FAKE NEWS DETECTION USING NLP

BATCH MEMBER

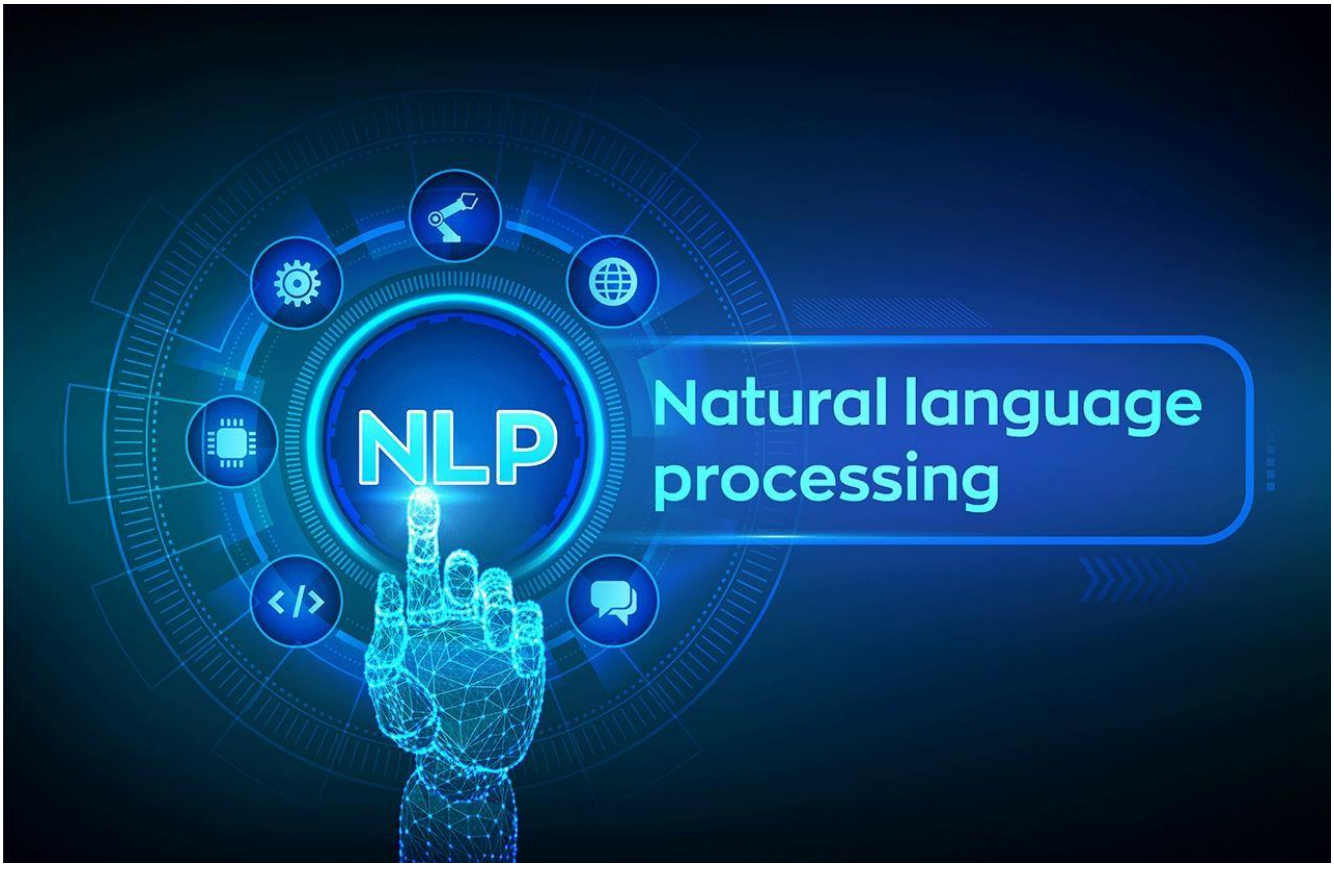
822621104307

P.THOOYAVAN

Phase 4 submission document

Project Title: FAKE NEWS DETECTION USING NLP

Phase 4: Development Part 2



Topic: *continue building the fake news detection model*

by feature engineering ,model training and evaluation

Introduction:

- The fake news has been rapidly increasing in numbers. It is not a new problem but recently it has been on a great rise. According to Wikipedia Fake news is false or misleading information presented as news.

