Real or Fake News Detection

BY GROUP 2:

1. Nampelly Abhinay.
2. Rutuja Sukhasare.
3. Subhash Dasari.
4. Sahil Kaladgi.
5. Hanmantagoud Patil.
6. Akshata Punde.

Introduction

Fake news has quickly become a society problem, being used to propagate false or rumour information in order to change peoples behaviour.

Misinformation can lead to public confusion, affecting opinions and behaviors on critical issues like health, politics and the economy.

The performance of attention mechanism for fake news detection on two datasets, one containing true data and second one news from fake dataset.

Objective:

This project aims to develop an NLP-based model to accurately classify news articles as real or fake.

The ultimate aim is to develop a robust and accurate system for detecting fake news using NLP techniques, thereby contributing to efforts in combating misinformation and promoting media literacy in society.

Datasets:

Dataset 1: True Data

Size: 21417 articles

Columns: Title, text, subject, date

Dataset 2: Fake Data

Size: 23502 articles

Columns: Title, text, subject, date

Exploratory Data Analysis

Data Analysis :

- Check for data types and null values present in datasets.

Data Visualization :

1. Pie chart: To check values present in both the dataset.
2. Histogram: To check text length.
3. Density Plot: To check density of both dataset.
4. Wordcloud: To get most frequent words from both the datasets.

```
true_data.head()
```

]:

	title	text	subject	date
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017

```
fake_data.head()
```

]:

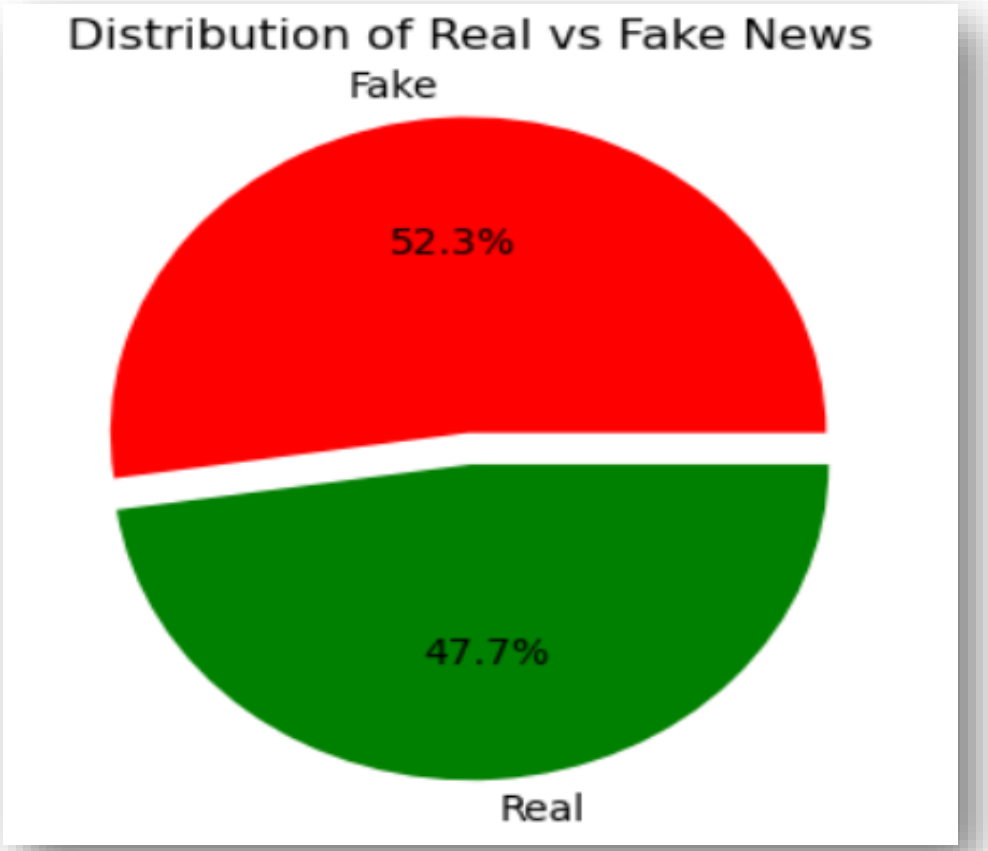
	title	text	subject	date
0	Donald Trump Sends Out Embarrassing New Year□...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017
3	Trump Is So Obsessed He Even Has Obama□s Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017

```
# Add labels for both datasets
true_data['label'] = 1  # Real news
fake_data['label'] = 0  # Fake news
```

