
4.3	Using Random Forest	
<b>5.</b>	<b>IMPLEMENTATION ( CODE WITH SCREENSHOTS)</b>	
<b>6.</b>	<b>COMPARISON BETWEEN MODELS ( RESULTS)</b>	
6.1	Confusion Matrix	
6.2	Visualization of Result	
<b>7.</b>	<b>CONCLUSION</b>	
<b>8.</b>	<b>REFERENCES</b>	

# 1. INTRODUCTION

Data or information is the most valuable asset. The most important problem to be solved is to evaluate whether the data is relevant or irrelevant. Fake data has a huge impact on lot of people and organizations.

Since fake news tends to spread fast than the real news there a need to classify news as fake or not. In the project the dataset used is from Kaggle website where real news and fake news are in two separate datasets we combined both the datasets into one and trained with different machine learning classification algorithms to classify the news as fake or not.

In this project different feature engineering methods for text data has been used like Bag of words model and word embedding model which is going to convert the text data into feature vectors which is sent into machine learning algorithms to classify the news as fake or not.

With different features and classification algorithms we are going to classify the news as fake or real and the algorithm with the feature which gives us the best result with that feature extraction method and that algorithm we are going to predict the news as fake or real.

In this project we will be ignoring attributes like the source of the news, whether it was reported online or in print, etc. and instead focus only the content matter being reported. We aim to use different machine learning algorithms and determine the best way to classify news .

## 1.1 PROBLEM STATEMENT

Our main aim of the project is to make a machine learning model, with the help of which news can be classified as fake or real with help of different machine learning classification algorithms, deep learning methods and text feature extraction methods for classifying news.

## 1.2 OBJECTIVE

To achieve our goal of developing machine learning model to classify news as fake or real, we need perform following tasks in the same order as stated.

- Data Collection and Analysis
- Preprocessing the data
- Text feature extraction
- Using different classification algorithms
- Taking the best classification algorithm and feature extraction method.

-Classifying the news as fake or real.

### 1.3 METHODOLOGY

Story ID	Requirement description	User stories/Task	Description
Requirement	Gathering project Ideas To work on .	Find if there already exist Implementations Of the chosen project idea. If implementation exists, study the existing Implementation. Based on the study, arrive at the Missing necessary Features we can build.	Collect data and ideas from various sources to find a potential project.
Planning	Prepare feature list. Technologies to be used to estimate effort	Decide on the important Functionalities we can add to our implementation	Planning is process which embraces a no of steps to be taken.
Development	High level API class design. Cloud based machines, Google Collab.	Decide and arrive at an overview of modules and classes. Using Collab we made the code much more readable.	Coding basic python, Sci-Kit learning library was majorly used.
Test Cases	Accuracy of random forest, Logistic regression, and Naive Bayes are compared.	Implement different features for the UI and make sure they are working properly.	Information and research on the algorithms Used for classification of data

### 1.4 LITERATURE SURVEY

In Today's world, anybody can post the content over the internet. Unfortunately, counterfeit news gathers a lot of consideration over the web, particularly via web-based networking media. Individuals get misdirected and don't reconsider before flowing such miseducational pieces to the

most distant part of the arrangement. Such type of activities are not good for the society where some rumors or vague news evaporates the negative thought among the people or specific category of people. As fast the technology is moving, on the same pace the preventive measures are required to deal with such activities. Broad communications assuming a gigantic job in impacting the general public and as it is normal, a few people attempt to exploit it. There are numerous sites which give false data.

They deliberately attempt to bring out purposeful publicity, deceptions and falsehood under the pretense of being true news. Their basic role is to control the data that can cause open to have confidence in it. There are loads of case of such sites everywhere throughout the world .Therefore, counterfeit news influences the brains of the individuals. As indicated by study Scientist accept that numerous man-made brainpower calculations can help in uncovering the bogus news.

Fake news detection is made to stop the rumors that are being spread through the various platforms whether it be social media or messaging platforms, this is done to stop spreading fake news which leads to activities like mob lynching, this has been a great reason motivating us to work on this project. We have been continuously seeing various news of mob lynching that leads to the murder of an individual; fake news detection works on the objective of detecting this fake news and stopping activities like this thereby protecting the society from these unwanted acts of violence. The digital news industry in the United States is facing a complex future.

On one hand, a steadily growing portion of Americans are getting news through the internet, many U.S. adults get news on social media, and employment at digital-native outlets has increased. On the other, digital news has not been immune to issues affecting the broader media environment, including layoffs, made-up news and public distrust.

## 2. DATA EXPLORATION

A good starting point for the analysis is to make some data exploration of the data set. The first thing to be done is statistical analysis such as counting the number of texts per class or counting the number of words per sentence. Then it is possible to try to get an insight of the data distribution by making dimensionality reduction and plotting data in 2D.

### 2.2 DATASET

Online news can be collected from different sources, such as news agency homepages, search engines, and social media websites. However, manually determining the veracity of news is a challenging task, usually requiring annotators with domain expertise who performs careful analysis of claims and additional evidence, context, and reports from authoritative sources. Generally, news data with annotations can be gathered in the following ways: Expert journalists, Fact-checking websites, Industry detectors, and Crowdsourced workers. However, there are no agreed upon benchmark datasets for the fake news detection problem.