- Detecting the fake news has been a challenging and a complex task. It is observed that humans have a tendency to believe the misleading information which makes the spreading of fake news even easier.

- Fake news is dangerous as it can deceive people easily and create a state of confusion among a community. This can further affect the society badly .The spread of fake news creates rumors circulating around and the victims could be badly impacted. Recent reports showed that due to the rise of fake news that was being created online it had impacted the US Presidential Elections.

- Text-Based Features: Leveraging the content of news articles or posts is paramount. Text-based features can include word frequencies, sentiment analysis, and the presence of specific keywords or phrases commonly associated with fake news.

- Source and Author Analysis: Features related to the credibility of the source and the history of the author can be informative. Metrics like source reputation and author reliability contribute to feature engineering.

- Social Network Analysis: For content on social media, features can be extracted from user interactions, such as the number of likes, shares, and comments. These social engagement metrics can reveal the potential virality of fake news.

- Linguistic and Style Analysis: Stylometric features, including writing style, readability, and grammar, can be indicative of fake news. Detecting anomalies in the linguistic characteristics of content is valuable.

Temporal and Historical Features: The time and context of a news article's publication can be vital. Features such as publication date, the history of the source, and trends in publishing times can help in detection.

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import re
import nltk as nlp
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split
# Reading datasets
fake_df = pd.read_csv("../input/fake-and-real-news-dataset/Fake.csv")
true_df = pd.read_csv("../input/fake-and-real-news-dataset/True.csv")
# Rows and columns of fake news dataset
```

```
print(f"Fake news dataset has: {fake_df.shape[0]} rows")
print(f"Fake news dataset has: {fake_df.shape[1]} columns")
```

Fake news dataset has: 23481 rows

Fake news dataset has: 4 column

Dealing with missing values if any

```
fake_df.isnull().sum()
```

Out[5]:

```
title      0
text       0
subject    0
date       0
dtype: int64
```

```
true_df.isnull().sum()
```

Out[6]:

```
title      0
text       0
subject    0
date       0
dtype: int64
```

Working on date column

```
# Adding 'Fake' column to our datasets then join them together
fake_df['Fake'] = 1
true_df['Fake'] = 0
df = pd.concat([fake_df, true_df])
# Extracting the year and the month
df['date'] = pd.to_datetime(df['date'], errors='coerce')
df['Year'] = df['date'].dt.year
```