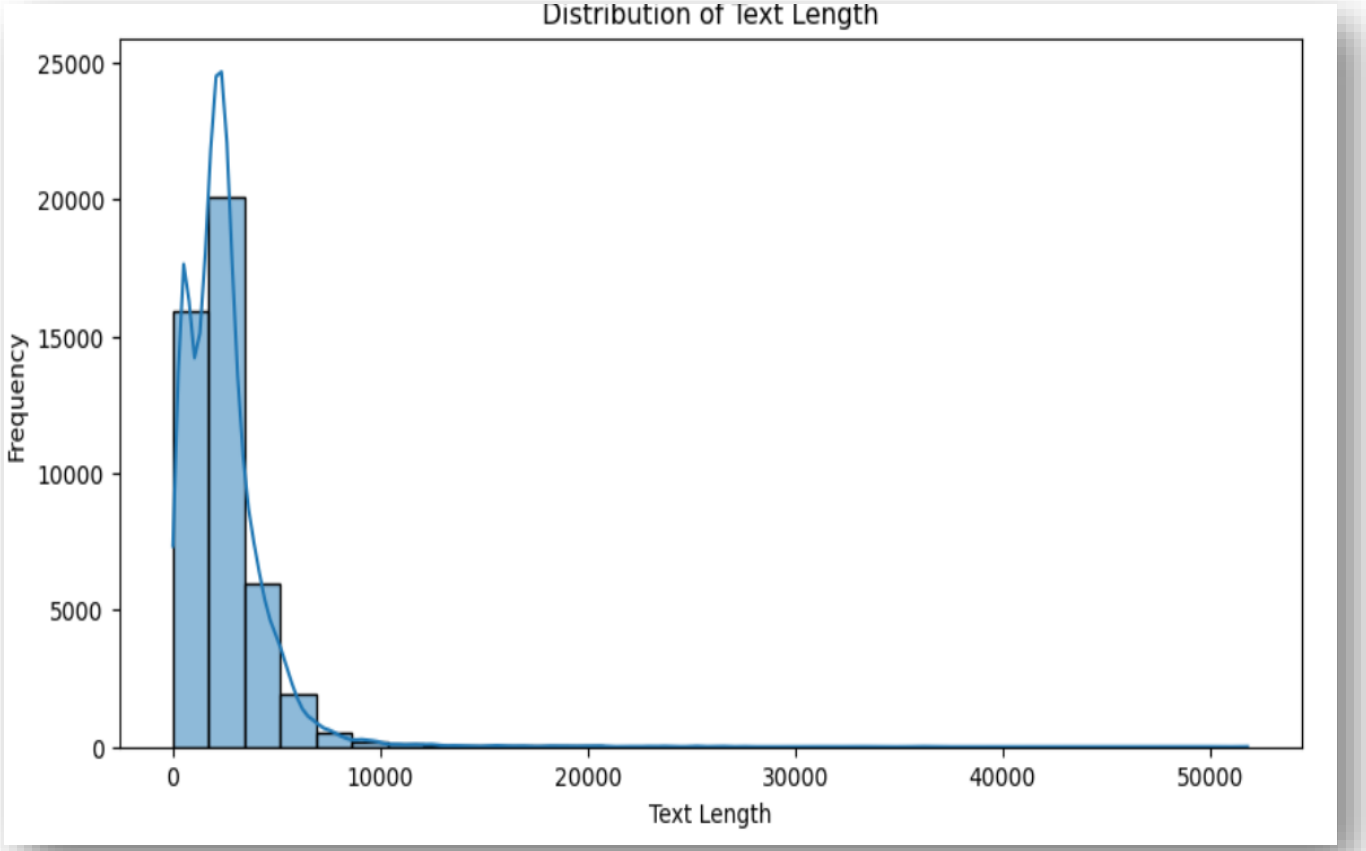
```
df=pd.concat([true_data, fake_data])
df
```

	title	text	subject	date	label
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017	1
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	1
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	1
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017	1
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017	1
...	...	...	...	...	...
23476	McPain: John McCain Furious That Iran Treated ...	21st Century Wire says As 21WIRE reported earl...	Middle-east	January 16, 2016	0
23477	JUSTICE? Yahoo Settles E-mail Privacy Class-ac...	21st Century Wire says It s a familiar theme. ...	Middle-east	January 16, 2016	0
23478	Sunnistan: US and Allied □Safe Zone□ Plan to T...	Patrick Henningsen 21st Century WireRemember ...	Middle-east	January 15, 2016	0
23479	How to Blow $700 Million: Al Jazeera America F...	21st Century Wire says Al Jazeera America will...	Middle-east	January 14, 2016	0
23480	10 U.S. Navy Sailors Held by Iranian Military ...	21st Century Wire says As 21WIRE predicted in ...	Middle-east	January 12, 2016	0