This dataset is collected from fact-checking website having address-

<https://www.kaggle.com/clmentbisailon/fake-and-real-news-dataset?select=True.csv>

We can get online news from different sources like social media websites, search engine, homepage of news agency websites or fact checking websites. On the Internet, there are a few publicly available datasets for Fake news classification like BuzzFeed News, LIAR, BS Detector etc. These datasets have been widely used in different research papers for determining the veracity of news.

Fake News Corpus:

This works uses multiple corpus in order to train and test different models. The main corpus used for training is called Fake News Corpus. This corpus has been automatically crawled using opensources.co labels. In other words, domains have been labeled with one or more labels in • Fake News • Satire • Extreme Bias • Conspiracy Theory • Junk Science • Hate News • Clickbait • Proceed With Caution • Political • Credible

Liar, Liar Pants on Fire:

The Liar, Liar Pants on Fire dataset, which is a collection of twelve thousand small sentences collected from various sources and hand labeled. They are divided in six classes:  
• pants-fire • false • barely-true • half-true • mostly-true • true

### 2.3 EXPLORATORY DATA ANALYSIS

The first step is to put everything in a database in order to be able to retrieve only the wanted piece of information. In order to do so, the file has been read line by line. It appears that some of the lines are badly formatted, preventing them from being read correctly, in this case they are dropped without being put in the database. Also, each line that is a duplicate of a line already read is also dropped.

We found the following values at the table for real data.

```
# Column Non-Null Count Dtype
---  ---  ---
0 title  21417 non-null object
1 text   21417 non-null object
2 subject 21417 non-null object
3 date   21417 non-null object
```

We found the following values at the table for fake data.

```
# Column Non-Null Count Dtype
---  ---  ---
0 title  23481 non-null object
1 text   23481 non-null object
2 subject 23481 non-null object
3 date   23481 non-null object
```

In addition, the number of words per text and the average number of words per sentences have been computed for each text categories using wordcloud after which both real and fake data sets are combined.

We found the following values at the table for the combined dataset.

```
# Column Non-Null Count Dtype
---  ---  ---
0 title  44898 non-null object
1 text   44898 non-null object
2 subject 44898 non-null object
3 date   44898 non-null object
4 Target 44898 non-null int64
```



### 3. PREPROCESSING

A. Pre-processing Data Social media data is highly unstructured – majority of them are informal communication with typos, slangs and bad-grammar etc. Quest for increased performance and reliability has made it imperative to develop techniques for utilization of resources to make informed decisions. To achieve better insights, it is necessary to clean the data before it can be used for predictive modeling. For this purpose, basic pre-processing was done on the News training data. This step consisted of Data Cleaning: While reading data, we get data in the structured or unstructured format. A structured format has a well-defined pattern whereas unstructured data has no proper structure. In between the 2 structures, we have a semi-structured format which is comparably better structured than unstructured format. Cleaning up the text data is necessary to highlight attributes that we're going to want our machine learning system to pick up on. Cleaning (or pre-processing) the data typically consists of a number of steps: a) Remove punctuation Punctuation can provide grammatical context to a sentence which supports our understanding. But for our vectorizer which counts the number of words and not the context, it does not add value, so we remove all special characters. eg: How are you?->How are you b) Tokenization Tokenizing separates text into units such as sentences or words. It gives structure to previously unstructured text. eg: Plata o Plomo-> 'Plata','o','Plomo'. c) Remove stopwords Stopwords are common words that will likely appear in any text. They don't tell us much about our data so we remove them. eg: silver or lead is fine for me-> silver, lead, fine. d) Stemming Stemming helps reduce a word to its stem form. It often makes sense to treat related words in the same way. It removes suffices, like "ing", "ly", "s", etc. by a simple rule-based approach. It reduces the corpus of words but often the actual words get neglected. eg: Entitling, Entitled -> Entitle. Note: Some search engines treat words with the same stem as synonyms.

#### 3.1 VECTORIZATION OF DATA

- The process of converting words or text into numbers or vectors are called Text-Vectorization.
  - can be done using :-
    - 1.CountVectorizer()
    - 2.TfidfVectorizer()

1. Vectorizing Data: Bag-Of-Words Bag of Words (BoW) or CountVectorizer describes the presence of words within the text data. It gives a result of 1 if present in the sentence and 0 if not present. It, therefore, creates a bag of words with a document-matrix count in each text document.

2. Vectorizing Data: TF-IDF It computes "relative frequency" that a word appears in a document compared to its frequency across all documents TF-IDF weight represents the relative importance of a term in the document and entire corpus. TF stands for Term Frequency: It calculates how frequently a term appears in a document. Since, every document size varies, a term may appear more in a long sized document than a short one.

Thus, the length of the document often divides Term frequency. Note: Used for search engine scoring, text summarization, document clustering.

$TF(t, d) = \text{Number of times } t \text{ occurs in document 'd'}$  Total word count of document 'd' IDF stands for Inverse Document Frequency: A word is not of much use if it is present in all the documents.

