FAKE NEWS CLASSIFICATION

Dissertation submitted in fulfilment of the requirements for the Degree of

B. Tech Computer Science and Engineering Data Science (AI and ML)

By

B Naga Padma Maanasa

12016660

Supervisor

Ved Prakash Chaubey

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

Month…………… Year ………

DECLARATION STATEMENT

I hereby declare that the research work reported in the dissertation/dissertation proposal entitled "FAKE NEWS CLASSIFICATION” in partial fulfilment of the requirement for the award of Degree for Master of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor Mr. Ved Prakash Chaubey. I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University’s Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

Signature of Candidate

B Naga Padma Maanasa

RK20CHA05.

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the B-Tech Dissertation/dissertation proposal entitled "**FAKE NEWS CLASSIFICATION**", submitted by **B Naga Padma Maanasa** at **Lovely Professional University, Phagwara, India** is a bonafide record of her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Signature of Supervisor

(Name of Supervisor)

Date:

Counter Signed by:

1) **Concerned HOD:**
   HoD's Signature: ________________

   HoD Name: ____________________

   Date: ___________________

2) **Neutral Examiners:**

   External Examiner

   Signature: _______________

   Name: __________________

   Affiliation: ______________

   Date: ___________________

   Internal Examiner

   Signature: _______________

   Name: __________________

   Date: ___________________

TABLE OF CONTENTS

Data Set Description.

Fake news classification refers to the process of identifying and categorizing news articles or other types of media that contain intentionally misleading, false, or fabricated information. This dataset contains 2 datasets fake dataset and real dataset. The fake dataset contains 4 columns and 23,481 rows and the real dataset contains 4 columns and 21,417 rows. The primary dataset of the both real and fake includes: Title of the news, Matter of the news, Subject of the news and the news release date. This dataset was created in 2020 by the authors Ahmed H, Traore I, Saad S.

Description of Columns

- **Title -** Title of the news given.
- **Text -** Content of the news
- **Subject-** Type of the news
- **Date-** The release date of the news

Objective of the Project.

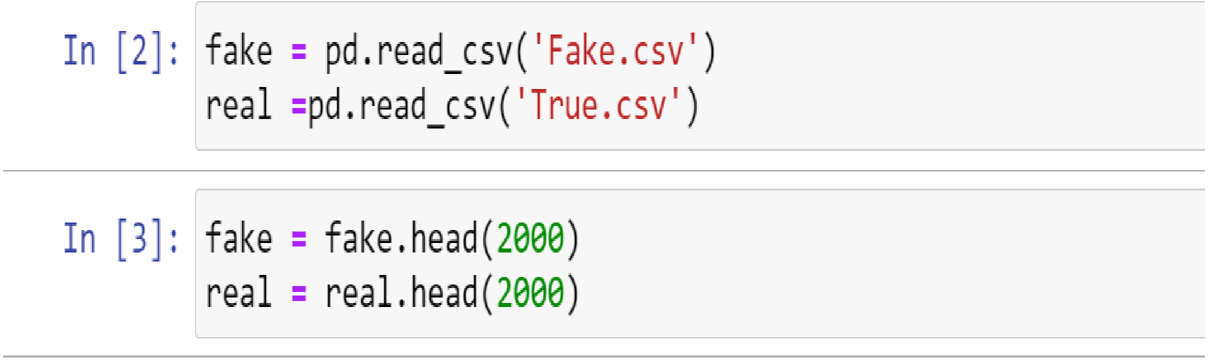
The main end goal of this project is to achieve maximum success rate while predicting whether the given news is fake or real. This includes the process of combining the both real and fake news datasets to form a single dataset and training it. Here in this dataset logistic regression is used to predict the accurate outcome of the dataset.

Why I chose this Project and Health Care Field?

The main objective of taking up this project is with the increasing prevalence of social media and online news sources fake news has become a serious problem that can have far-reaching consequences. It can be used to spread misinformation, influence public opinion, and even manipulate political outcomes. Fake news can have serious consequences in various areas, such as politics, public health, and security. It can spread misinformation, undermine trust in the media and other information sources, propagate biased or discriminatory views, and even be used to manipulate public opinion or behaviour. By detecting and labelling fake news, we can reduce the impact of these negative effects and help people make more informed decisions based on accurate and trustworthy information.

Statistical Insights of the Dataset.

Reading the datasets

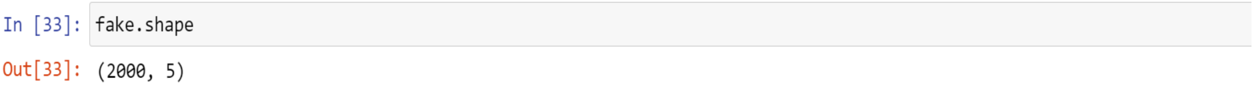


```
In [2]: fake = pd.read_csv('Fake.csv')
        real =pd.read_csv('True.csv')

In [3]: fake = fake.head(2000)
        real = real.head(2000)
```

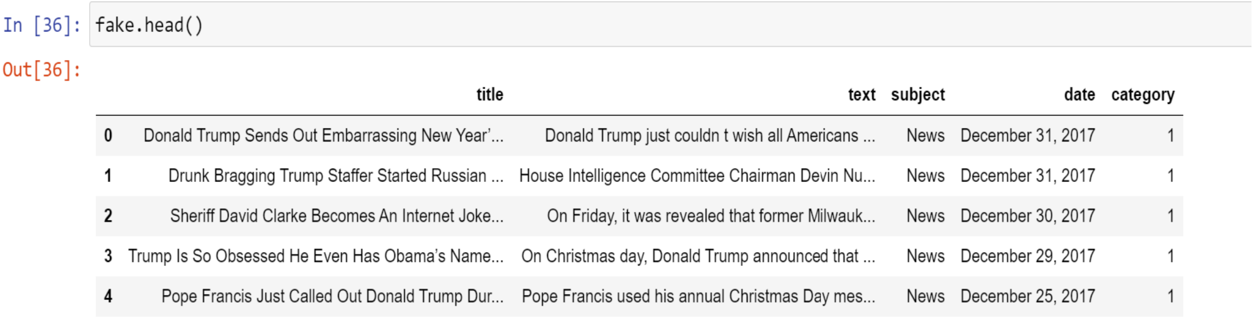
- Here after reading the data sets as original dataset contains 21+ thousand rows, I have only taken 2000 rows to make the compilation process much eas

Cleaning and Pre-processing

- Fake news:

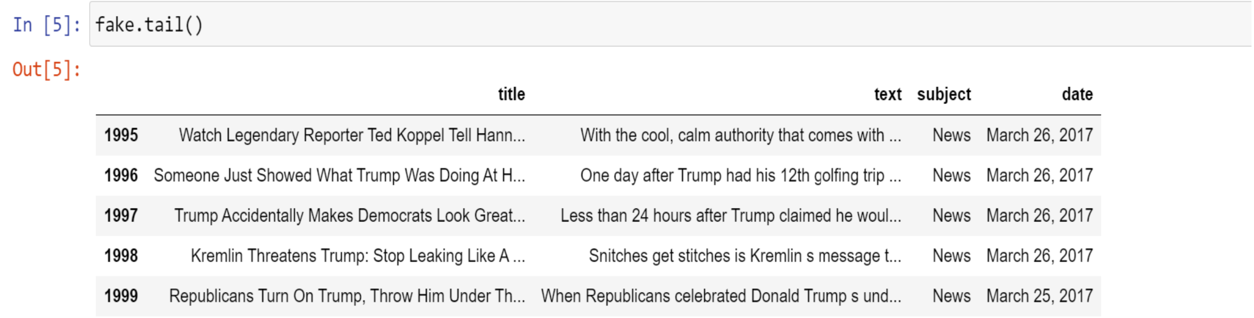


```
In [33]: fake.shape
Out[33]: (2000, 5)
```



```
In [36]: fake.head()
Out[36]:
```