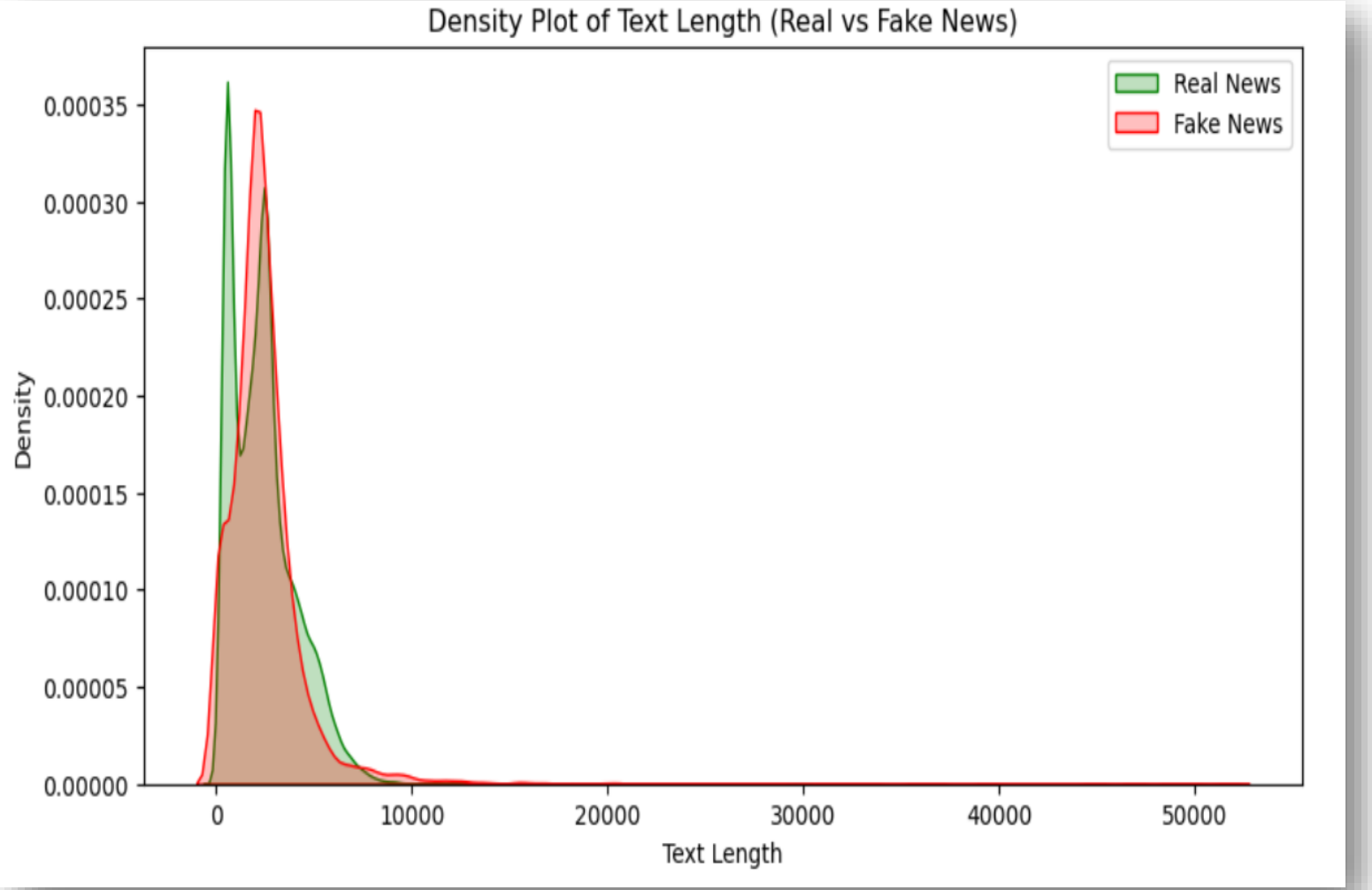
44897 rows × 5 columns



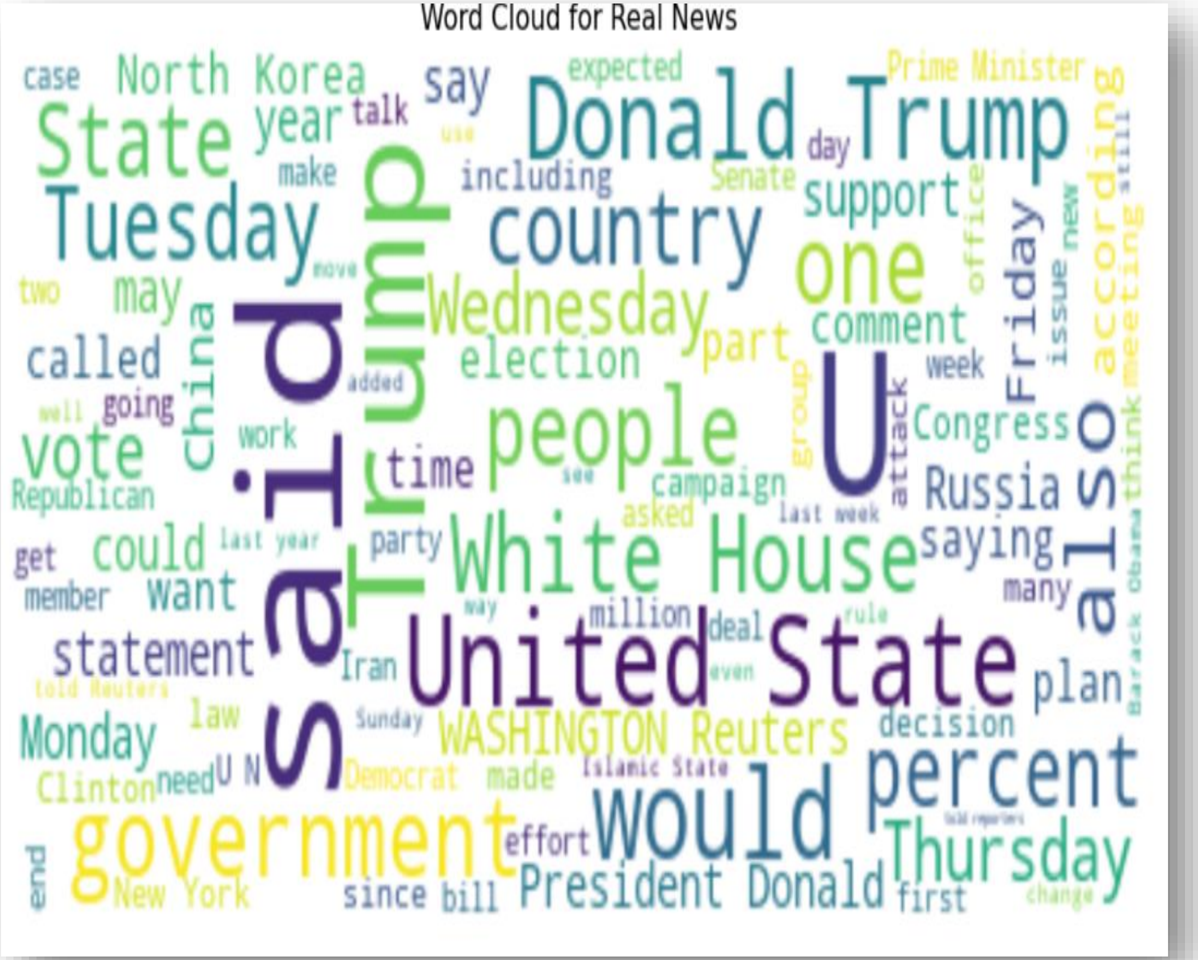
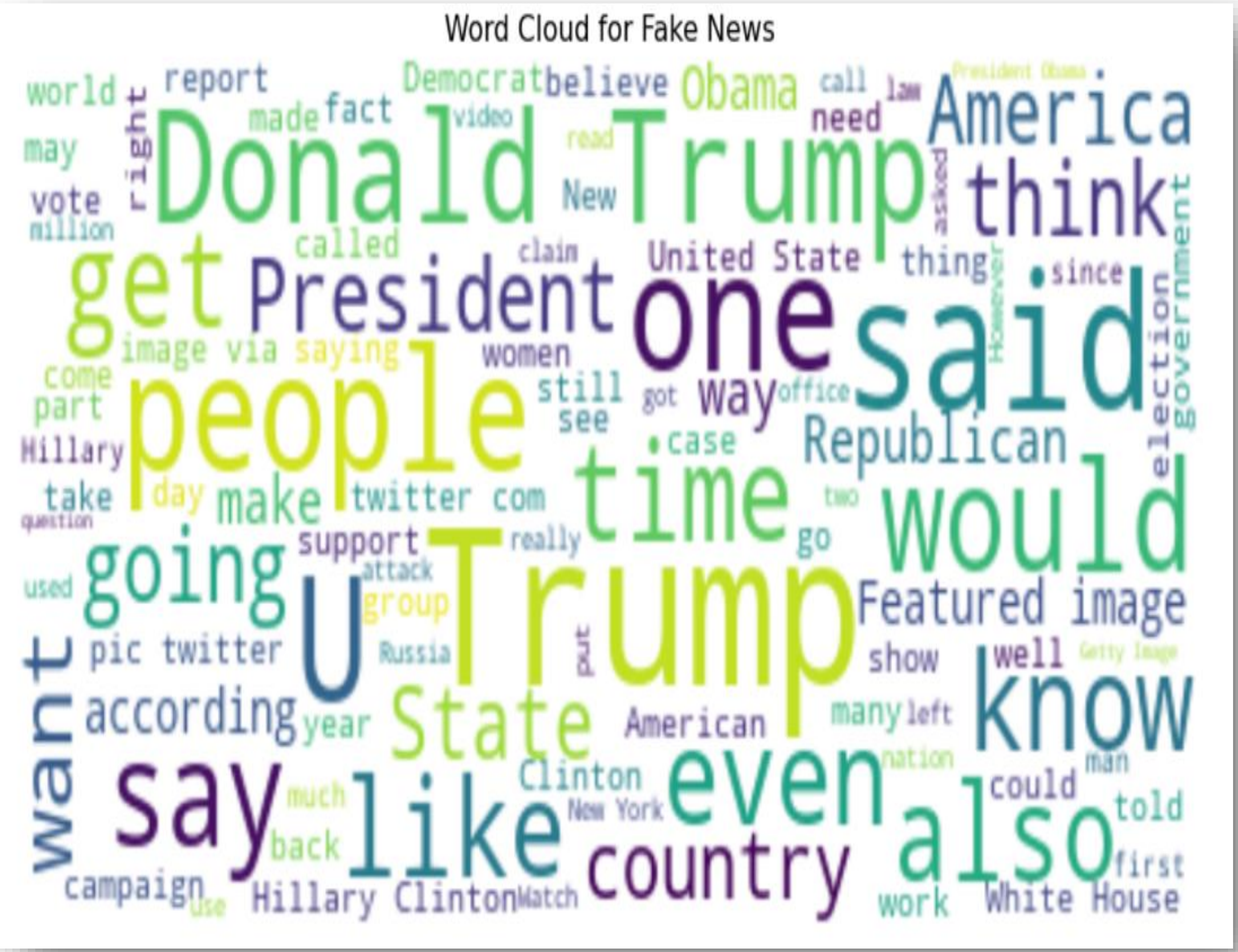
Pie Chart



Histogram



Density Plot


Word Cloud for Real News
said
Trump
Donald Trump
United State
White House
U
State
Tuesday
country
one
Wednesday
people
government
would
percent
Thursday
also
North Korea
President Donald
WASHINGTON Reuters
election
support
comment
Friday
Congress
Russia
statement
Monday
vote
called
China
Republican
could
want
time
say
year
may
plan
saying
decision