```
df['Month'] = df['date'].dt.month
```

EDA - Exploratory Data Analysis

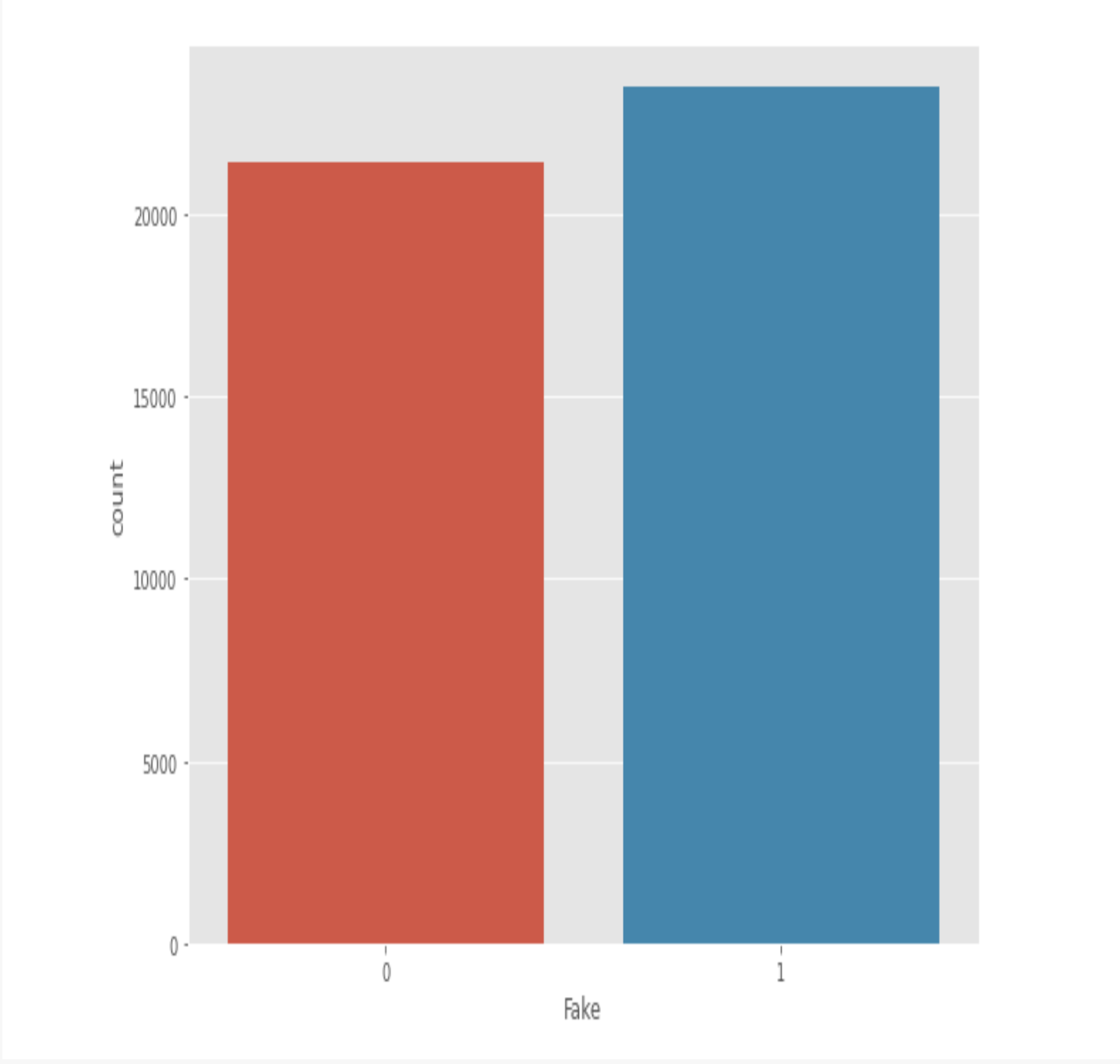
```
plt.style.use('ggplot')

