FAKE NEWS DETECTION USING NLP IN ARTIFICIAL INTELLIGENCE

TEAM MEMBER

au820421205071: SUBANU R S

Phase-3 SUBMISSION DOCUMENT

Project: Fake News Detection

Phase 3: ***Development Part 1***

TOPIC: Begin building the fake news detection model by loading and preprocessing the dataset. Load the fake news dataset and preprocess the textual data.

FAKE NEWS DETECTION

ABSTRACT

- The proliferation of fake news in the digital age has become a significant challenge in maintaining the integrity of information dissemination.
- To address this issue, the development of a fake news detection model is essential.
- This abstract outlines the initial steps involved in building such a model, focusing on the critical processes of loading and preprocessing the dataset.

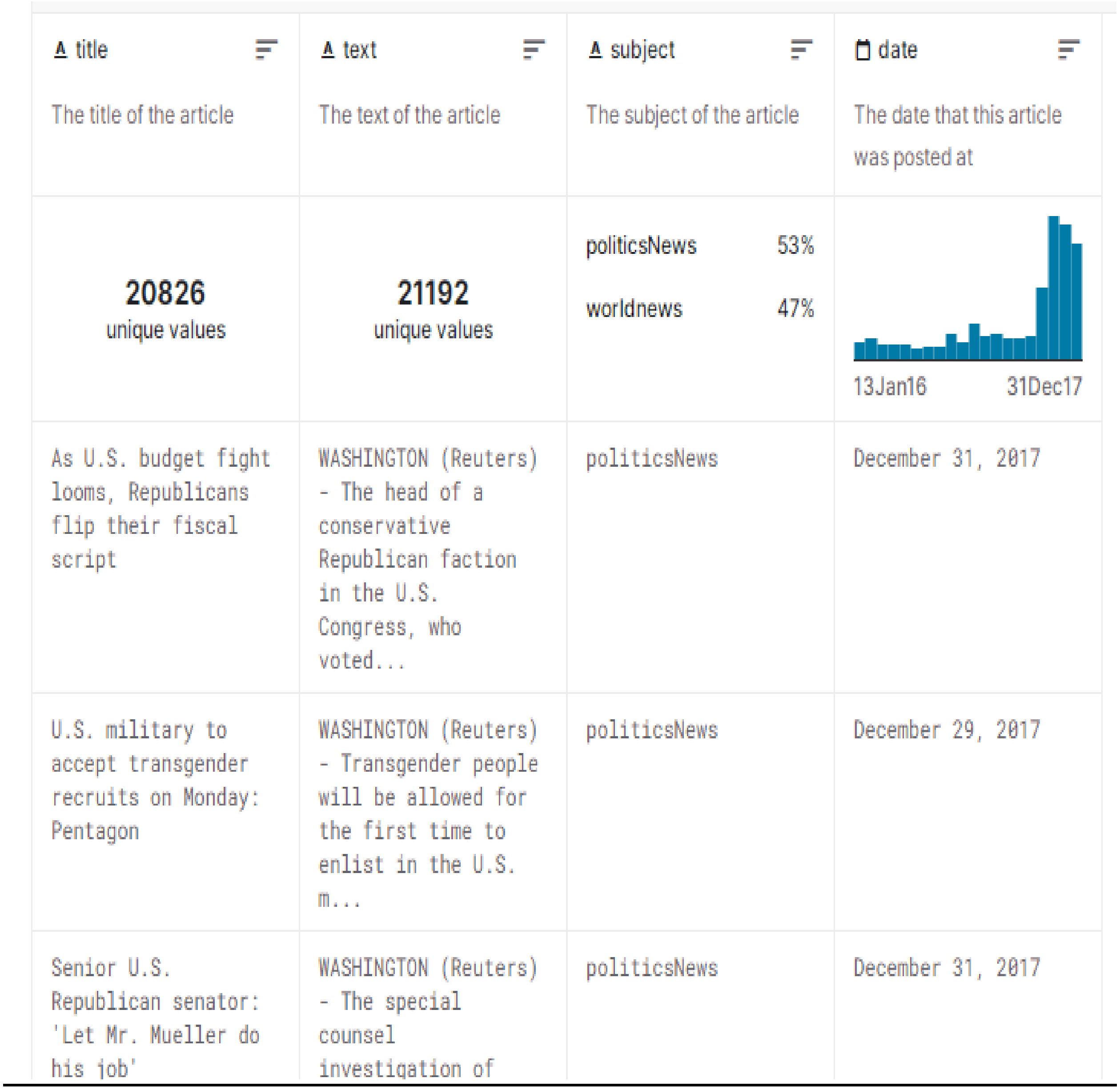
INTRODUCTION

Building a fake news detection model involves several steps, with one of the initial steps being loading and preprocessing the dataset. In this example, I'll provide a general outline of how to load and preprocess a fake news dataset using Python. For this purpose, we'll use Python and popular libraries like Pandas and NLTK. Please note that you'll need to have your dataset in a format suitable for this process.