Word Cloud for Fake News
Donald Trump
Trump
said
people
one
President
get
would
time
know
America
think
even
also
like
say
country
State
going
want
U
Republican
Featured image
United State
Obama
according
make
twitter com
pic twitter
image via saying
Hillary Clinton
White House
Democrat
campaign
way
still

Wordcloud

```
# Preprocessing the Data:
```

```
import re
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
import nltk

# Download necessary resources
nltk.download('punkt')
nltk.download('stopwords')

# Text preprocessing function
def preprocess_text(text):
    text = re.sub(r'[^A-Za-z\s]', '', text)  # Remove punctuation and numbers
    text = text.lower()  # Convert to lowercase
    words = word_tokenize(text)  # Tokenize the text
    stop_words = set(stopwords.words('english'))
    words = [word for word in words if word not in stop_words]  # Remove stopwords
    return ' '.join(words)

# Apply preprocessing to both datasets
true_data['clean_text'] = true_data['text'].apply(preprocess_text)
fake_data['clean_text'] = fake_data['text'].apply(preprocess_text)

# Check results
print(true_data[['text', 'clean_text']].head())
print(fake_data[['text', 'clean_text']].head())
```

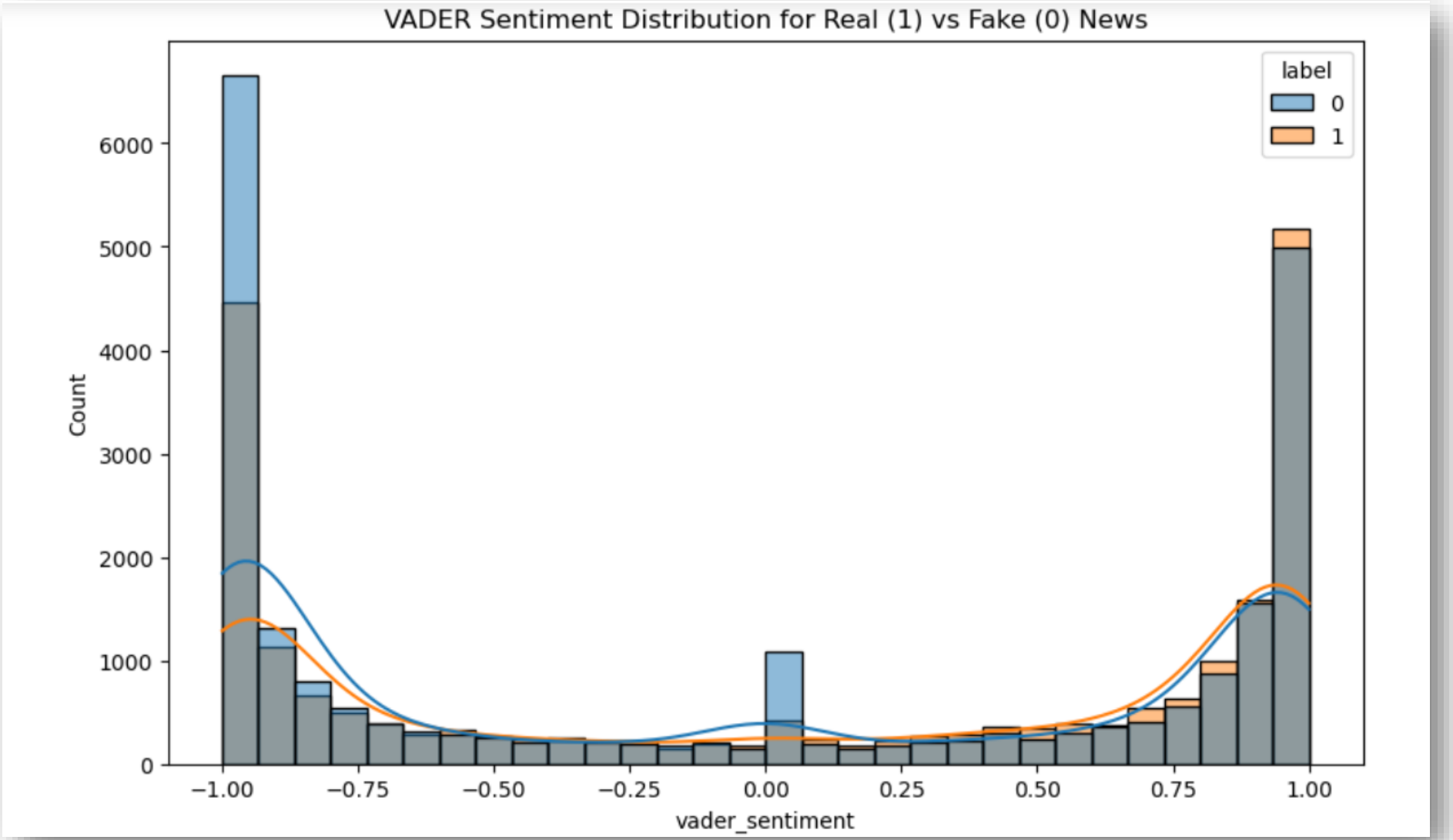
```
                                                text  \
0  WASHINGTON (Reuters) - The head of a conservat...
1  WASHINGTON (Reuters) - Transgender people will...
2  WASHINGTON (Reuters) - The special counsel inv...
3  WASHINGTON (Reuters) - Trump campaign adviser ...
4  SEATTLE/WASHINGTON (Reuters)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

                                          clean_text
0  washington reuters head conservative republica...
1  washington reuters transgender people allowed ...
2  washington reuters special counsel investigati...
3  washington reuters trump campaign adviser geor...
4  seattlewashington reuters president donald tru...
                                                text  \
0  Donald Trump just couldn t wish all Americans ...
1  House Intelligence Committee Chairman Devin Nu...
2  On Friday, it was revealed that former Milwauk...
3  On Christmas day, Donald Trump announced that ...
4  Pope Francis used his annual Christmas Day mes...

                                          clean_text
0  donald trump wish americans happy new year lea...
1  house intelligence committee chairman devin nu...
2  friday revealed former milwaukee sheriff david...
3  christmas day donald trump announced would bac...
4  pope francis used annual christmas day message...
```