Certain terms like “a”, “an”, “the”, “on”, “of” etc. appear many times in a document but are of little importance. IDF weighs down the importance of these terms and increases the importance of rare ones. The more the value of IDF, the more unique the word is.  $IDF(t, d) = \frac{\text{Total number of documents}}{\text{Number of documents with term } t \text{ in it}}$  TF-IDF is applied on the body text, so the relative count of each word in the sentences is stored in the document matrix.

$$TFIDF(t, d) = TF(t, d) * IDF(t)$$

Note: Vectorizers output sparse matrices. Sparse Matrix is a matrix in which most entries are 0 [21]

### 3.2 TRAIN TEST SPLITTING OF DATA

- Splitting dataset into Training and Testing Datasets.
  - Training as 70%
  - Testing as 30%

## 4. MODELING

This section deals with training the classifier. Different classifiers were investigated to predict the class of the text. We explored specifically four different machine-learning algorithms – Multinomial Naïve Bayes Passive Aggressive Classifier and Logistic regression. The implementations of these classifiers were done using Python library Sci-Kit Learn.

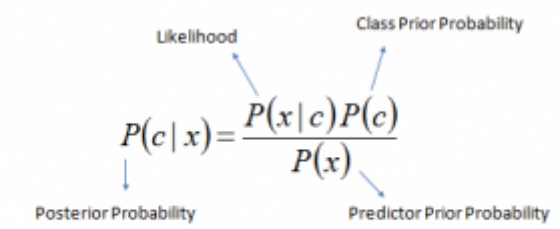
### 4.1 USING NATIVE BAYES CLASSIFIER

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

Bayes theorem provides a way of calculating posterior probability  $P(c|x)$  from  $P(c)$ ,  $P(x)$  and  $P(x|c)$ . Look at the equation below:



The diagram shows the equation  $P(c|x) = \frac{P(x|c)P(c)}{P(x)}$  with arrows pointing from labels to the terms: 'Likelihood' points to  $P(x|c)$ , 'Class Prior Probability' points to  $P(c)$ , 'Posterior Probability' points to  $P(c|x)$ , and 'Predictor Prior Probability' points to  $P(x)$ .

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$
$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$

Above,

- $P(c|x)$  is the posterior probability of *class* ( $c$ , *target*) given *predictor* ( $x$ , *attributes*).
- $P(c)$  is the prior probability of *class*.
- $P(x|c)$  is the likelihood which is the probability of *predictor* given *class*.
- $P(x)$  is the prior probability of *predictor*.

### 4.2 USING LOGISTIC REGRESSION

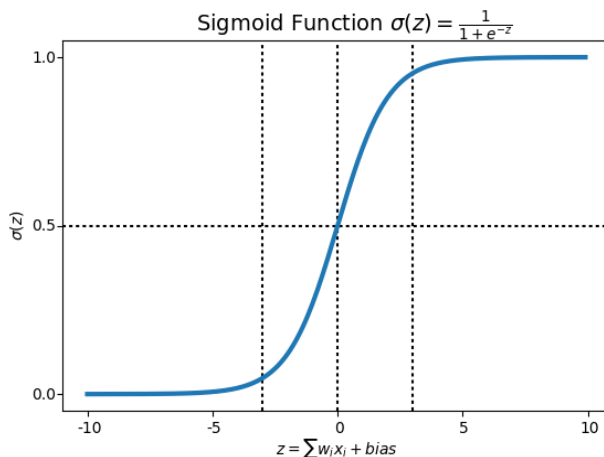
Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability.

We can call a Logistic Regression a Linear Regression model but the Logistic Regression uses a more complex cost function, this cost function can be defined as the '**Sigmoid function**' or also known as the 'logistic function' instead of a linear function. The hypothesis of logistic regression tends it to limit the cost function between 0 and 1. Therefore linear functions fail to represent it as

it can have a value greater than 1 or less than 0 which is not possible as per the hypothesis of logistic regression.

$$0 \leq h_{\theta}(x) \leq 1$$

In order to map predicted values to probabilities, we use the Sigmoid function. The function maps any real value into another value between 0 and 1. In machine learning, we use sigmoid to map predictions to probabilities.