plt.figure(figsize=(10,7))

sns.countplot(data=df, x='Fake')
```

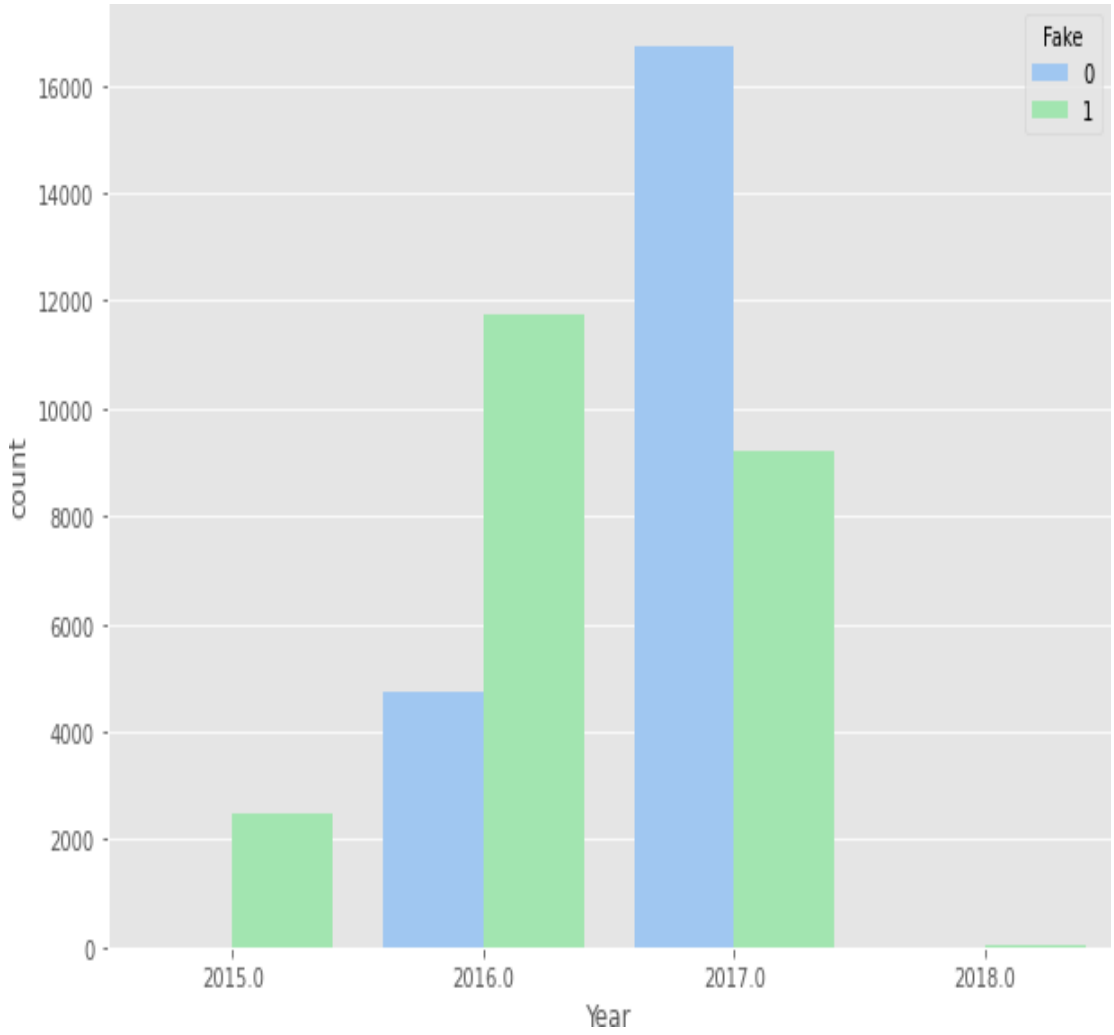
Out[9]:

<AxesSubplot:xlabel='Fake', ylabel='count'>