Sentiment Analysis: Reveal whether fake news articles tend to be more positive, negative or neutral compared to real news.

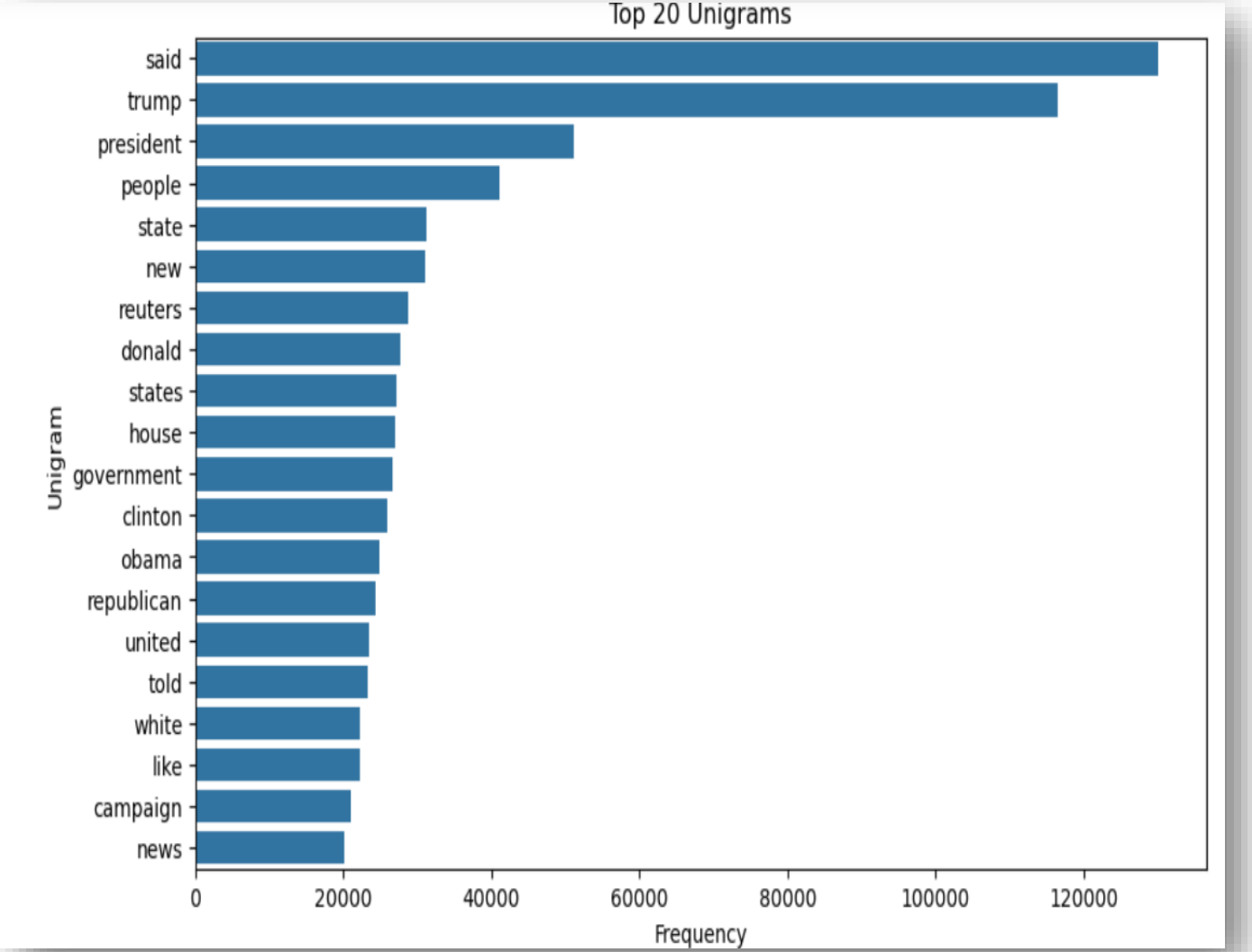
```python
# vader:
from nltk.sentiment import SentimentIntensityAnalyzer
import nltk

# Download the VADER lexicon (if not already done)
nltk.download('vader_lexicon')

# Initialize the VADER sentiment analyzer
sia = SentimentIntensityAnalyzer()

# Function to get VADER sentiment scores
def get_vader_sentiment(text):
    sentiment = sia.polarity_scores(text)
    return sentiment['compound']  # Compound score is the overall sentiment

# Apply sentiment analysis to the 'clean_text' column in the merged dataset
fdf['vader_sentiment'] = fdf['clean_text'].apply(get_vader_sentiment)

# Check the sentiment analysis results
print(fdf[['clean_text', 'vader_sentiment']].head())
```