GIVEN DATASET

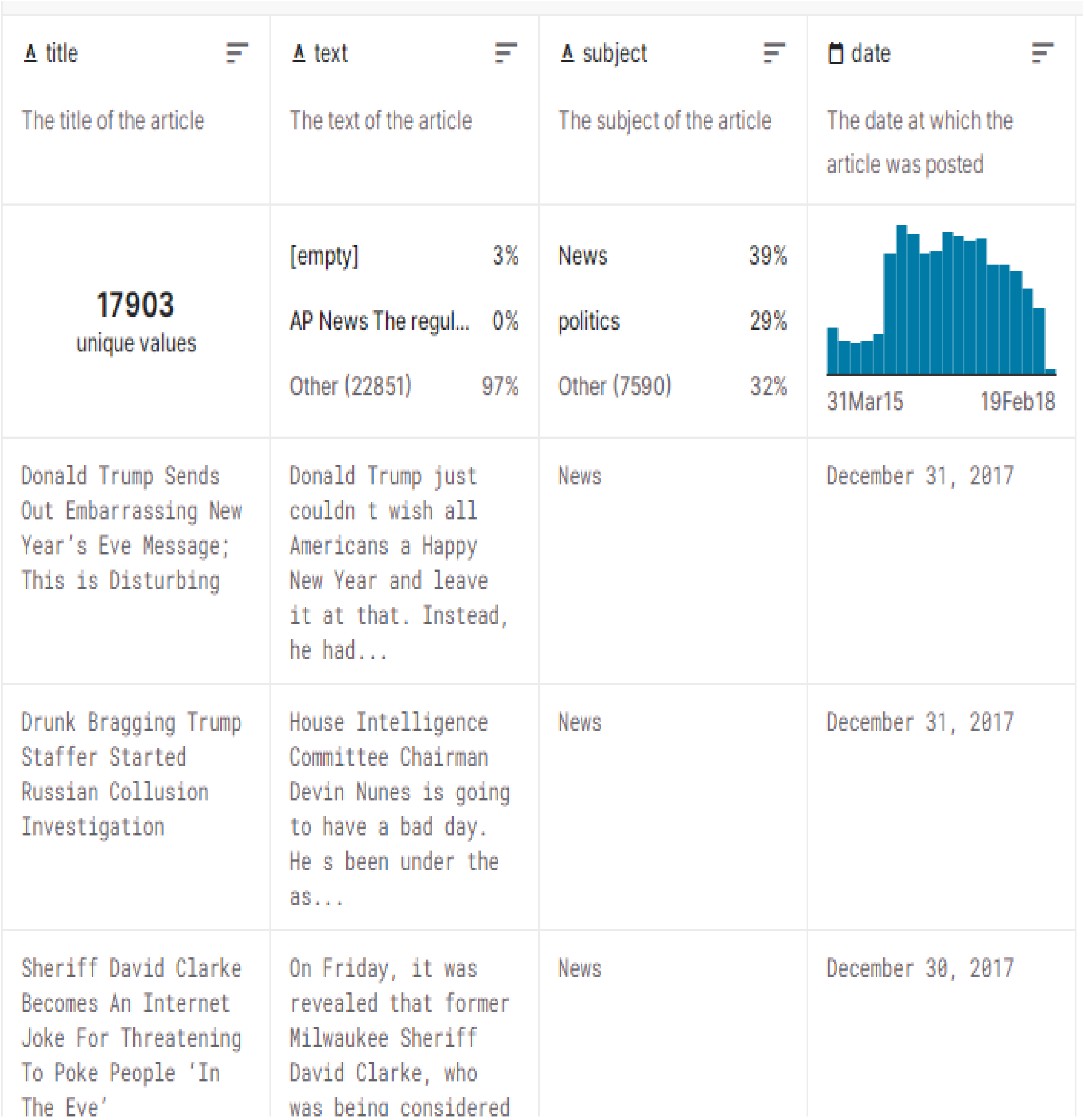
Dataset Link: https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset

real.csv



title	text	subject	date
The title of the article	The text of the article	The subject of the article	The date that this article was posted at
20826 unique values	21192 unique values	politicsNews 53% worldnews 47%	13Jan16 31Dec17
As U.S. budget fight looms, Republicans flip their fiscal script	WASHINGTON (Reuters) - The head of a conservative Republican faction in the U.S. Congress, who voted...	politicsNews	December 31, 2017
U.S. military to accept transgender recruits on Monday: Pentagon	WASHINGTON (Reuters) - Transgender people will be allowed for the first time to enlist in the U.S. m...	politicsNews	December 29, 2017
Senior U.S. Republican senator: 'Let Mr. Mueller do his job'	WASHINGTON (Reuters) - The special counsel investigation of	politicsNews	December 31, 2017

fake.csv



title	text	subject	date
The title of the article	The text of the article	The subject of the article	The date at which the article was posted
17903 unique values	[empty] 3% AP News The regul... 0% Other (22851) 97%	News 39% politics 29% Other (7590) 32%	31Mar15 19Feb18
Donald Trump Sends Out Embarrassing New Year's Eve Message; This is Disturbing	Donald Trump just couldn t wish all Americans a Happy New Year and leave it at that. Instead, he had...	News	December 31, 2017
Drunk Bragging Trump Staffer Started Russian Collusion Investigation	House Intelligence Committee Chairman Devin Nunes is going to have a bad day. He s been under the as...	News	December 31, 2017
Sheriff David Clarke Becomes An Internet Joke For Threatening To Poke People 'In The Eve'	On Friday, it was revealed that former Milwaukee Sheriff David Clarke, who was being considered	News	December 30, 2017