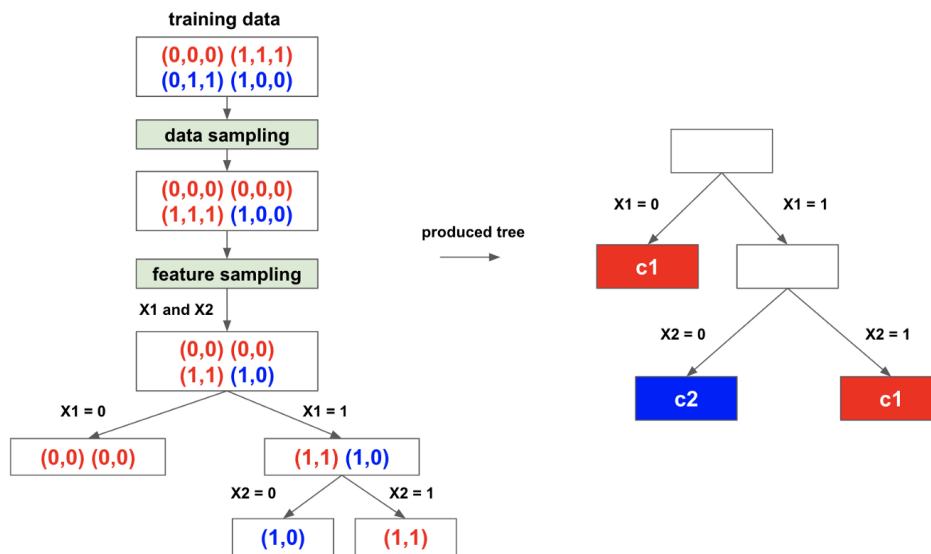
$$f(x) = \frac{1}{1+e^{-(x)}}$$

### 4.3 USING RANDOM FOREST

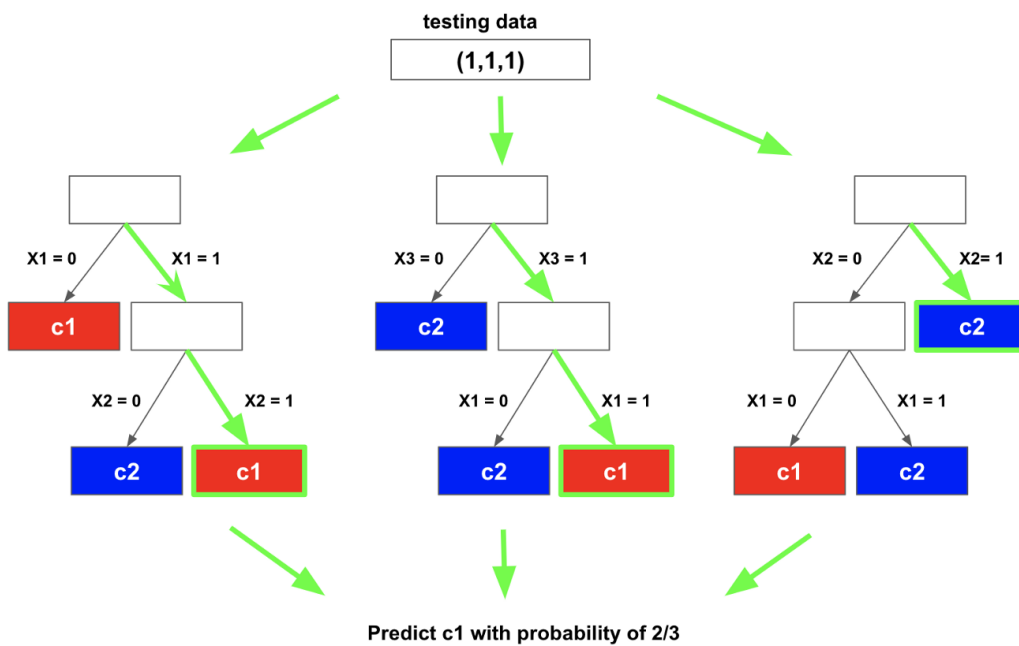
A random forest consists of multiple random decision trees. Two types of randomnesses are built into the trees. First, each tree is built on a random sample from the original data. Second, at each tree node, a subset of features are randomly selected to generate the best split. We use the dataset below to illustrate how to build a random forest tree. Note Class = XOR(X1,X2). X3 is made identical as X2 (for illustrative purposes in later sections).

X1	X2	X3	Class
0	0	0	c1
1	1	1	c1
0	1	1	c2
1	0	0	c2

The figure below demonstrates how to build a random forest tree.



The same process is applied to build multiple trees. The figure below illustrates the flow of applying a random forest with three trees to a testing data instance.



## 5. IMPLEMENTATION( CODE AND SCREENSHOTS)

### 1.Loading Libraries and datasets

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from wordcloud import WordCloud
```

```
from google.colab import drive
drive.mount('/content/drive')
```

Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force\_remount=True).

```
real_df=pd.read_csv("/content/drive/MyDrive/Fake News.zip (Unzipped
Files)/Fake News/True.csv")
fake_df=pd.read_csv("/content/drive/MyDrive/Fake News.zip (Unzipped
Files)/Fake News/Fake.csv")
```

### 2. Exploratory Data Analysis

```
real_df.head()
```

```
                                title \
0  As U.S. budget fight looms, Republicans flip t...
1  U.S. military to accept transgender recruits o...
2  Senior U.S. Republican senator: 'Let Mr. Muell...
3  FBI Russia probe helped by Australian diplomat...
4  Trump wants Postal Service to charge 'much mor...

                                text      subject \
0  WASHINGTON (Reuters) - The head of a conservat...  politicsNews
1  WASHINGTON (Reuters) - Transgender people will...  politicsNews
2  WASHINGTON (Reuters) - The special counsel inv...  politicsNews
3  WASHINGTON (Reuters) - Trump campaign adviser ...  politicsNews
4  SEATTLE/WASHINGTON (Reuters) - President Donal...  politicsNews

                                date
0  December 31, 2017
1  December 29, 2017
2  December 31, 2017
3  December 30, 2017
4  December 29, 2017
```

```
fake_df.head()
```

```
                                title \
0  Donald Trump Sends Out Embarrassing New Year'...
1  Drunk Bragging Trump Staffer Started Russian ...
2  Sheriff David Clarke Becomes An Internet Joke...
3  Trump Is So Obsessed He Even Has Obama's Name...
4  Pope Francis Just Called Out Donald Trump Dur...