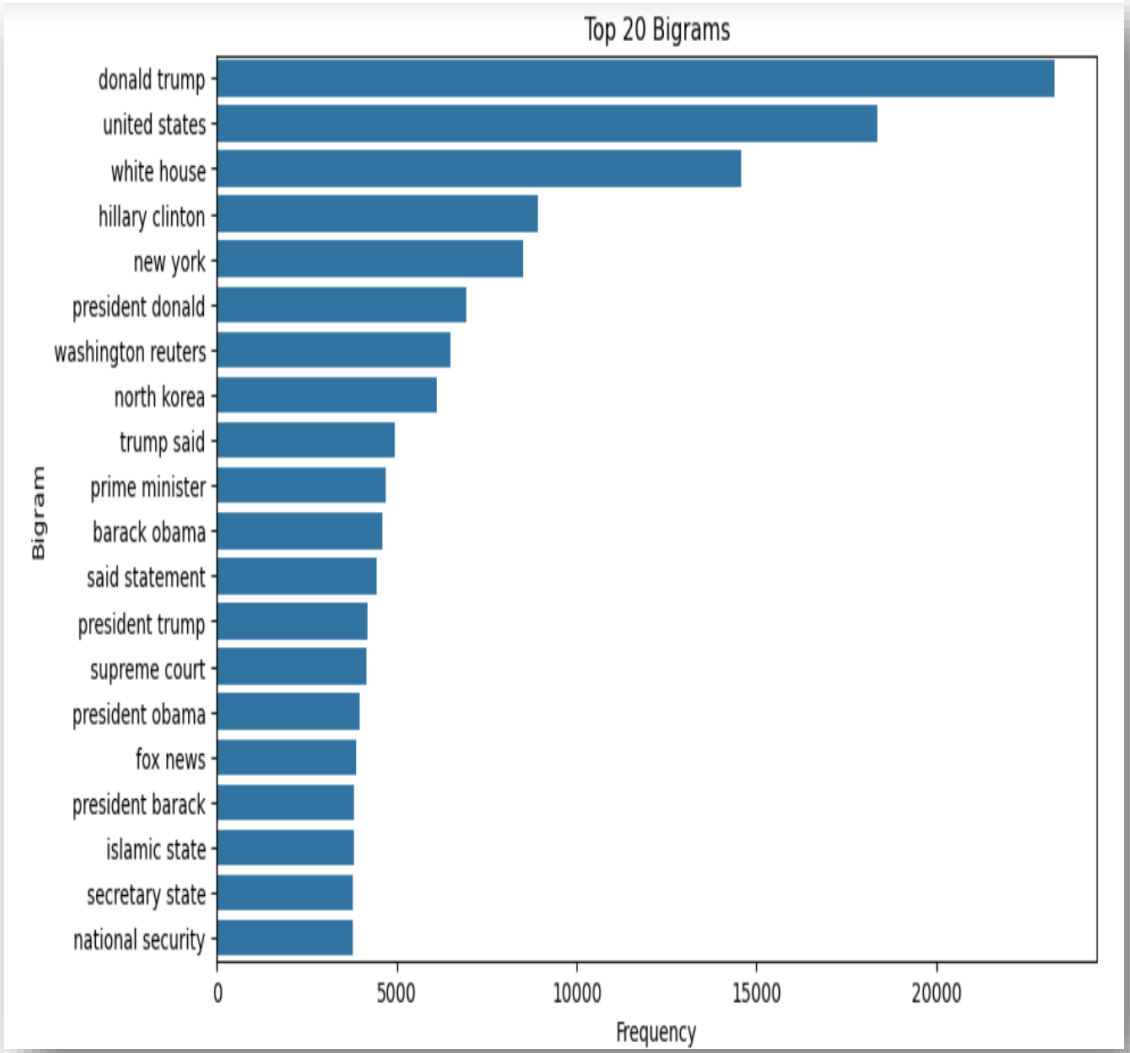
```
[nltk_data] Downloading package vader_lexicon to
[nltk_data]     C:\Users\subha\AppData\Roaming\nltk_data...
[nltk_data]   Package vader_lexicon is already up-to-date!
```

	clean_text	vader_sentiment
0	washington reuters head conservative republica...	0.9831
1	washington reuters transgender people allowed ...	0.9578
2	washington reuters special counsel investigati...	0.5719
3	washington reuters trump campaign adviser geor...	-0.1761
4	seattlewashington reuters president donald tru...	0.9670