DATA DESCRIPTION:

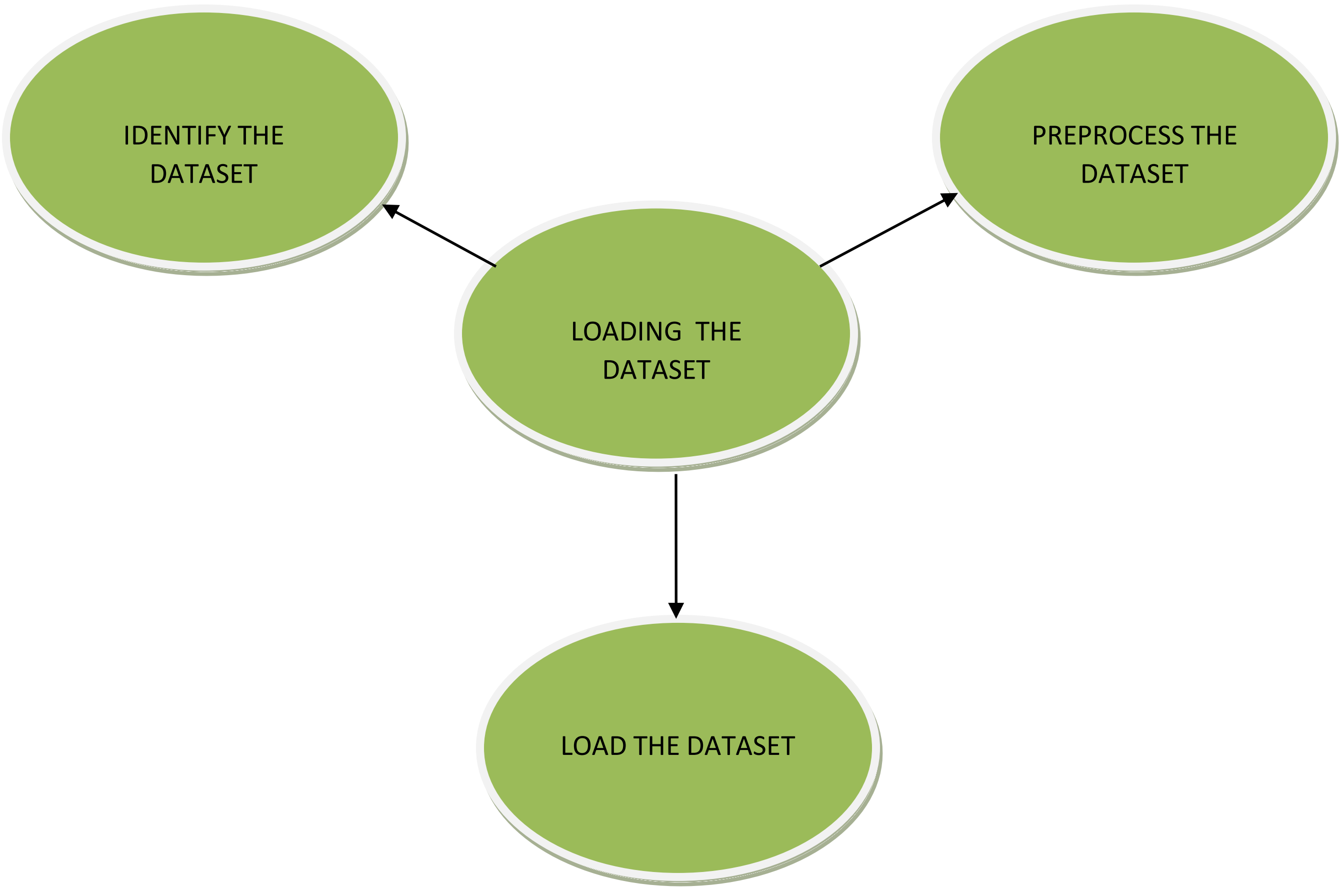
Title: the title of a news article

Text: the text of the article; could be incomplete

Subject: display the field of the news

Date: publish date

NECESSARY STEPS TO FOLLOW



IMPORTING LIBRARIES

Before begin working with the dataset,import essential Python libraries,such as pandas for data manipulation,nltkor spaCy for text preprocessing, and sklearn for machine learning.