```
aborn-pastel') plt.style
plt.figure(figsize=(10, 7))
sns.countplot(data=df, x='Year', hue='Fake')
```

<AxesSubplot:xlabel='Year', ylabel='count'>




```
plt.style.use('bmh')
plt.figure(figsize=(12, 7))
sns.countplot(data=df, x='Month', hue='Fake')
```

Out[11]:

<AxesSubplot:xlabel='Month', ylabel='count'>

 # Subjects count

```
df.subject.value_counts()
```

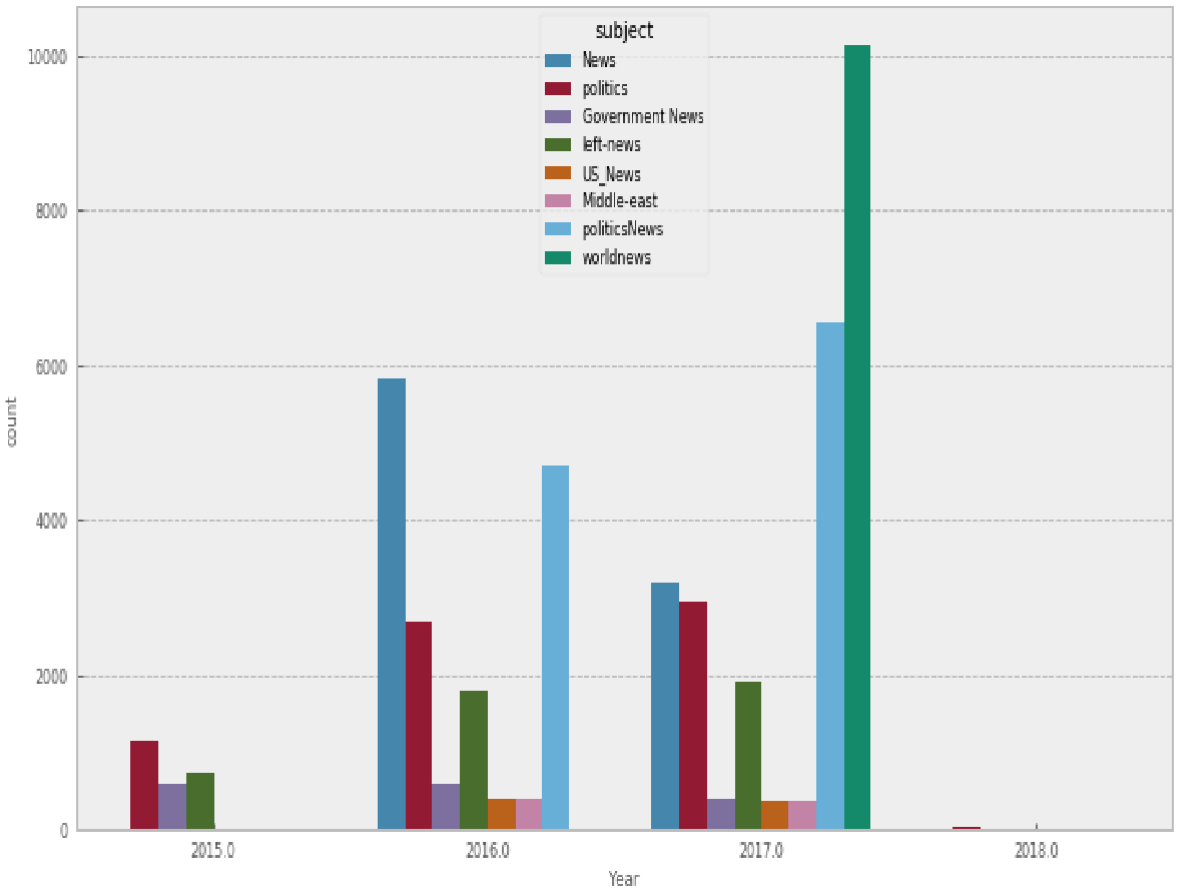
Out[12]:

```
politicsNews       11272
worldnews          10145
News                9050
politics            6841
left-news           4459
Government News     1570
US_News              783
Middle-east          778
Name: subject, dtype: int64
```

```
plt.style.use('seaborn-paper')
plt.figure(figsize=(12, 7))
sns.countplot(data=df, x='Year', hue='subject')
```

Out[13]:

<AxesSubplot:xlabel='Year', ylabel='count'>



A. Dataset

There are very few datasets which are available publicly for the detection of fake news. In this paper we have used three different datasets which are available online. The first d ataset ISOT Fake News dataset is obtained from a website[17]. The second data that is u sed in the project is the Fake News Detection dataset from Kaggle[18]. The third dataset used is the Real and Fake News Dataset which is obtained from Kaggle[19].

Merging Dataset

The first dataset ISOT Fake News dataset is obtained from a website[9]. This dataset wa s created using data from real world news sources. This dataset consists of two types of articles: fake and real. The dataset consists of two CSV files. First file contains all the n ews which is true and the second file contains the news which is fake.

Each article contains the following information: article title, text ,type and the date the a rticle was published on. The second dataset used in the project is the Fake News Detecti on dataset from Kaggle. This dataset consists of 4 columns which are the URLs of the n

ews source, the Headline of the news, the Body of the news that is the content of the ne ws and the last column contains the Label of the news which tells whether the news is fa ke or not. Next, the two datasets are merged together to obtain a single dataset. After the merge we obtained a dataset with 10344 records. Finally, we obtain a master dataset by merging the first dataset with the above merged dataset [dataset with 10344 records], he nce the final obtained master dataset consists of 54726 records and three columns , Title , text and Class. Data preprocessing is a data mining technique that involves transformin g raw data into understandable form. In natural language processing, text preprocessing is the practice of cleaning and preparing text data. Real-world data is often incomplete, i nconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many e rrors. Data preprocessing is a proven method of resolving such issues. Data preprocessin g methods such as tokenization, lemmatization, stop word removal and lowercasing

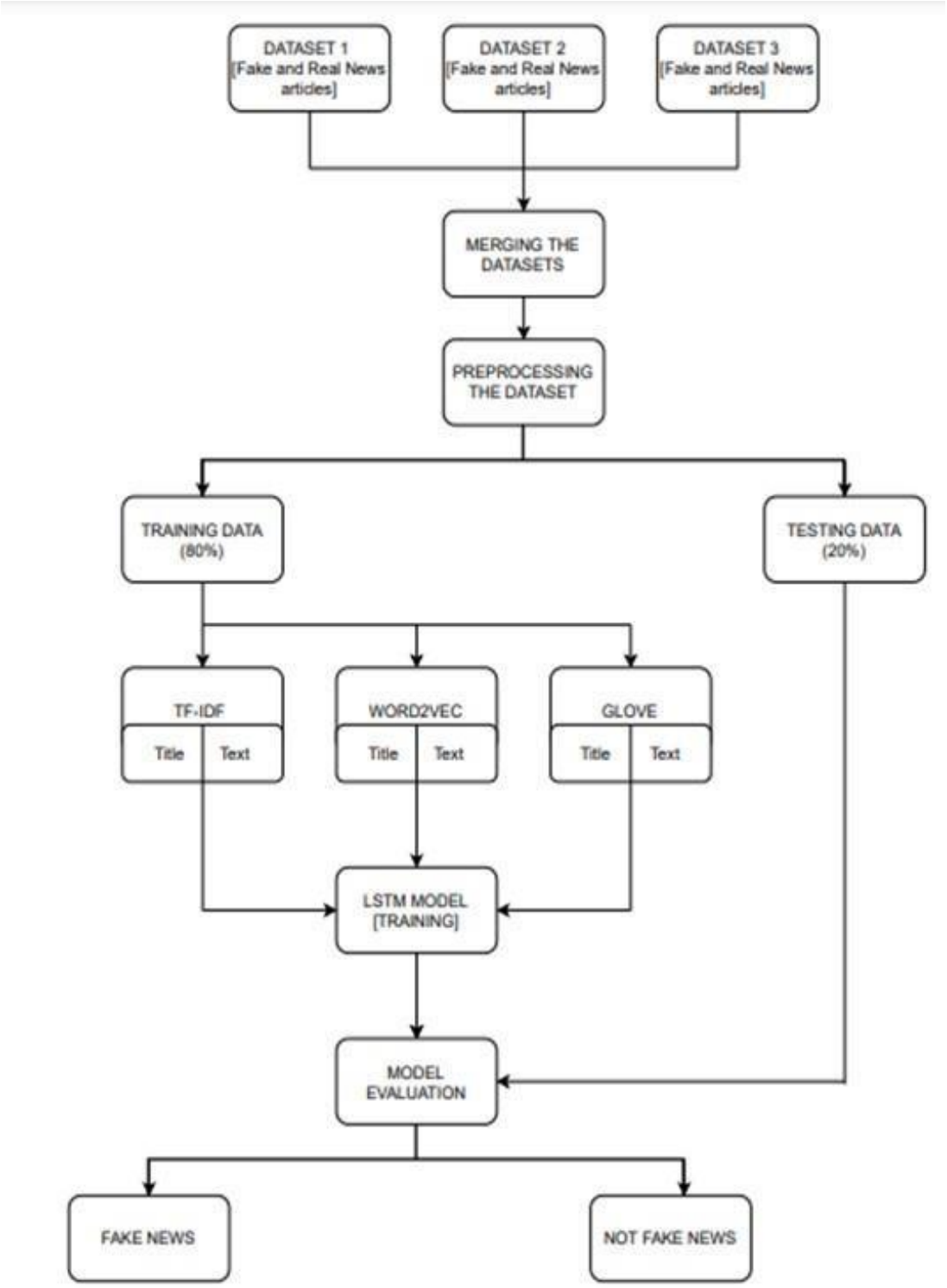


Fig.1 Proposed System Model for Detection of Fake News.

‘Fig.2’ shows thee number of fake news and real news in the dataset. We have used wor d clouds to check which are the words which appear frequently in the fake and real new s. ‘Fig.3’ shows the Word Cloud for real news and Fig. 5.4 shows the Word Cloud for f ake news.

RESULTS

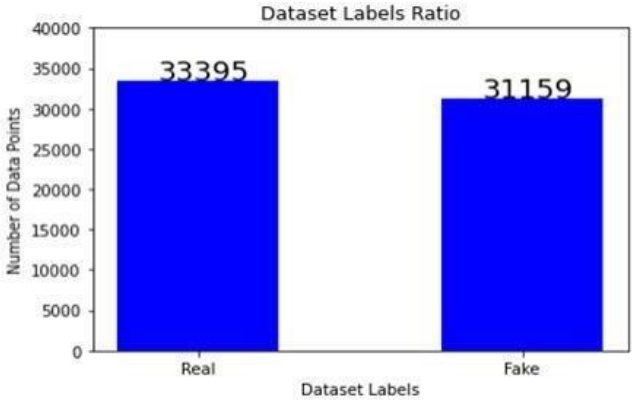


Fig.2. Dataset Labels Ratio.

Fig.3. Word Cloud for Real News.

After the models were trained we calculated the performance metrics accuracy, precision and recall. 'Table.I' shows the performance metrics of the models.

Model	Accuracy	Precision	Recall
MODEL 1 [Fed with text vectors of 'Title' obtained by GloVe]	89.71%	91.84%	86.61%
MODEL 2 [Fed with text vectors of 'Text' obtained by GloVe]	92.8%	92.3%	92.9%
MODEL 3 [Fed with text vectors of 'Title' obtained by Word2Vec]	86.49%	87.14%	84.04%
MODEL 4 [Fed with text vectors of 'Text' obtained by Word2Vec]	96.26%	95.40%	96.86%
MODEL 5 [Fed with text vectors of 'Title' obtained by TF-IDF]	80%	70%	70%
MODEL 6 [Fed with text vectors of 'Text' obtained by TF-IDF]	81%	71%	76%

The performance is measured using the accuracy , precision and recall.

Accuracy: It shows the overall accuracy of the instances which are correctly classified to the total number of the instances. It is calculated by the following formula:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (1)$$