                                text subject \
```

0	Donald Trump just couldn t wish all Americans ...	News
1	House Intelligence Committee Chairman Devin Nu...	News
2	On Friday, it was revealed that former Milwauk...	News
3	On Christmas day, Donald Trump announced that ...	News
4	Pope Francis used his annual Christmas Day mes...	News

	date
0	December 31, 2017
1	December 31, 2017
2	December 30, 2017
3	December 29, 2017
4	December 25, 2017

```
real_df.columns,fake_df.columns
```

```
(Index(['title', 'text', 'subject', 'date'], dtype='object'),
Index(['title', 'text', 'subject', 'date'], dtype='object'))
```

```
real_df.shape,fake_df.shape
```

```
((21417, 4), (23481, 4))
```

```
real_df.info()
print('\n')
fake_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21417 entries, 0 to 21416
Data columns (total 4 columns):
#   Column      Non-Null Count  Dtype
---  ---
0    title      21417 non-null  object
1    text       21417 non-null  object
2    subject    21417 non-null  object
3    date       21417 non-null  object
dtypes: object(4)
memory usage: 669.4+ KB
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23481 entries, 0 to 23480
Data columns (total 4 columns):
#   Column      Non-Null Count  Dtype
---  ---
0    title      23481 non-null  object
1    text       23481 non-null  object
2    subject    23481 non-null  object
3    date       23481 non-null  object
dtypes: object(4)
memory usage: 733.9+ KB
```

```
real_df.describe()
```

count	title \
	21417

```
unique                20826
top    Factbox: Trump fills top jobs for his administ...
freq                14
```

```
count                text                subject \
unique                21417                21417
top    (Reuters) - Highlights for U.S. President Dona... politicsNews
freq                8                11272
```

```
count                date
unique                21417
top    December 20, 2017
freq                716
```

```
fake_df.describe()
```

```
count                title    text    subject
unique                23481    23481    23481
top    MEDIA IGNORES Time That Bill Clinton FIRED His...
freq                6        626    9050
```

```
count                date
unique                23481
top    May 10, 2017
freq                1681
```

```
real_df['subject'].unique()
```

```
array(['politicsNews', 'worldnews'], dtype=object)
```

```
fake_df['subject'].unique()
```

```
array(['News', 'politics', 'Government News', 'left-news', 'US_News',
      'Middle-east'], dtype=object)
```

### Visualization of most frequent words

```
real_words = " ".join([x for x in real_df['title']])
wordcloud1 = WordCloud(width=500, height=500,
random_state=40).generate(real_words)
```

```
plt.figure(figsize=(20,8))
plt.imshow(wordcloud1,interpolation='bilinear')
plt.axis('off')
plt.show()
```







	date	Target
0	December 31, 2017	1
1	December 29, 2017	1
2	December 31, 2017	1
3	December 30, 2017	1
4	December 29, 2017	1

df.tail()

	title \
23476	McPain: John McCain Furious That Iran Treated ...
23477	JUSTICE? Yahoo Settles E-mail Privacy Class-ac...
23478	Sunnistan: US and Allied 'Safe Zone' Plan to T...
23479	How to Blow \$700 Million: Al Jazeera America F...
23480	10 U.S. Navy Sailors Held by Iranian Military ...

	text	subject \
23476	21st Century Wire says As 21WIRE reported earl...	Middle-east
23477	21st Century Wire says It s a familiar theme. ...	Middle-east
23478	Patrick Henningsen 21st Century WireRemember ...	Middle-east
23479	21st Century Wire says Al Jazeera America will...	Middle-east
23480	21st Century Wire says As 21WIRE predicted in ...	Middle-east

	date	Target
23476	January 16, 2016	0
23477	January 16, 2016	0
23478	January 15, 2016	0
23479	January 14, 2016	0
23480	January 12, 2016	0

df.info()

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 44898 entries, 0 to 23480
Data columns (total 5 columns):
#   Column      Non-Null Count  Dtype
---  -
0    title      44898 non-null  object
1    text       44898 non-null  object
2    subject    44898 non-null  object
3    date       44898 non-null  object
4    Target     44898 non-null  int64
dtypes: int64(1), object(4)
memory usage: 2.1+ MB
```

df.shape

(44898, 5)