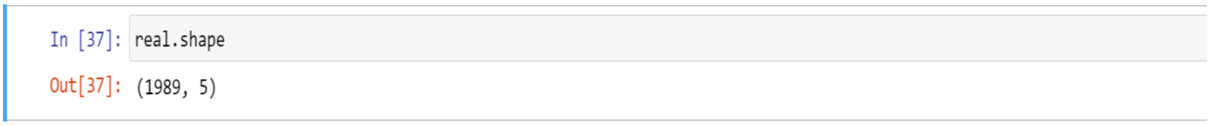
	title	text	subject	date	category
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	1
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	1
2	Sheriff David 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



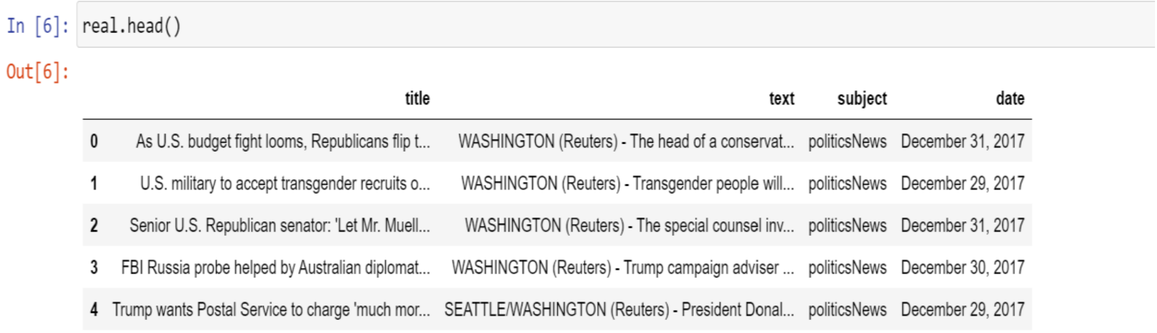
```
In [5]: fake.tail()
Out[5]:
```

	title	text	subject	date
1995	Watch Legendary Reporter Ted Koppel Tell Hann...	With the cool, calm authority that comes with ...	News	March 26, 2017
1996	Someone Just Showed What Trump Was Doing At H...	One day after Trump had his 12th golfing trip ...	News	March 26, 2017
1997	Trump Accidentally Makes Democrats Look Great...	Less than 24 hours after Trump claimed he woul...	News	March 26, 2017
1998	Kremlin Threatens Trump: Stop Leaking Like A ...	Snitches get stitches is Kremlin s message t...	News	March 26, 2017
1999	Republicans Turn On Trump, Throw Him Under Th...	When Republicans celebrated Donald Trump s und...	News	March 25, 2017

- True news:

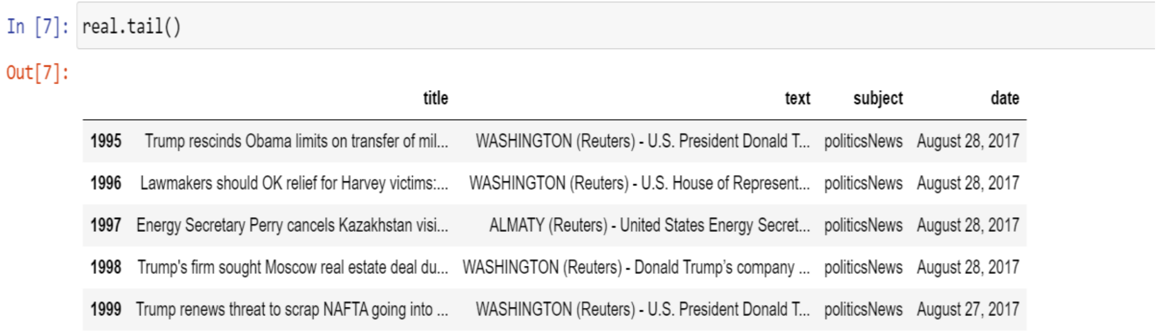


```
In [37]: real.shape
Out[37]: (1989, 5)
```



```
In [6]: real.head()
Out[6]:
```

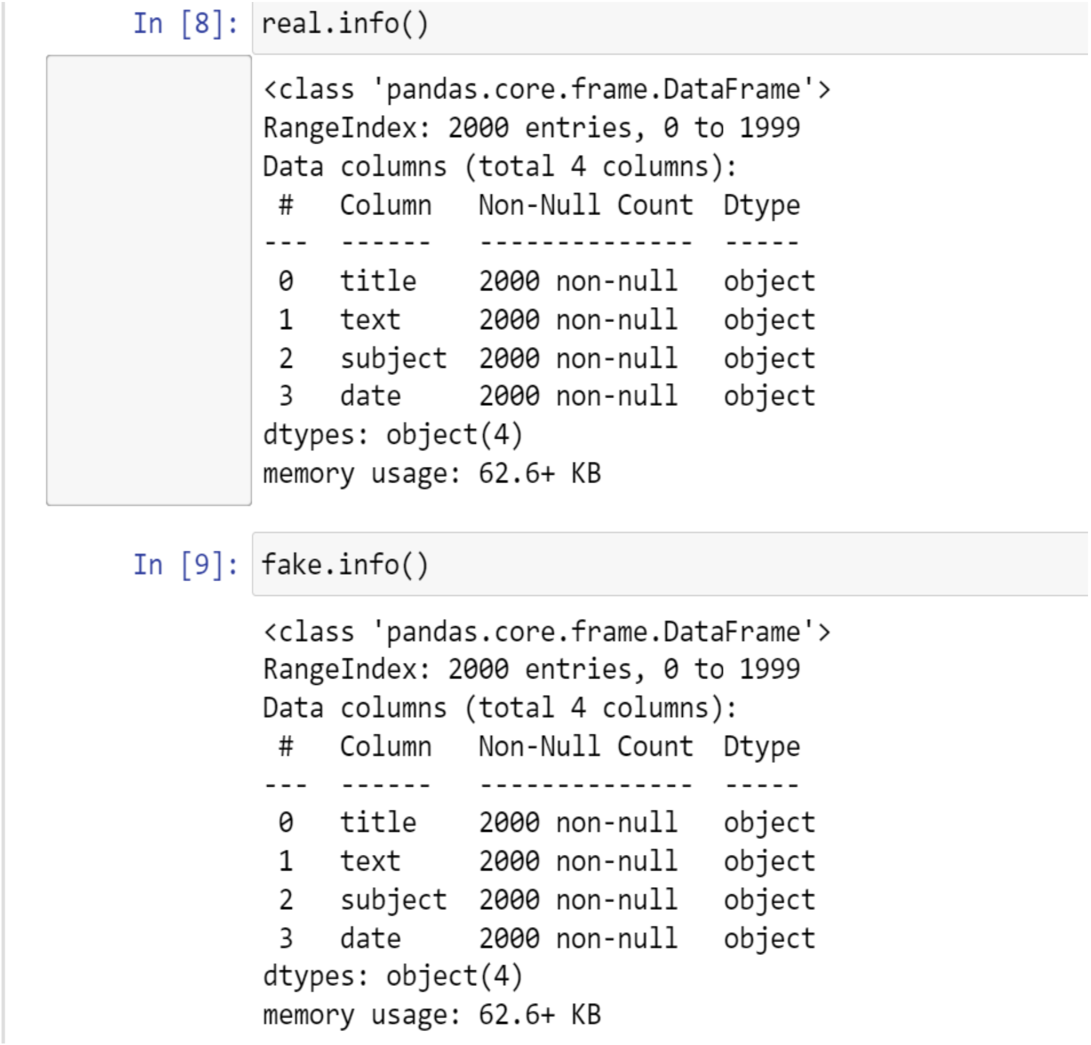
	title	text	subject	date
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017
2	Senior U.S. 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



```
In [7]: real.tail()
Out[7]:
```