Where, TP = true positive, TN = true negative, FP = false positive, FN = false negative.

Precision: It represents the percentage of relevant sarcastic headlines. That is, it measures the amount of headlines categorized as sarcastic against the total number of headlines classified as sarcastic. It is calculated by the following formula:

$$Precision = \frac{TP}{TP+FP} \qquad (2)$$

Recall: It represents the percentage of relevant sarcastic headlines that have been searched. That is, against the total number of sarcastic headlines, measured the number of headlines that are normally classified as sarcastic. It is calculated by the following formula:

$$Recall = \frac{TP}{TP+FN} \qquad (3)$$

"Table 1" shows the results obtained by the models. From the results obtained we can observe that the model trained using the content of the news gives better output than the other models. Also, we can see that the models which have used GloVe and WordVec method work better than the models using TF-IDF.

ACKNOWLEDGMENT

The heading of the Acknowledgment section and the References section must not be nu mbered.Causal Productions wishes to acknowledge Michael Shell and other contributors for developing and maintaining the IEEE LaTeX style files which have been used in the preparation of this template. To see the list of contributors, please refer to the top of file IEEETran.cls in the IEEE LaTeX distribution.

A. Text Font of Entire Document

The entire document should be in Times New Roman or Times font. Type 3 fonts must not be used. Other font types may be used if needed for special purposes.
Recommended font sizes are shown in Table 1.

B. Title and Author Details

Title must be in 24 pt Regular font. Author name must be in 11 pt Regular font. Author affiliation must be in 10 pt Italic. Email address must be in 9 pt Courier Regular font.

TABLE II
FONT SIZES FOR PAPERS

Font Size	Appearance (in Time New Roman or Times)		
	Regular	Bold	Italic
8	table caption (in Small Caps), figure caption, reference item		reference item (partial)
9	author email address (in Courier), cell in a table	abstract body	abstract heading (also in Bold)
10	level-1 heading (in Small Caps), paragraph		level-2 heading, level-3 heading, author affiliation
11	author name		
24	title		

All title and author details must be in single-column format and must be centered.
Every word in a title must be capitalized except for short minor words such as "a", "an", "and", "as", "at", "by", "for", "from", "if", "in", "into", "on", "or", "of", "the", "to", "with".
Author details must not show any professional title (e.g. Managing Director), any academic title (e.g. Dr.) or any membership of any professional organization (e.g. Senior Member IEEE).
To avoid confusion, the family name must be written as the last part of each author name (e.g. John A.K. Smith).
Each affiliation must include, at the very least, the name of the company and the name of the country where the author is based (e.g. Causal Productions Pty Ltd, Australia).
Email address is compulsory for the corresponding author.

C. Section Headings

No more than 3 levels of headings should be used. All headings must be in 10pt font. Every word in a heading must be capitalized except for short minor words as listed in Section III-B.

1) *Level-1 Heading*: A level-1 heading must be in Small Caps, centered and numbered using uppercase Roman numerals. For example, see heading "III. Page Style" of this document. The two level-1 headings which must not be numbered are "Acknowledgment" and "References".
2) *Level-2 Heading:* A level-2 heading must be in Italic, left-justified and numbered using an uppercase alphabetic letter followed by a period. For example, see heading "C. Section Headings" above
3) *Level-3 Heading:* A level-3 heading must be indented, in Italic and numbered with an Arabic numeral followed by a right parenthesis. The level-3 heading must end with a colon. The body of the level-3 section immediately follows the level-3 heading in the same paragraph. For example, this paragraph begins with a level-3 heading.