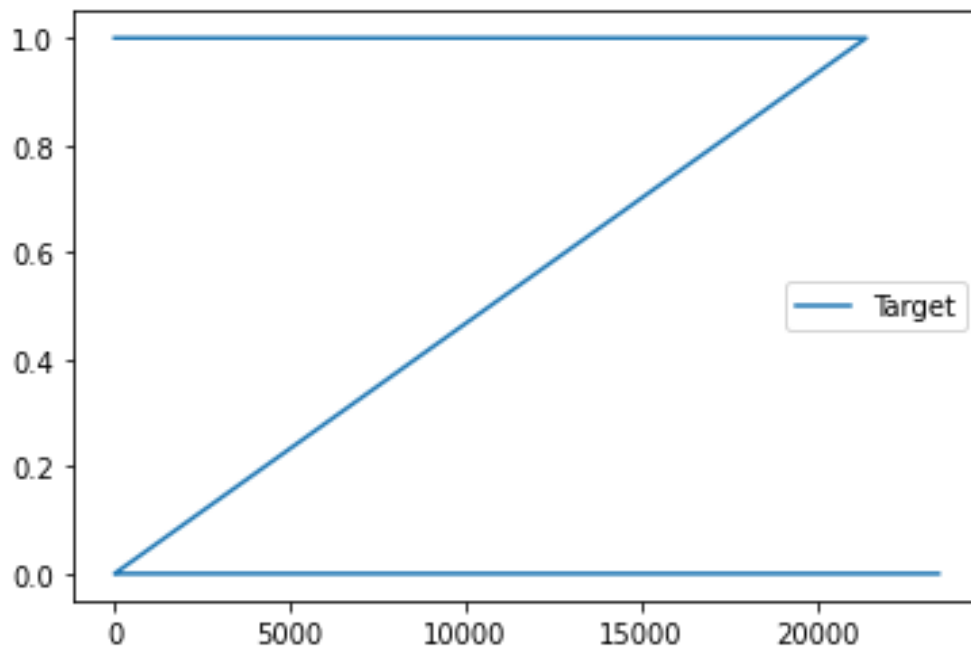
df.describe()

	Target
count	44898.000000
mean	0.477015

```
std      0.499477
min      0.000000
25%      0.000000
50%      0.000000
75%      1.000000
max      1.000000
```

```
df.plot()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fcf9c318810>
```



```
df['Text_len']=df['text'].apply(len)
df['Title_len']=df['title'].apply(len)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 44898 entries, 0 to 23480
Data columns (total 7 columns):
#   Column      Non-Null Count  Dtype
---  -
0   title       44898 non-null  object
1   text        44898 non-null  object
2   subject     44898 non-null  object
3   date        44898 non-null  object
4   Target      44898 non-null  int64
5   Text_len    44898 non-null  int64
6   Title_len   44898 non-null  int64
dtypes: int64(3), object(4)
memory usage: 2.7+ MB
```

```
df.groupby('Target').describe()
```

Text_len	count	mean	std	min	25%	50%	75%
----------	-------	------	-----	-----	-----	-----	-----

\

Target							
0	23481.0	2547.396235	2532.884399	1.0	1433.0	2166.0	3032.0
1	21417.0	2383.278517	1684.835730	1.0	914.0	2222.0	3237.0

	Title_len							
\	max	count	mean	std	min	25%	50%	75%
Target								
0	51794.0	23481.0	94.198032	27.184433	8.0	77.0	90.0	105.0
1	29781.0	21417.0	64.667881	9.168999	26.0	59.0	64.0	70.0

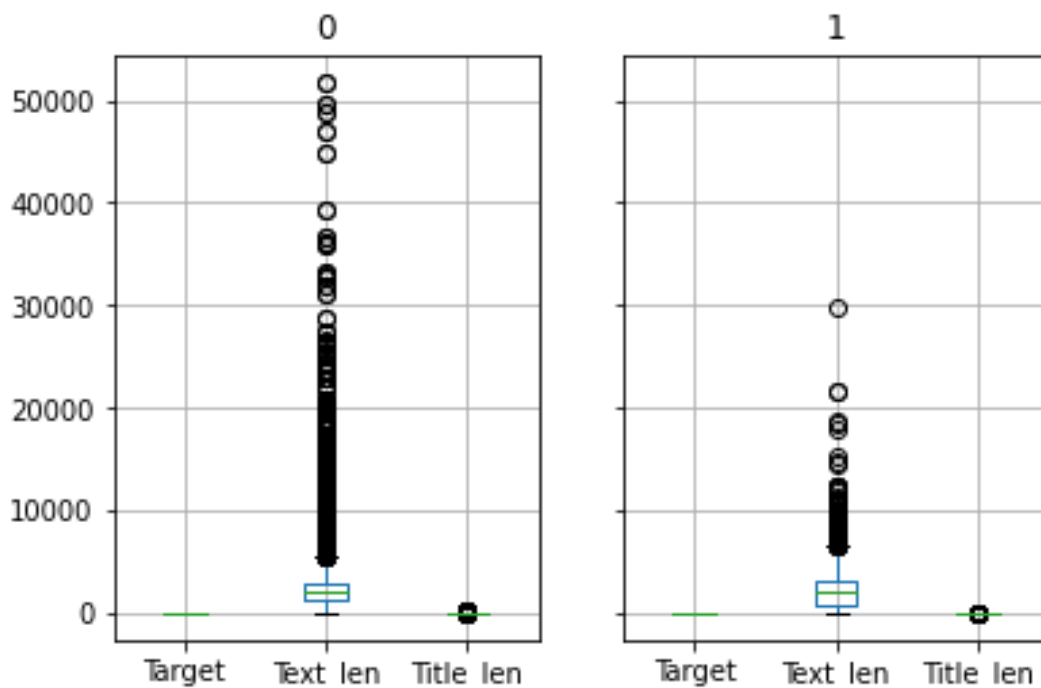
	max
Target	
0	286.0
1	133.0

```
df.groupby('Target').median()
```

	Text_len	Title_len
Target		
0	2166.0	90.0
1	2222.0	64.0

```
df.groupby('Target').boxplot()
```

```
0 AxesSubplot(0.1,0.15;0.363636x0.75)
1 AxesSubplot(0.536364,0.15;0.363636x0.75)
dtype: object
```



### Insights

- Target=0 is fake data Target=1 is real data.