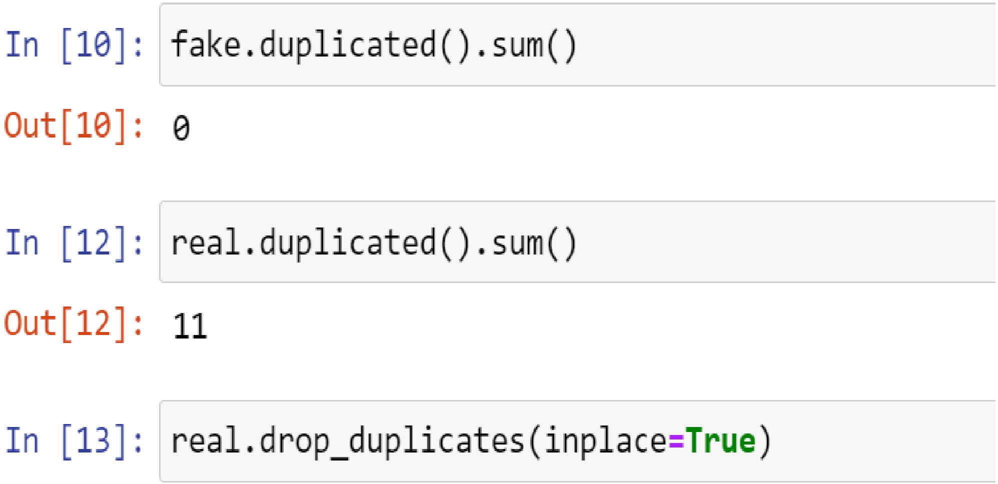
	title	text	subject	date
1995	Trump rescinds Obama limits on transfer of mil...	WASHINGTON (Reuters) - U.S. President Donald T...	politicsNews	August 28, 2017
1996	Lawmakers should OK relief for Harvey victims:...	WASHINGTON (Reuters) - U.S. House of Represent...	politicsNews	August 28, 2017
1997	Energy Secretary Perry cancels Kazakhstan visi...	ALMATY (Reuters) - United States Energy Secret...	politicsNews	August 28, 2017
1998	Trump's firm sought Moscow real estate deal du...	WASHINGTON (Reuters) - Donald Trump's company ...	politicsNews	August 28, 2017
1999	Trump renews threat to scrap NAFTA going into ...	WASHINGTON (Reuters) - U.S. President Donald T...	politicsNews	August 27, 2017

- Information regarding both the news



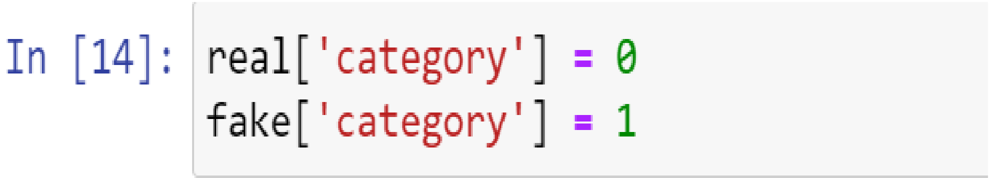
```
In [8]: real.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 4 columns):
 #   Column   Non-Null Count  Dtype 
---  ------   --------------  ----- 
 0   title    2000 non-null   object
 1   text     2000 non-null   object
 2   subject  2000 non-null   object
 3   date     2000 non-null   object
dtypes: object(4)
memory usage: 62.6+ KB

In [9]: fake.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 4 columns):
 #   Column   Non-Null Count  Dtype 
---  ------   --------------  ----- 
 0   title    2000 non-null   object
 1   text     2000 non-null   object
 2   subject  2000 non-null   object
 3   date     2000 non-null   object
dtypes: object(4)
memory usage: 62.6+ KB
```

- Checking duplicates and dropping them



```
In [10]: fake.duplicated().sum()

Out[10]: 0

In [12]: real.duplicated().sum()

Out[12]: 11

In [13]: real.drop_duplicates(inplace=True)
```

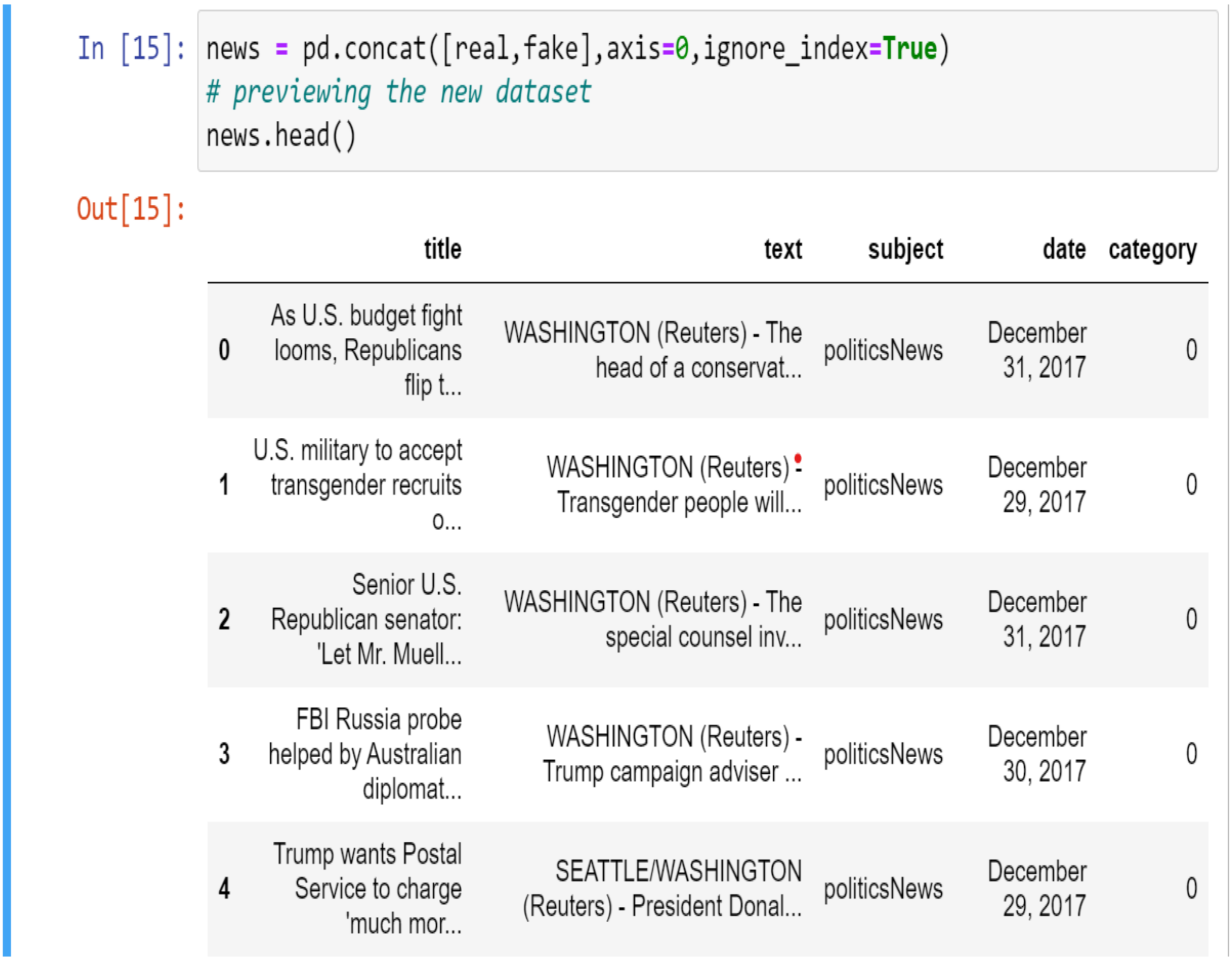
- Creating a new column “Category” in both of the datasets



```
In [14]: real['category'] = 0
         fake['category'] = 1
```