PROGRAM

```
importnumpy as np
import pandas as Pd
importmatplotlib.pyplotasplt
importseabornassns
fromsklearn.feature_extraction.text import TfidfVectorizer
fromsklearn.model_selection import train_test_split
fromsklearn.metrics import accuracy_score
fromsklearn.metrics import classification_report
fromsklearn.metrics
importconfusion_matrix,ConfusionMatrixDisplay
import re
import string
```

SPLIT THE DATA

Make separate training and testing sets from your dataset. This aids in the performance evaluation of your model afterwards.

PROGRAM

```
fromsklearn.model_selection import train_test_split
```

FEATURE SCALING

Apply feature scaling to normalize your data, ensuring that all features have similar scales. Standardization (scaling to mean=0 andstd=1) is a common choice.

PROGRAM

```
fromsklearn.feature_extraction.text import TfidfVectorizer
```

LOADING THE DATASET

Using a library like pandas to load the chosen dataset into a DataFrame . This allows us to easily manipulate and analyze the data.

PROGRAM

```
true_data=pd.read_csv("C:\\Users\\gokul\\OneDrive\\Desktop\\
machine_learning\\projects\\True.csv")
fake_data=pd.read_csv("C:\\Users\\gokul\\OneDrive\\Desktop\\
machine_learning\\projects\\Fake.csv")
```

DATA PREPROCESSING

Preprocessing steps are essential in fake news detection using Natural Language Processing (NLP). These steps help clean and prepare the text data for analysis, making it easier for machine learning models to identify patterns and features that distinguish fake news from real news.

ADDING TARGET ATTRIBUTE TO DATASET

PROGRAM

```
true_data['class']=1
fake_data['class']=0
```

CONCATENATION OF TRUE AND FAKE DATASET

PROGRAM

```
data=pd.concat([true_data,fake_data],axis=0)
```

DATA EXPLORATION

checking for missing values, exploring the data's statistics, andvisualizing it to identify patterns.

PROGRAM

Removing unwanted columns

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

Removing NULL values:

```
data.isnull().sum()
```

RANDOM SHUFFLING THE DATAFRAME

To make sure that the data order does not induce biases throughout the model training phase, shuffle the dataset at random.

PROGRAM

```
data=data.sample(frac=1)
```

Tokenization and Text Preprocessing

Tokenize the text data and apply common preprocessing steps (e.g., lowercasing, stop word removal, stemming) to prepare the text for analysis. Tokenization involves breaking the text into individual words or tokens, which is crucial for NLP analysis.

PROGRAM

```
defwordopt(text):

text = text.lower()

text = re.sub('\[.*?\]', '', text)

text = re.sub("\\W"," ",text)

text = re.sub('https?://\S+|www\.\S+', '', text)

text = re.sub('<.*?>+', '', text)

text = re.sub('[%s]' % re.escape(string.punctuation), '', text)

text = re.sub('\n', '', text)

text = re.sub('\w*\d\w*', '', text)

return text

data['text']=data['text'].apply(wordopt)
```

SPLITTING OF DATA

```
x=data['text']
y=data['class']
```

TEXT TO VECTOR:

```
tfv=TfidfVectorizer()
x=tfv.fit_transform(x)
```

SPLITTING DATA TO TRAIN AND TEST DATA

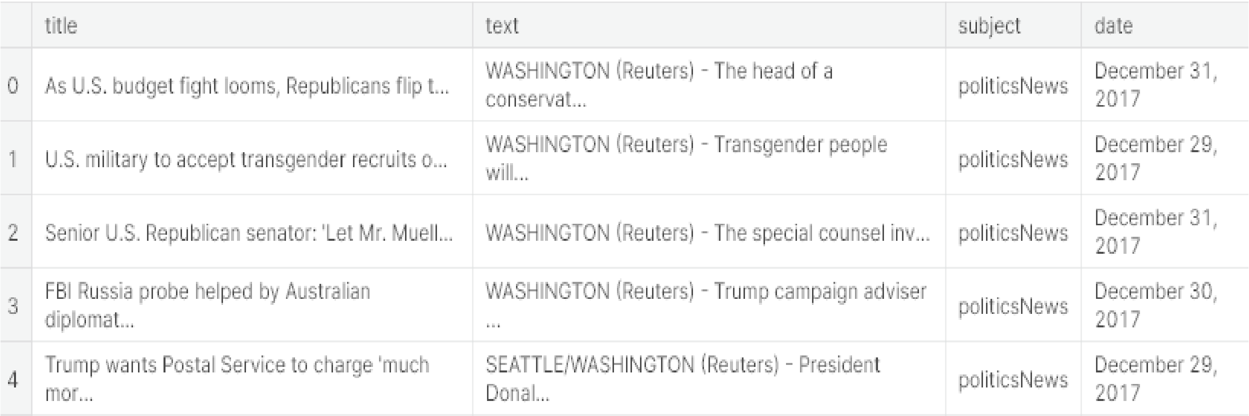
```
X_train,X_test,y_train,y_test=train_test_split(x,y,test_size=0.20)
```