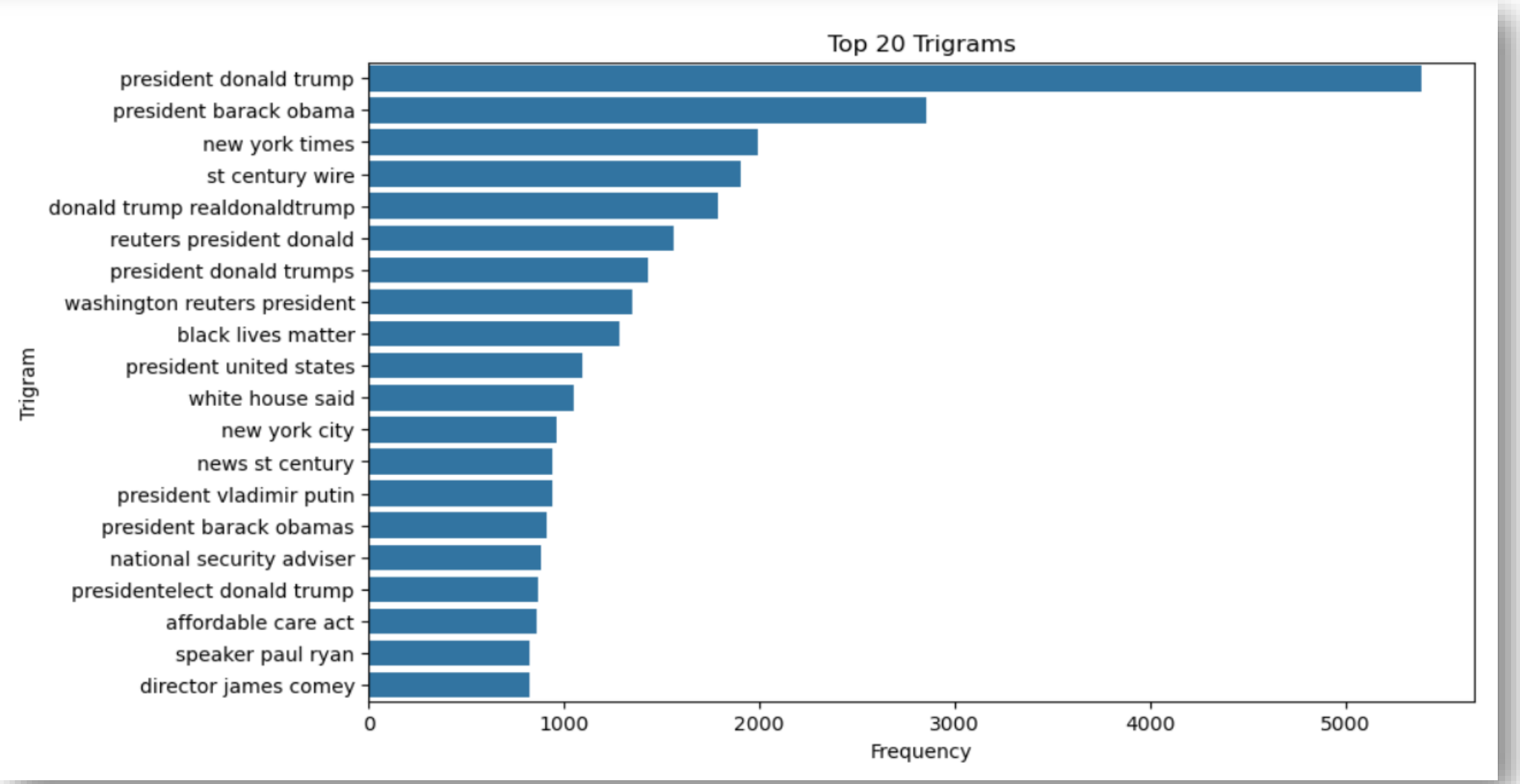

VADER Sentiment Distribution for Real (1) vs Fake (0) News
label
0
1
Count
6000
5000
4000
3000
2000
1000
0
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
vader_sentiment

N- Gram Analysis: Used to examine contiguous sequences of words in a given text.





N- Gram Analysis:



Model Building:

Logistic Regression: is a classification algorithm that models the probability that an input belongs to a particular class.

```python
# 1. Logistic Regression:
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report, accuracy_score

# Initialize the Logistic Regression model
logistic_model = LogisticRegression(max_iter=1000)

# Train the model
logistic_model.fit(X_train, y_train)

# Predict on the test set
y_pred_logistic = logistic_model.predict(X_test)

# Evaluate the model
print("Logistic Regression Accuracy:", accuracy_score(y_test, y_pred_logistic))
print("Classification Report for Logistic Regression:\n", classification_report(y_test, y_pred_logistic))
```

```
Logistic Regression Accuracy: 0.9891982182628062
Classification Report for Logistic Regression:
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      4644
           1       0.99      0.99      0.99      4336

    accuracy                           0.99      8980
   macro avg       0.99      0.99      0.99      8980
weighted avg       0.99      0.99      0.99      8980
```

Random Forest: An ensemble learning method that constructs multiple decision trees and outputs the majority vote for classification.

```python
# 2. Random Forest:
from sklearn.ensemble import RandomForestClassifier

# Initialize the Random Forest model
rf_model = RandomForestClassifier(n_estimators=100, random_state=42)

# Train the model
rf_model.fit(X_train, y_train)

# Predict on the test set
y_pred_rf = rf_model.predict(X_test)

# Evaluate the model
print("Random Forest Accuracy:", accuracy_score(y_test, y_pred_rf))
print("Classification Report for Random Forest:\n", classification_report(y_test, y_pred_rf))
```

```
Random Forest Accuracy: 0.9981069042316258
Classification Report for Random Forest:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      4644
           1       1.00      1.00      1.00      4336

    accuracy                           1.00      8980
   macro avg       1.00      1.00      1.00      8980
weighted avg       1.00      1.00      1.00      8980
```

Conclusion:

The **Random Forest model** provided more accurate and reliable classification of fake vs real news due to its ability to capture complex patterns and its robustness to noise in the data.

Thank You