➢ Category helps us to understand weather the news is real or fake where real is represented with 0 and fake is 1.

- Creating a new data set combining the real and fake datasets

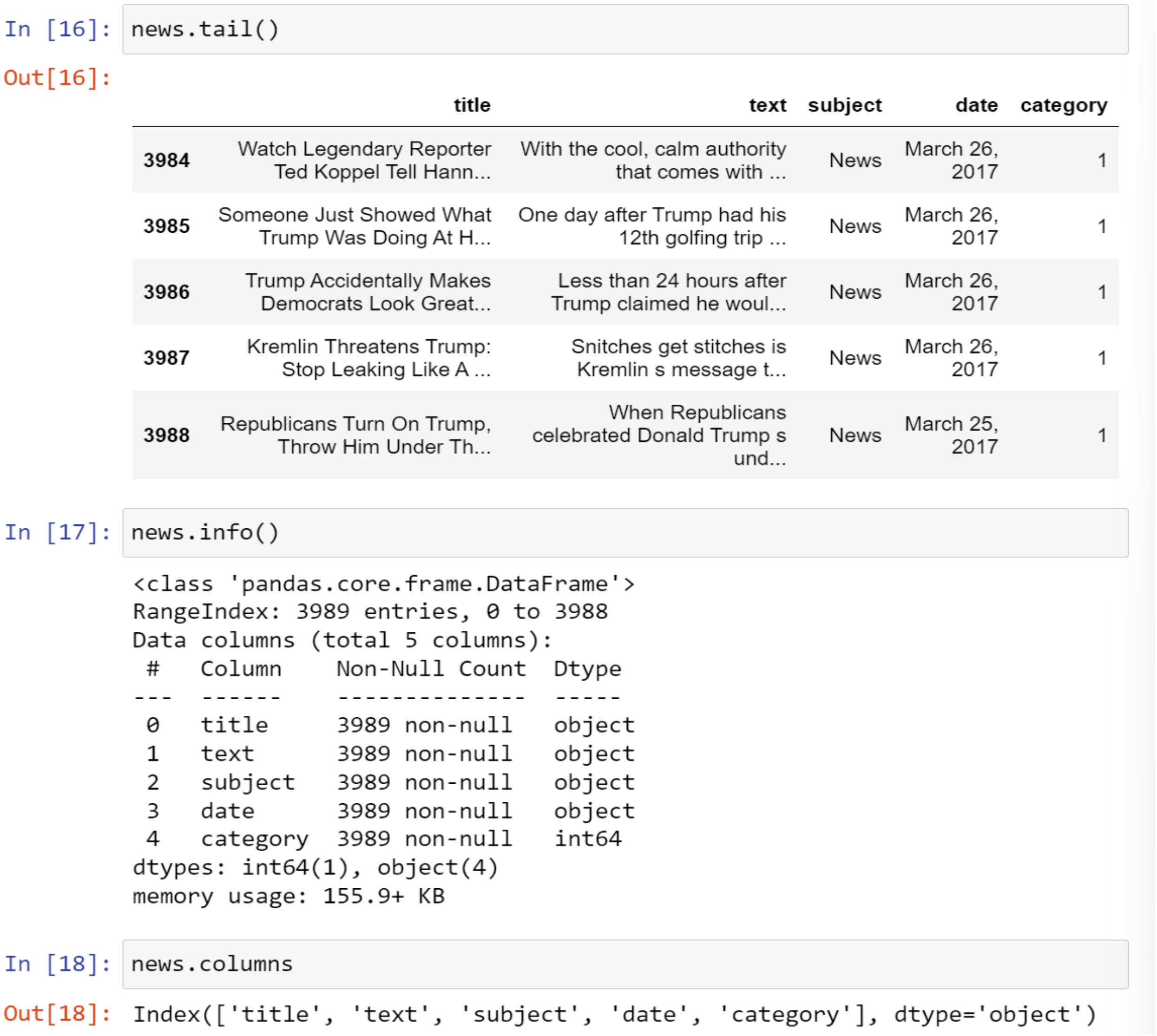


```
In [15]: news = pd.concat([real,fake],axis=0,ignore_index=True)
         # previewing the new dataset
         news.head()
```

Out[15]:

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

- Tail, Information and columns of the news:

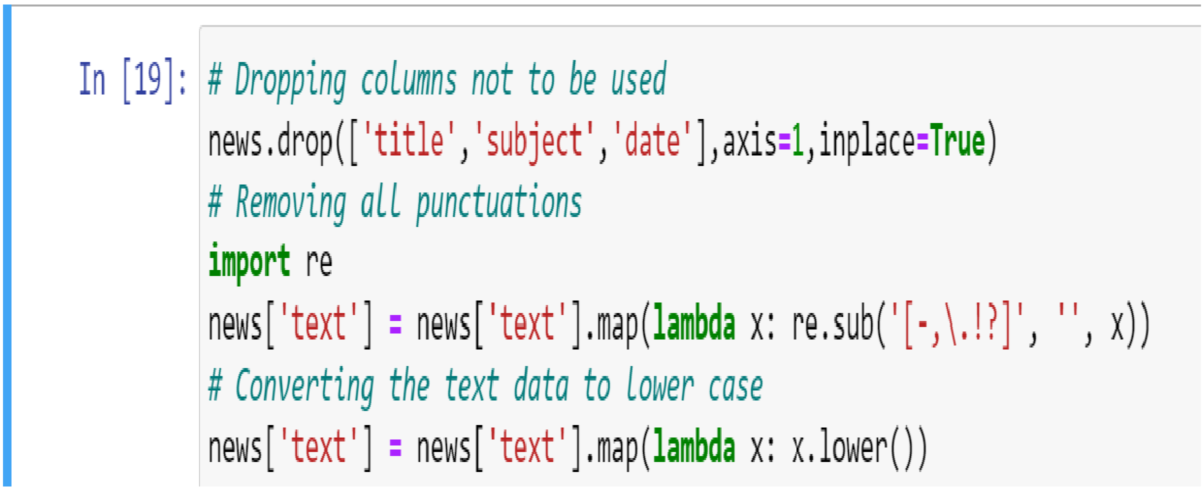


```
In [16]: news.tail()
Out[16]:
```

	title	text	subject	date	category
3984	Watch Legendary Reporter Ted Koppel Tell Hann...	With the cool, calm authority that comes with ...	News	March 26, 2017	1
3985	Someone Just Showed What Trump Was Doing At H...	One day after Trump had his 12th golfing trip ...	News	March 26, 2017	1
3986	Trump Accidentally Makes Democrats Look Great...	Less than 24 hours after Trump claimed he woul...	News	March 26, 2017	1
3987	Kremlin Threatens Trump: Stop Leaking Like A ...	Snitches get stitches is Kremlin s message t...	News	March 26, 2017	1
3988	Republicans Turn On Trump, Throw Him Under Th...	When Republicans celebrated Donald Trump s und...	News	March 25, 2017	1

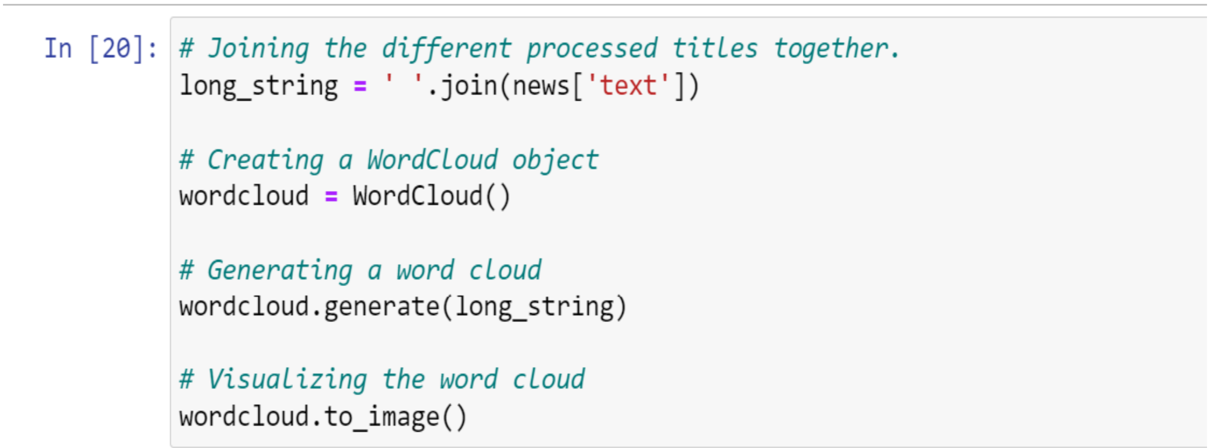
```
In [17]: news.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3989 entries, 0 to 3988
Data columns (total 5 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   title     3989 non-null   object
 1   text      3989 non-null   object
 2   subject   3989 non-null   object
 3   date      3989 non-null   object
 4   category  3989 non-null   int64 
dtypes: int64(1), object(4)
memory usage: 155.9+ KB

In [18]: news.columns

Out[18]: Index(['title', 'text', 'subject', 'date', 'category'], dtype='object')
```

- Dropping the columns not to be used

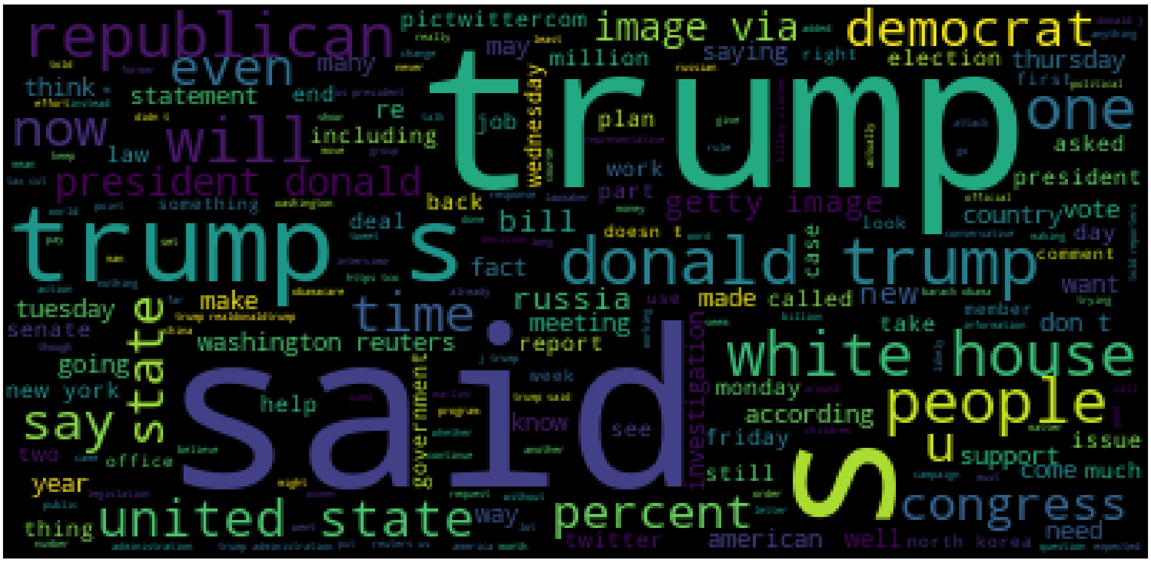


```
In [19]: # Dropping columns not to be used
         news.drop(['title','subject','date'],axis=1,inplace=True)
         # Removing all punctuations
         import re
         news['text'] = news['text'].map(lambda x: re.sub('[-,\.!?]', '', x))
         # Converting the text data to lower case
         news['text'] = news['text'].map(lambda x: x.lower())
```

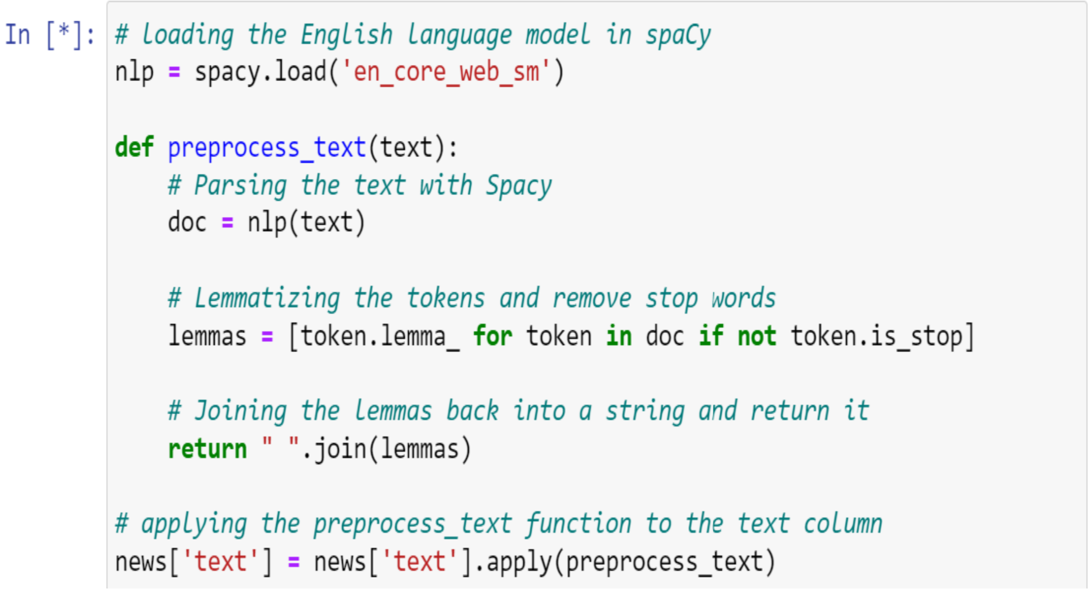
- Creating a Word Cloud to understand about the data:



```
In [20]: # Joining the different processed titles together.
         long_string = ' '.join(news['text'])

         # Creating a WordCloud object
         wordcloud = WordCloud()

         # Generating a word cloud
         wordcloud.generate(long_string)

         # Visualizing the word cloud
         wordcloud.to_image()
```