PROGRAM

```
importnumpy as np
import pandas as pd
importmatplotlib.pyplot as plt
importseaborn as sns
fromsklearn.model_selection import train_test_split
fromsklearn.feature_extraction.text import TfidfVectorizer
fromsklearn.metrics import accuracy_score
fromsklearn.metrics import classification_report
fromsklearn.metrics import confusion_matrix,ConfusionMatrixDisplay
import re
import string
```

```
true_data=pd.read_csv("C:\\Users\\gokul\\OneDrive\\Desktop\\machine_learning\\projects\\True.csv")   #reading true news dataset

fake_data=pd.read_csv("C:\\Users\\gokul\\OneDrive\\Desktop\\machine_learning\\projects\\Fake.csv")  #reading fake news dataset

print(true_data.head())
```

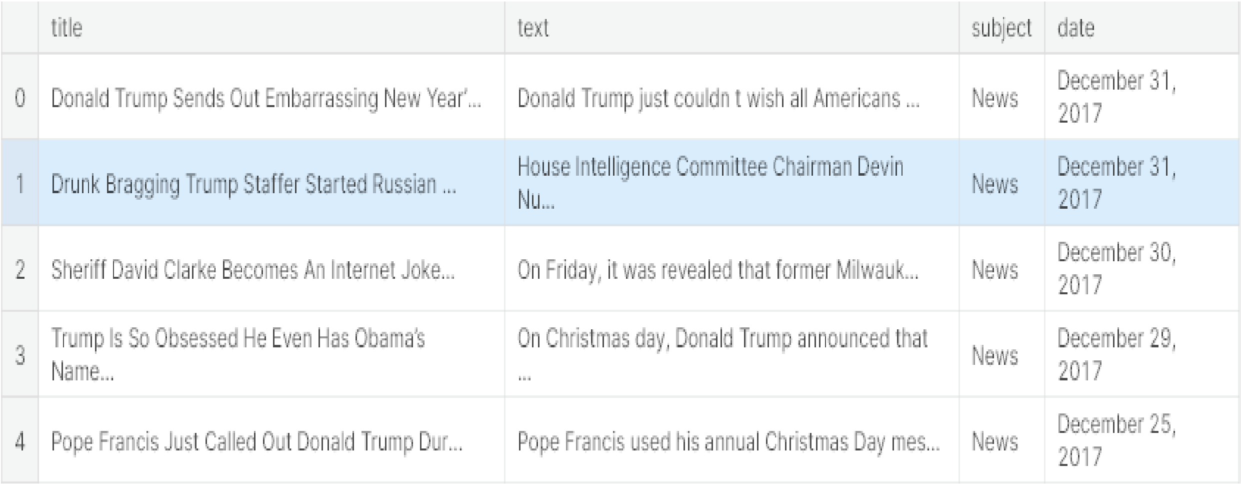
OUTPUT



	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

```
print(fake_data.head())
```