- Average length of Titles of real data is 64.66 and fake data is 94.19 .
- Length of title fake data is more than that of real data.

## 2.Data Preprocessing

- Delete attributes which are of no use.
  - e.g. "Date" and "Subject" in this case.

```
df.drop(columns=['date', 'subject'])
```

```

                                title \
0      As U.S. budget fight looms, Republicans flip t...
1      U.S. military to accept transgender recruits o...
2      Senior U.S. Republican senator: 'Let Mr. Muell...
3      FBI Russia probe helped by Australian diplomat...
4      Trump wants Postal Service to charge 'much mor...
...
23476  McPain: John McCain Furious That Iran Treated ...
23477  JUSTICE? Yahoo Settles E-mail Privacy Class-ac...
23478  Sunnistan: US and Allied 'Safe Zone' Plan to T...
23479  How to Blow $700 Million: Al Jazeera America F...
23480  10 U.S. Navy Sailors Held by Iranian Military ...

                                text  Target  Text_len
\
0      WASHINGTON (Reuters) - The head of a conservat...      1      4659
1      WASHINGTON (Reuters) - Transgender people will...      1      4077
2      WASHINGTON (Reuters) - The special counsel inv...      1      2789
3      WASHINGTON (Reuters) - Trump campaign adviser ...      1      2461
4      SEATTLE/WASHINGTON (Reuters) - President Donal...      1      5204
...
23476  21st Century Wire says As 21WIRE reported earl...      0      3237
23477  21st Century Wire says It s a familiar theme. ...      0      1684
23478  Patrick Henningsen 21st Century WireRemember ...      0     25065
23479  21st Century Wire says Al Jazeera America will...      0      2685
23480  21st Century Wire says As 21WIRE predicted in ...      0      5251

                                Title_len
0                                64
1                                64
2                                60
3                                59
4                                69
...
23476                            61
23477                            81
23478                            85
23479                            67
23480                            81

```

```
[44898 rows x 5 columns]
```

### 3. Text-Vectorization and Train-Test split of data

#### Vectorization

- The process of converting words or text into numbers or vectors are called Text-Vectorization.
  - can be done using :-
    - 1.CountVectorizer()
    - 2.TfidfVectorizer()

```
from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer()
```

```
X = df['text']
Y = df['Target']
```

```
X = cv.fit_transform(X)
```

#### Train-Test split of data

- Splitting dataset into Training and Testing Datasets.
  - Training as 70%
  - Testing as 30%

```
from sklearn.model_selection import train_test_split
```

```
X_train,X_test,Y_train,Y_test = train_test_split(X,Y,test_size =
0.3,random_state = 101)
```

### 4.Modeling

#### a. Naive Bayes

```
from sklearn.naive_bayes import MultinomialNB
nb = MultinomialNB()
nb.fit(X_train, Y_train)
```

```
MultinomialNB()
```

```
from sklearn.metrics import classification_report, confusion_matrix
```

```
nb_prediction = nb.predict(X_test)
```

```
from sklearn.metrics import accuracy_score
accuracy_score(Y_test,nb_prediction)
```

```
0.9512249443207127
```

*~ Accuracy is 95% for Naive Bayes.*

#### b.Logistic Regression

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
model1=LogisticRegression(solver='lbfgs', max_iter=1000)
model1.fit(X_train,Y_train)
```

```
LogisticRegression(max_iter=1000)
prediction=model1.predict(X_test)
accuracy_score(Y_test,prediction)
0.9952487008166295
```

*~ Accuracy is 99.53% for Logistic Regression.*

### *c. Random Forest*