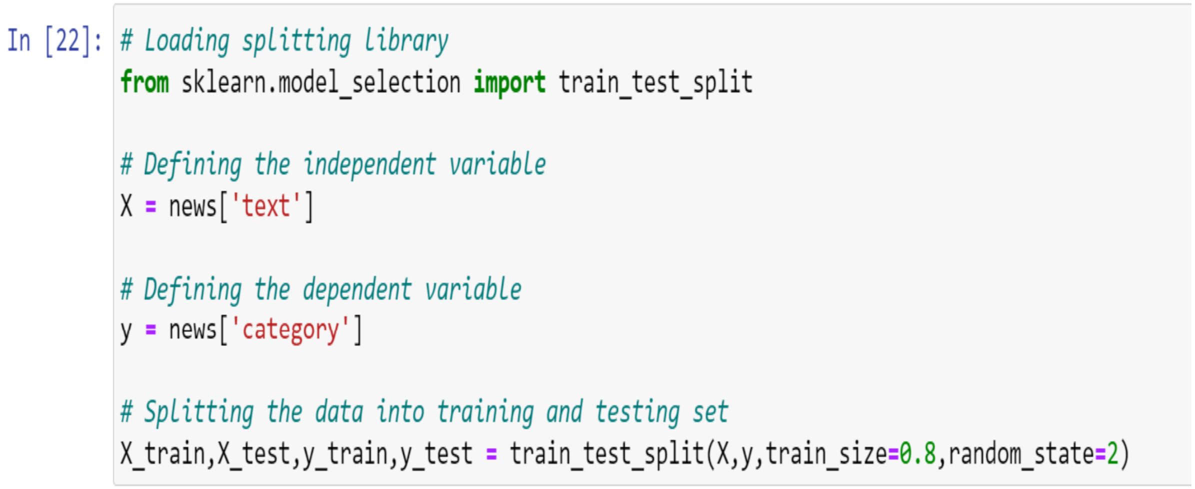


- Removing the Stop words



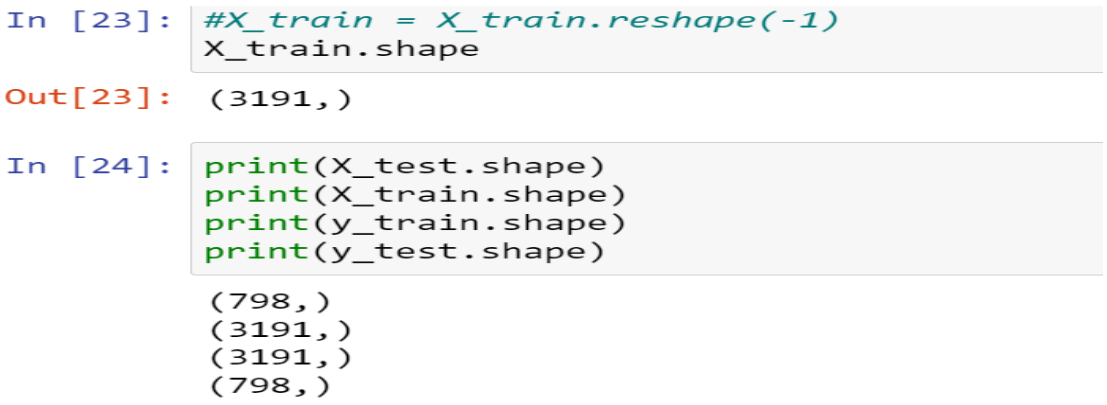
```
In [*]: # Loading the English language model in spaCy
        nlp = spacy.load('en_core_web_sm')

        def preprocess_text(text):
            # Parsing the text with Spacy
            doc = nlp(text)

            # Lemmatizing the tokens and remove stop words
            lemmas = [token.lemma_ for token in doc if not token.is_stop]

            # Joining the Lemmas back into a string and return it
            return " ".join(lemmas)

        # applying the preprocess_text function to the text column
        news['text'] = news['text'].apply(preprocess_text)
```

- Splitting the data into testing and training data.



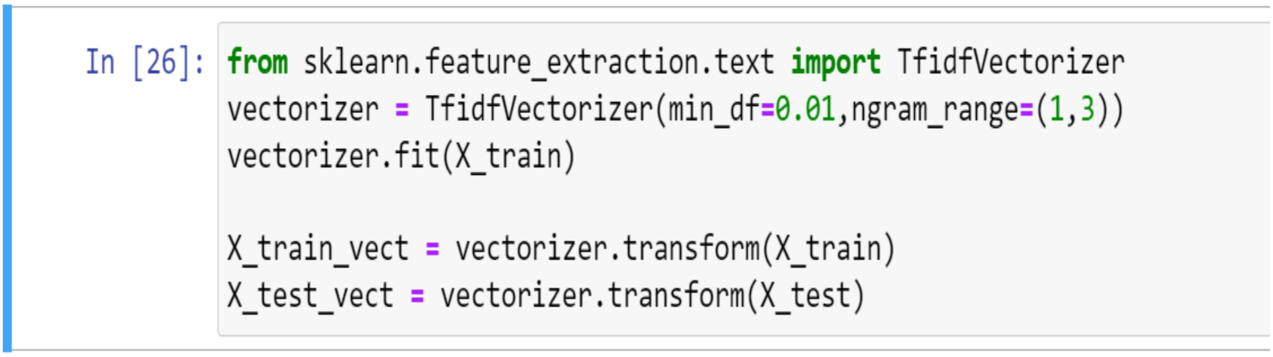
```
In [22]: # Loading splitting library
         from sklearn.model_selection import train_test_split

         # Defining the independent variable
         X = news['text']

         # Defining the dependent variable
         y = news['category']

         # Splitting the data into training and testing set
         X_train,X_test,y_train,y_test = train_test_split(X,y,train_size=0.8,random_state=2)
```

- Shape of the dataset:

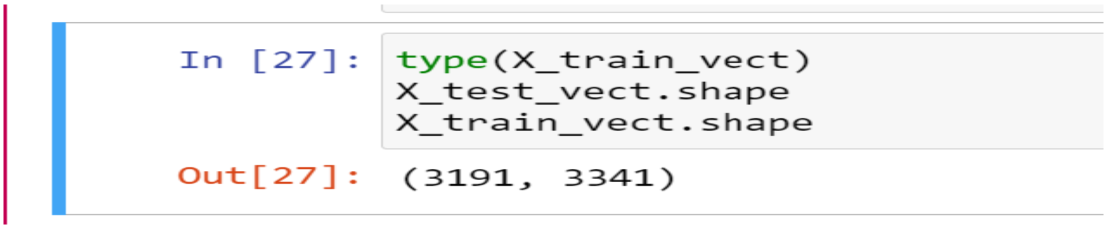


```
In [23]: #X_train = X_train.reshape(-1)
         X_train.shape

Out[23]: (3191,)

In [24]: print(X_test.shape)
         print(X_train.shape)
         print(y_train.shape)
         print(y_test.shape)

         (798,)
         (3191,)
         (3191,)
         (798,)
```

- Checking Term and Inverse Document Frequencies



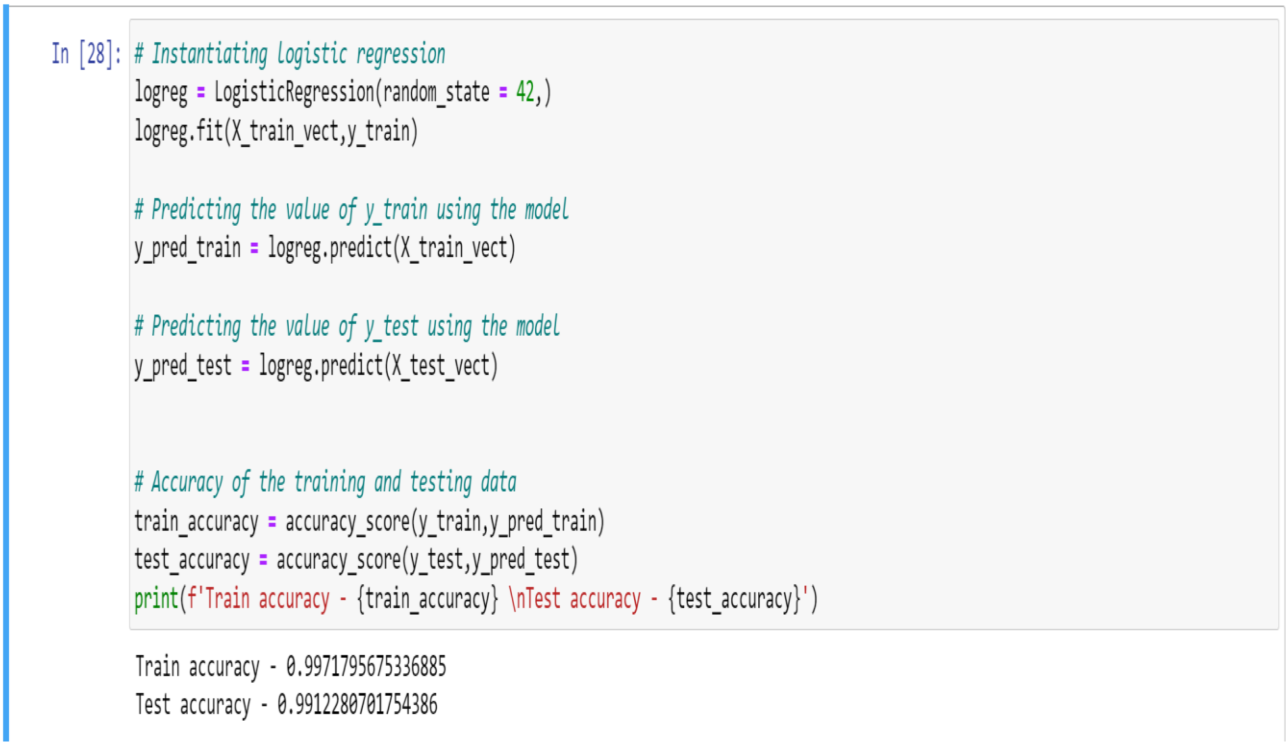
```
In [26]: from sklearn.feature_extraction.text import TfidfVectorizer
         vectorizer = TfidfVectorizer(min_df=0.01,ngram_range=(1,3))
         vectorizer.fit(X_train)

         X_train_vect = vectorizer.transform(X_train)
         X_test_vect = vectorizer.transform(X_test)
```



```
In [27]: type(X_train_vect)
         X_test_vect.shape
         X_train_vect.shape

Out[27]: (3191, 3341)
```

Modelling The Data.

- Using Logistic Regression

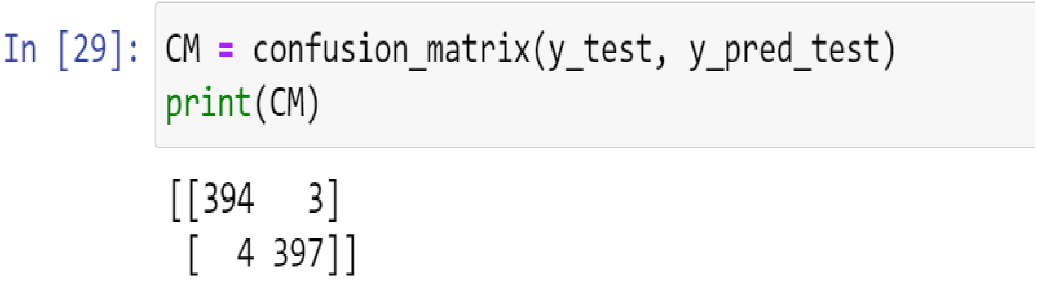


```
In [28]: # Instantiating logistic regression
         logreg = LogisticRegression(random_state = 42,)
         logreg.fit(X_train_vect,y_train)

         # Predicting the value of y_train using the model
         y_pred_train = logreg.predict(X_train_vect)

         # Predicting the value of y_test using the model
         y_pred_test = logreg.predict(X_test_vect)


         # Accuracy of the training and testing data
         train_accuracy = accuracy_score(y_train,y_pred_train)
         test_accuracy = accuracy_score(y_test,y_pred_test)
         print(f'Train accuracy - {train_accuracy} \nTest accuracy - {test_accuracy}')

         Train accuracy - 0.9971795675336885
         Test accuracy - 0.9912280701754386
```