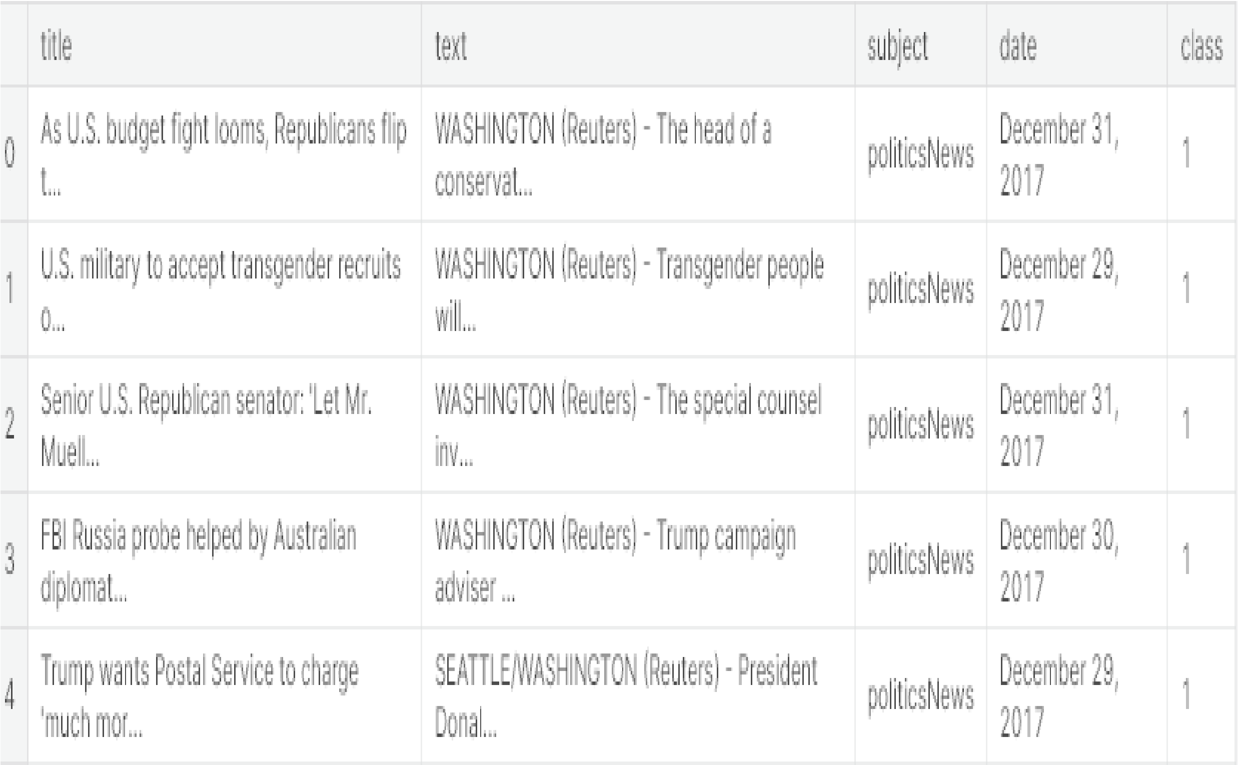
OUTPUT



	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

```
true_data['class']=1
fake_data['class']=0
print(true_data.head())
```

OUTPUT



	title	text	subject	date	class
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

```
print(true_data.shape , fake_data.shape)
```

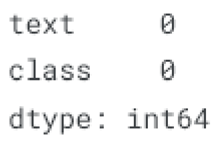
OUTPUT

```
 ((21417, 5), (23481, 5))
```

CONCATINATING FAKE AND REAL DATASET

```
data=pd.concat([true_data,fake_data],axis=0)
data.drop(['title','subject','date'],axis=1,inplace=True)
print(data.isnull().sum())
```

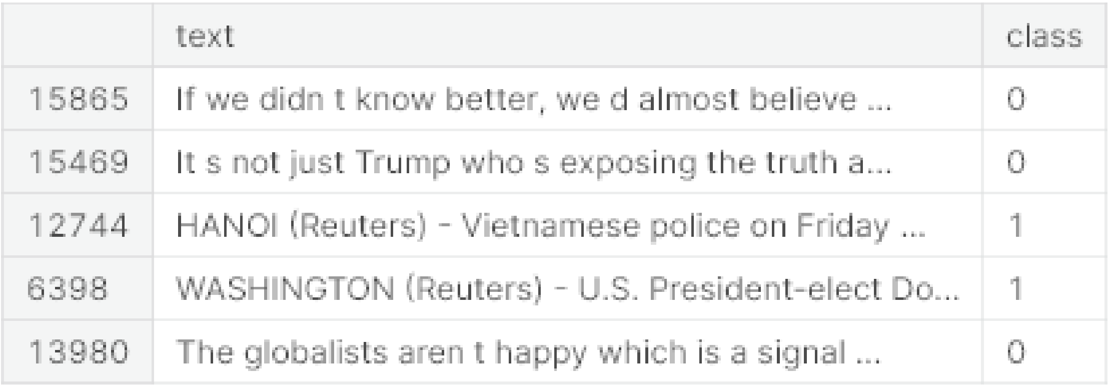
OUTPUT



```
text     0
class    0
dtype: int64
```

```
data=data.sample(frac=1)
Print(data.head())
```

OUTPUT



	text	class
15865	If we didn t know better, we d almost believe ...	0
15469	It s not just Trump who s exposing the truth a...	0
12744	HANOI (Reuters) - Vietnamese police on Friday ...	1
6398	WASHINGTON (Reuters) - U.S. President-elect Do...	1
13980	The globalists aren t happy which is a signal ...	0

```
defwordopt(text):
text=text.lower()
text=re.sub('\[.*?\]', '', text)
text=re.sub("\\W"," ",text)
text=re.sub('https?://\S+|www\.\S+', '', text)
```

```
text=re.sub('<.*?>+', '', text)
text=re.sub('[%s]'%re.escape(string.punctuation), '', text)
text=re.sub('\n', '', text)
text=re.sub('\w*\d\w*', '', text)
returntext
data['text']=data['text'].apply(wordopt)
x=data['text']
y=data['class']
tfv=TfidfVectorizer()
x=tfv.fit_transform(x)
X_train,X_test,y_train,y_test=train_test_split(x,y,test_size=0.20)
fromsklearn.linear_model import LogisticRegression
lr_model=LogisticRegression()
lr_model.fit(X_train,y_train)
```

OUTPUT

```
LogisticRegression()
```

```
y_pred_lr=lr_model.predict(X_test)
y_pred_lr
```