Figures and Tables

Figures and tables must be centered in the column. Large figures and tables may span a cross both columns. Any table or figure that takes up more than 1 column width must b e positioned either at the top or at the bottom of the page.

Graphics may be full color. All colors will be retained on the CDROM. Graphics must not use stipple fill patterns because they may not be reproduced properly. Please use on ly *SOLID FILL* colors which contrast well both on screen and on a black-and-white hard copy

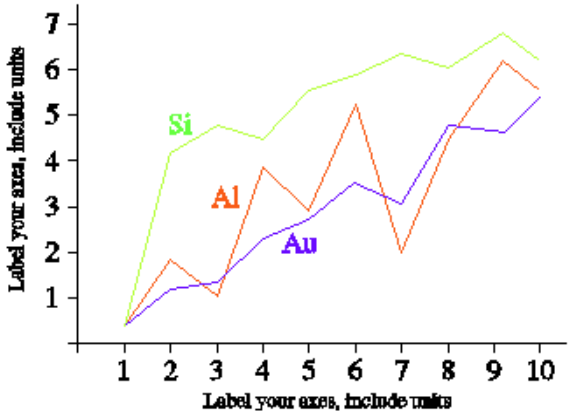


Fig. 1 A sample line graph using colors which contrast well both on screen and on a black-and-white hardcopy

Fig. 2 shows an example of a low-resolution image which would not be acceptable, whe reas Fig. 3 shows an example of an image with adequate resolution. Check that the reso lution is adequate to reveal the important detail in the figure.

Please check all figures in your paper both on screen and on a black-and-white hardcopy When you check your paper on a black-and-white hardcopy, please ensure that.

- The colors used in each figure contrast well,
- The image used in each figure is clear,
- All text labels in each figure are legible

E. Figure Captions

Figures must be numbered using Arabic numerals. Figure captions must be in 8 pt Regu lar font. Captions of a single line (e.g. Fig. 2) must be centered whereas multi-line capti ons must be justified (e.g. Fig. 1). Captions with figure numbers must be placed after th eir associated figures, as shown in Fig. 1.

Fig. 3 Example of an image with acceptable resolution

Fig. 2 Example of an unacceptable low-resolution image

Fig. 3 Example of an image with acceptable resolution

F. *Table Captions*

Tables must be numbered using uppercase Roman numerals. Table captions must be ce ntred and in 8 pt Regular font with Small Caps. Every word in a table caption must be c apitalized except for short minor words as listed in Section III-B. Captions with table n umbers must be placed before their associated tables, as shown in Table 1.

G. *Page Numbers, Headers and Footers*

Page numbers, headers and footers must not be used.

H. *Links and Bookmarks*

All hypertext links and section bookmarks will be removed from papers during the proc essing of papers for publication. If you need to refer to an Internet email address or UR L in your paper, you must type out the address or URL fully in Regular font.

I. *References*

The heading of the References section must not be numbered. All reference items must be in 8 pt font. Please use Regular and Italic styles to distinguish different fields as sho wn in the References section. Number the reference items consecutively in square brac kets (e.g. [1]).

When referring to a reference item, please simply use the reference number, as in [2]. D o not use “Ref. [3]” or “Reference [3]” except at the beginning of a sentence, e.g. “Refe rence [3] shows …”. Multiple references are each numbered with separate brackets (e.g . [2], [3], [4]–[6]

Conclusion:

- In the quest to build a house price prediction model, we have embarked on a critical journey that begins with loading and preprocessing the dataset.We have traversed through essential steps, starting with importing the necessary libraries to facilitate data manipulation and analysis.
- Understanding the data's structure, characteristics, and any potential issues through exploratory data analysis (EDA) is essential for informed decision-making.
- Data preprocessing emerged as a pivotal aspect of this process. It involves cleaning, transforming, and refining the dataset to ensure that it aligns with the requirements of machine learning algorithms.
- With these foundational steps completed, our dataset is now primed for the subsequent stages of building and training a house price prediction model.