- The training accuracy is: 0.9971
- The testing accuracy is: 0.9912

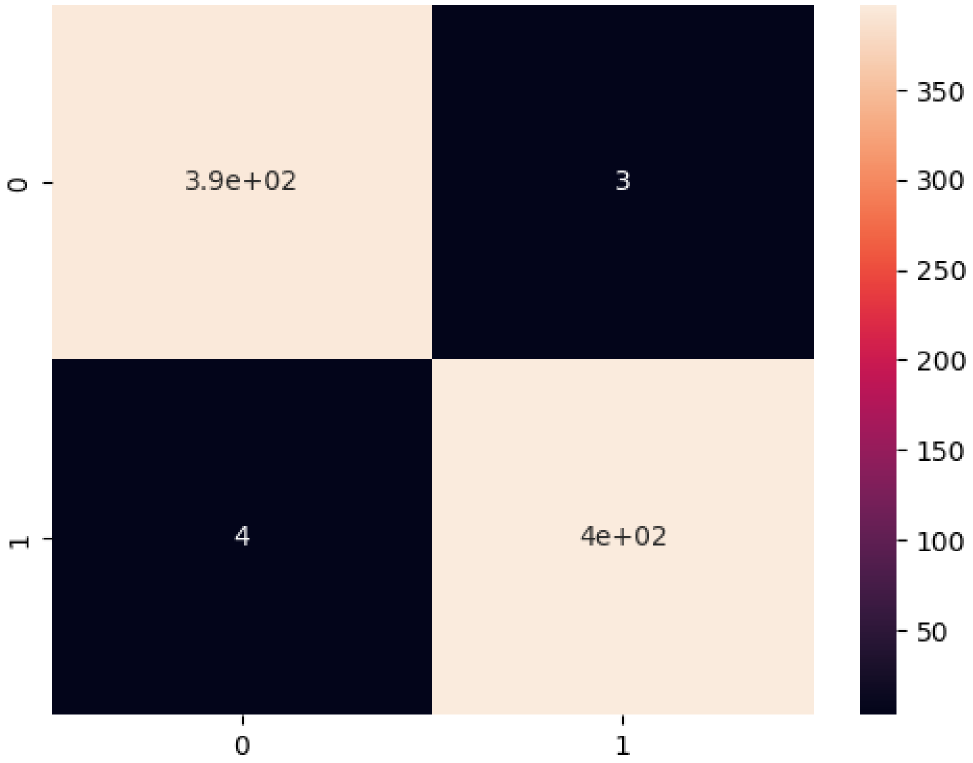
Model Evaluation.

- Confusion Matrix of Logistic Regression:

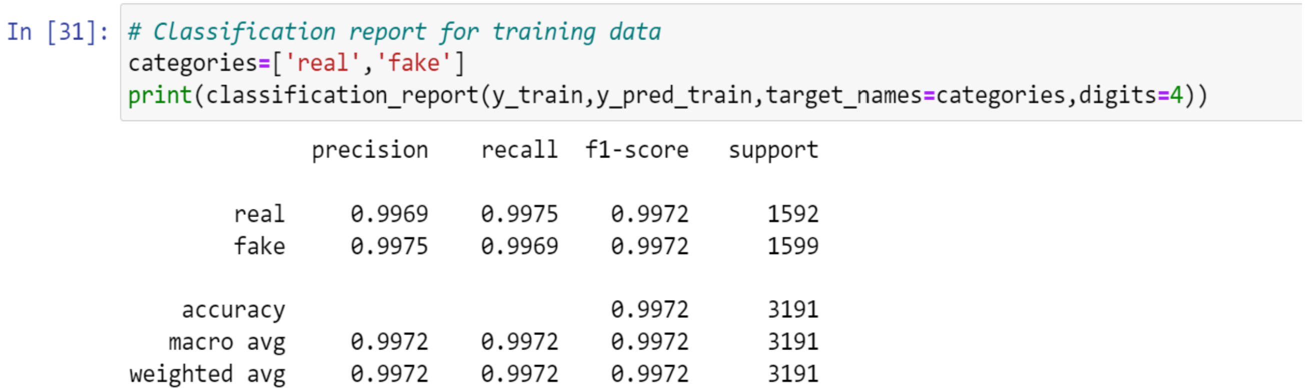


```
In [29]: CM = confusion_matrix(y_test, y_pred_test)
         print(CM)

         [[394   3]
          [  4 397]]
```

- Heat Map

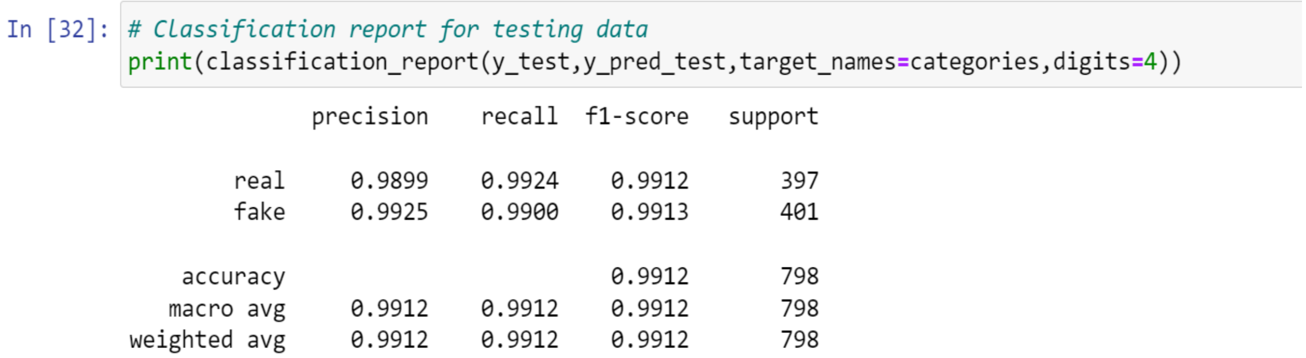


- Classification Report for both the training and testing data



```
In [31]: # Classification report for training data
         categories=['real','fake']
         print(classification_report(y_train,y_pred_train,target_names=categories,digits=4))
```

```
              precision    recall  f1-score   support

        real     0.9969    0.9975    0.9972      1592
        fake     0.9975    0.9969    0.9972      1599

    accuracy                         0.9972      3191
   macro avg     0.9972    0.9972    0.9972      3191
weighted avg     0.9972    0.9972    0.9972      3191
```



```
In [32]: # Classification report for testing data
         print(classification_report(y_test,y_pred_test,target_names=categories,digits=4))
```

```
              precision    recall  f1-score   support

        real     0.9899    0.9924    0.9912       397
        fake     0.9925    0.9900    0.9913       401

    accuracy                         0.9912       798
   macro avg     0.9912    0.9912    0.9912       798
weighted avg     0.9912    0.9912    0.9912       798
```

Conclusion.

In conclusion, the results of the model made using Logistic Regression algorithm has given outstanding training results of 99.71% and testing results of 99.12% so that I don't even have to take any algorithm for modeling.

Overall, the field of fake news classification is constantly evolving, as researchers and experts work to develop new and more effective methods for detecting and classifying fake news. By continuing to invest in this field, we can help combat the spread of fake news and promote a more informed and democratic society that values accuracy, transparency, and accountability.

References.

GitHub: https://github.com/B-Maanasa/Fake-News-Classification_

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

Udemy: https://www.udemy.com/course/machinelearning

Checklist for Dissertation-III Supervisor

Name: _______________________ UID: ________ Domain: _____
Registration No: ____________Name of student:__________________

Title of Dissertation:

____

□ Front pages are as per the format.

□ Topic on the PAC form and title page are same.

□ Front page numbers are in roman and for report, it is like 1, 2, 3…….

□ TOC, List of Figures, etc. are matching with the actual page numbers in the report.

□ Font, Font Size, Margins, line Spacing, Alignment, etc. are as per the guidelines.

□ Color prints are used for images and implementation snapshots.

□ Captions and citations are provided for all the figures, tables etc. and are numbered and center aligned.

□ All the equations used in the report are numbered.

□ Citations are provided for all the references.

□ **Objectives are clearly defined.**

□ Minimum total number of pages of report is 50.

□ Minimum references in report are 30.

Here by, I declare that I had verified the above mentioned points in the final dissertation report.

Signature of Supervisor with UID