OUTPUT

```
array([1, 1, 0, ..., 1, 0, 0], dtype=int64)
```

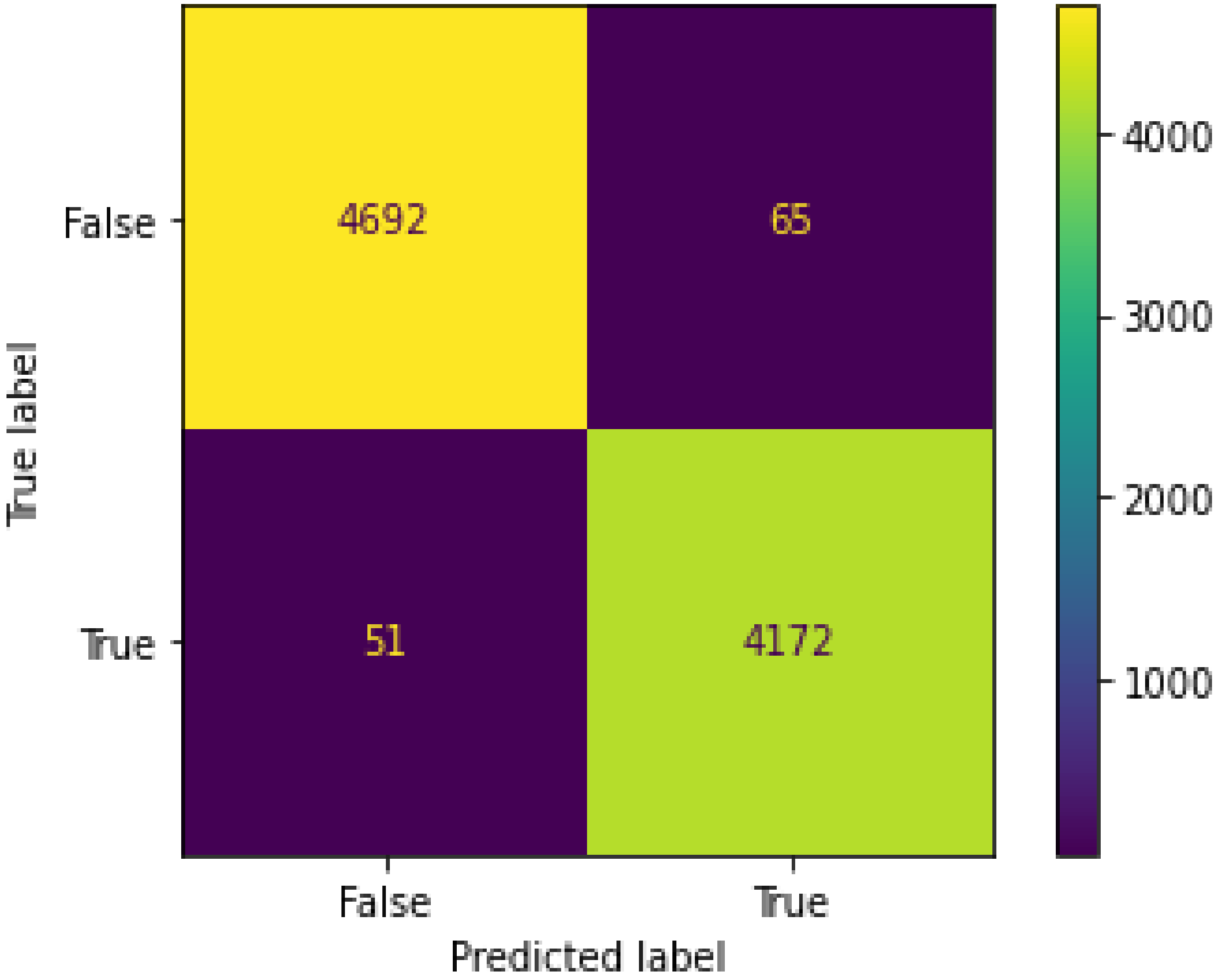
```
accuracy_score(y_pred_lr,y_test)
```

OUTPUT

```
0.9870824053452116
```

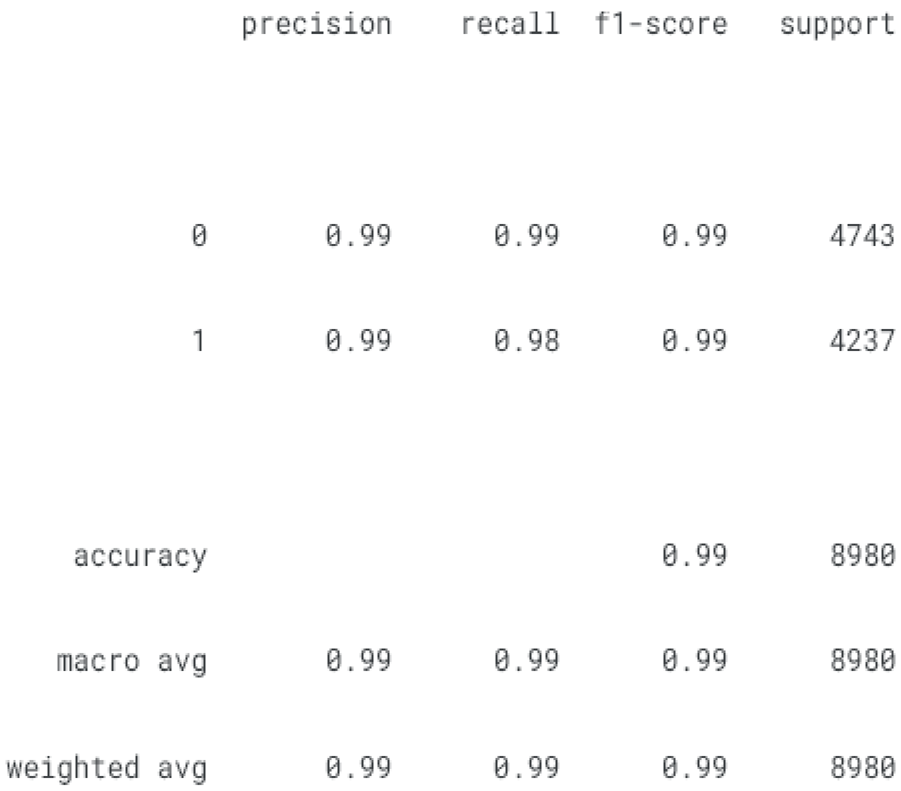
```
cm = confusion_matrix(y_test, y_pred_lr)
cm_display = ConfusionMatrixDisplay(confusion_matrix=cm,
display_labels=[False, True])
cm_display.plot()
plt.show()
```