```
from sklearn.ensemble import RandomForestClassifier
ref=RandomForestClassifier(n_estimators=135,
    criterion='gini',
    max_depth=None,
    max_features='auto',
    max_leaf_nodes=None,
    bootstrap=True,
    n_jobs=None,
    random_state=25,
)
ref.fit(X_train,Y_train)
RandomForestClassifier(n_estimators=135, random_state=25)
ref.score(X_test,Y_test)
0.9874536005939124
```

*~ Accuracy is 98.74% for Random Forest.*



## 6. COMPARISON BETWEEN MODELS ( RESULTS)

### 6.1 CONFUSION MATRIX

After applying three different classifiers (Naïve bayes, Logistic Regression and Random Forest), their confusion matrix showing actual set and predicted sets are mentioned below:

Table 1 : Confusion Matrix for Naive Bayes Classifier

<b>Total = 13470</b>	<b>True (Predicted)</b>	<b>Fake (Predicted)</b>
<b>True (Actual)</b>	<b>6664</b>	<b>314</b>
<b>Fake (Actual)</b>	<b>343</b>	<b>6419</b>

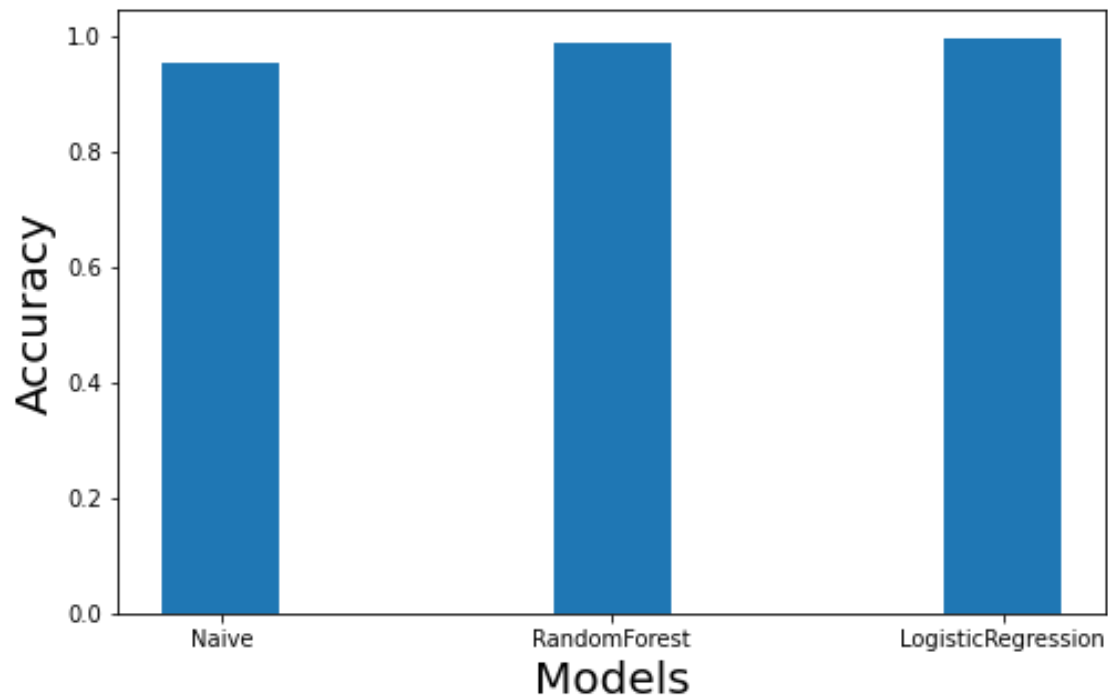
Table 2 : Confusion Matrix for Logistic Regression

<b>Total = 13470</b>	<b>True (Predicted)</b>	<b>Fake (Predicted)</b>
<b>True (Actual)</b>	<b>6946</b>	<b>32</b>
<b>Fake (Actual)</b>	<b>32</b>	<b>6460</b>

Table 3 : Confusion Matrix for Random Forest

<b>Total = 13470</b>	<b>True (Predicted)</b>	<b>Fake (Predicted)</b>
<b>True (Actual)</b>	<b>6901</b>	<b>77</b>
<b>Fake (Actual)</b>	<b>92</b>	<b>6400</b>

## 6.2 VISUALIZATION OF RESULT



### Result

- Logistic Regression gives best results for this.
  - ~ Accuracy=99.53%

## 7. CONCLUSION

In this project three different feature extraction methods like Count Vectorizer has been used. And also different classification algorithms like Logistic Regression Classifier, Naive Bayes Classifier, Random Forest Classifier have been used to classify the news as fake or real.

By using the classification algorithms we got highest accuracy with SVM Linear classification algorithm and with TF-IDF feature extraction with 0.94 accuracy. Even though we got the same accuracy with Neural Network with Count Vectorizer, Neural Networks take more time to train and are complex, so we used Linear SVC which is not so complex and takes less time to compute.

## 8. REFERENCES

- [1] Fake news detection Akshay Jain, Amey Kasbe 2018 IEEE International Students Conference on Electrical, Electronics and Computer Sciences.
- [2] A Smart System For Fake News Detection Using Machine Learning 2019 2nd International Conference on Issues and Challenges in Intelligent Computing Techniques.
- [3] Research on Text Classification for Identifying Fake News 2018 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC).
- [4] <https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9>
- [5] <https://towardsdatascience.com/nlp-text-preprocessing-a-practical-guide-and-templated80874676e79>
- [6] <https://machinelearningmastery.com/prepare-text-data-machine-learning-scikit-learn/>
- [7] Chaitra K Hiramath and Prof. G.C Deshpande "Fake News Detection Using Deep Learning Techniques" 2019 1st International Conference on Advances in Information Technology
- [8] Abhishek Verma, Vanshika Mittal and Suma Dawn "FIND: Fake Information and News Detections using Deep Learning"
- [9] Mykhailo Granik, Volodymyr Mesyura "Fake News Detection Using Naive Bayes Classifier" 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)
- [10] Shenhao Zhang, Yihui Wang and Chengxiang Tan "Research on Text Classification for Identifying Fake News" 2018 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)