OUTPUT



print(classification_report(y_pred_lr,y_test))

OUTPUT



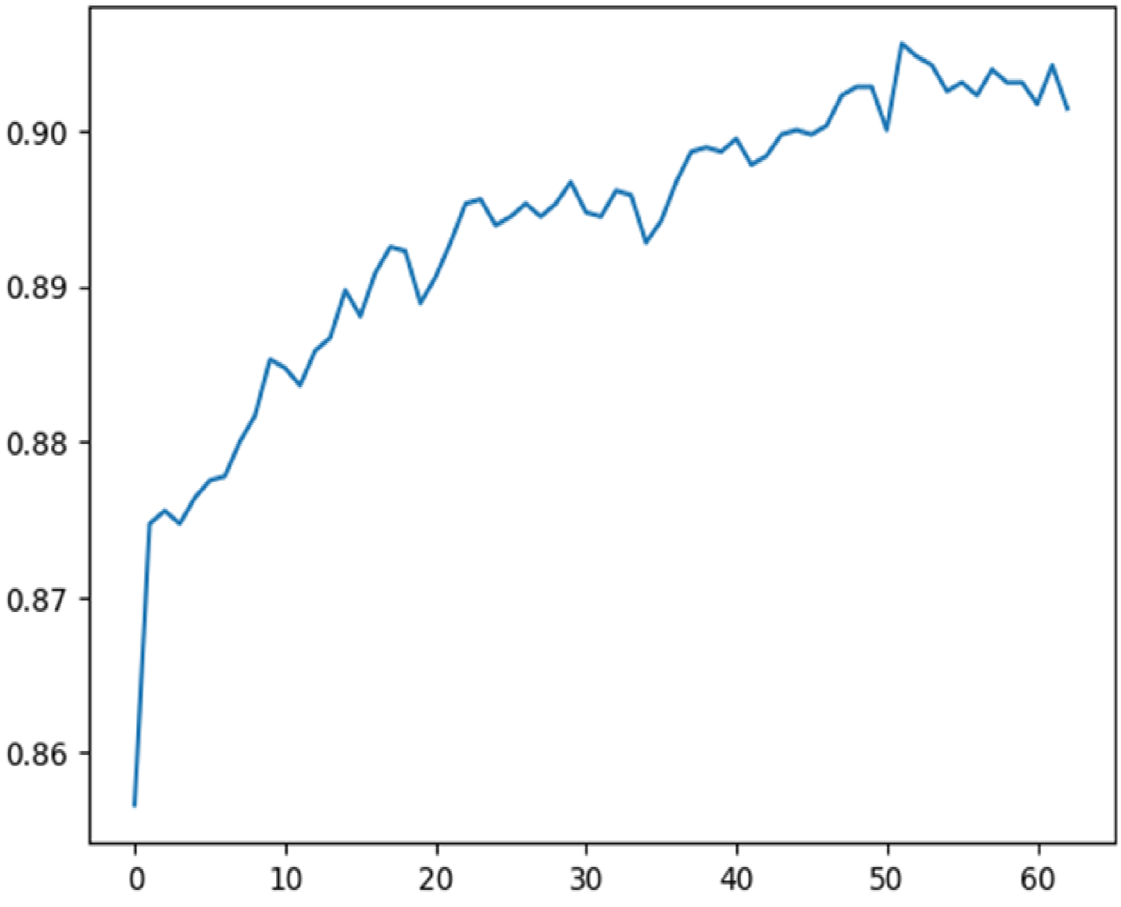
```
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      4743
           1       0.99      0.98      0.99      4237

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

DATA VISUALIZATION

```
import matplotlib.pyplot as plt
plt.plot(clf.validation_scores_)
```

OUTPUT

```
[<matplotlib.lines.Line2D at 0x7aa26f0901f0>]
```



CONCLUSION:

Fake news detection using NLP is a pivotal application of technology in safeguarding the reliability of information in the digital era. By leveraging natural language processing and machine learning, we can develop effective tools to distinguish between genuine and deceptive content.