FAKE NEWS DETECTION USING NLP

TEAM MEMBERS

513121106061

513121106094

513121106097

513121106312

513121106310

Phase 3 submission document

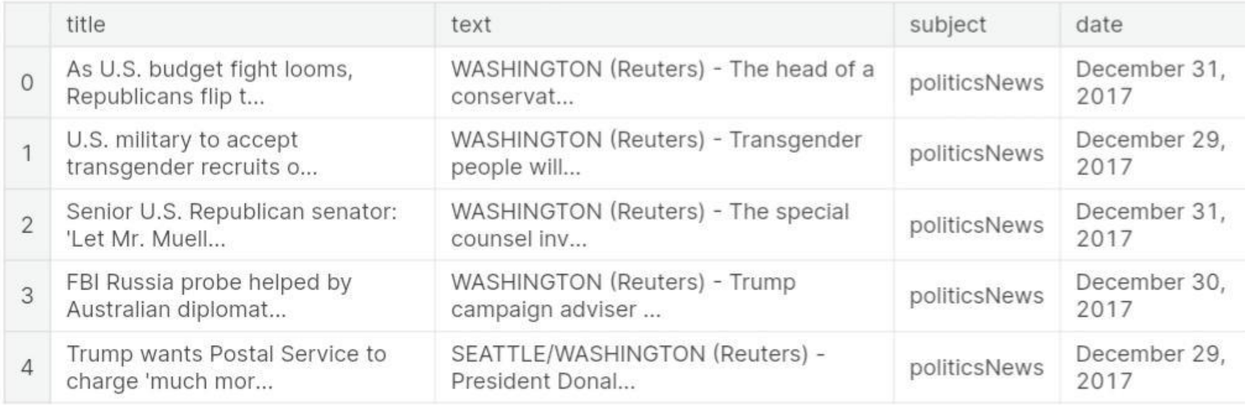
Phase 3: Development Part 1

Introduction:

- ✓ Detecting fake news using Natural Language Processing (NLP) is a critical application that leverages machine learning and linguistic analysis to identify misleading or fabricated information in text.

- ✓ Effective preprocessing can improve the quality of the dataset and the performance of the fake news detection model.

Given data set:



	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

Necessary steps to follow:

1.Import libraries:

Start by importing the necessary libraries:

Program:

```
Import pandas as pd

Import numpy as

npFrom sklearn.model_selection import train_test_split

From sklearn.preprocessing import StandardScaler
```

2.Load the Dataset:

Load your dataset into a Pandas DataFrame. You can typically find fake News datasets in CSV format, but you can adapt this code to other Formats as needed.

Program:

```
Real = pd.read_csv("/kaggle/input/fake-and-real-news-dataset/True.csv")

Fake = pd.read_csv("/kaggle/input/fake-and-real-news-dataset/Fake.csv")
```

```
Print(“real: “,real.shape,”\nfake: “,fake.shape)
```

3. Exploratory Data Analysis (EDA):

Perform EDA to understand your data better. This includes Checking for missing values, exploring the data’s statistics, and Visualizing it to identify patterns.

Program:

```
Print(“real:\n”,real.subject.value_counts(),”\n\nfake:\n”,fake.subject.value_counts())
```

4. Feature Engineering:

Depending on your dataset, you may need to create new features or Transform existing ones. This can involve one-hot encoding categorical Variables, handling date/time data, or scaling numerical features.

Program:

```
# Label Encoding

Real[‘label’] = 1

Fake[‘label’] = 0

Import spacy

Nlp = spacy.load(“en_core_web_lg”)

Real[‘vector’] = real.title.apply(lambda x: nlp(x).vector)

Fake[‘vector’] = fake.title.apply(lambda x: nlp(x).vector)

Concat_data = pd.concat([real[[‘vector’,’label’]],fake[[‘vector’,’label’]]])

Concat_data[:5]
```

5. Split the Data:

Split your dataset into training and testing sets. This helps you evaluate Your model's performance later.

Program:

```
  X_train, X_test, y_train, y_test = train_test_split(concat_data.vector,
  Concat_data.label,
  Test_size = 0.2,
    Random_state = 1,
   Stratify = concat_data.label)
 Print(X_train.shape, y_train.shape)
 Print(X_test.shape, y_test.shape)
```

Importance of loading and processing dataset:

Loading and processing datasets are critical for fake news detection using NLP because they ensure data quality, enable feature extraction and normalization, facilitate text vectorization, help balance data, and prepare text for NLP models, ultimately leading to more accurate and efficient fake news detection.

Challenges involved in loading and preprocessing a fake news detection Dataset:

The fake news datasets may contain a mixture of reliable and unreliable sources, making it challenging to ensure data quality. Preprocessing should include data cleaning, such as removing duplicates, correcting errors, and dealing with missing values.

How to overcome the challenges of loading and preprocessing the fake news detection ?

•Carefully curate and clean the dataset, removing duplicates and correcting errors.

- Verify the reliability of data sources and consider using labeled datasets from trusted sources.
- Use techniques like oversampling (e.g., SMOTE), undersampling, or generating synthetic data to balance the classes.
- Consider using regular expressions or libraries like NLTK or spaCy for text preprocessing.
- Fine-tune the choice of embeddings based on your dataset’s characteristics.
- Apply data augmentation techniques sparingly to avoid introducing biases.
- Pad or truncate text to a fixed length.
- Optimize preprocessing and feature extraction for efficiency, considering batch processing or parallelization.
- Use production-ready libraries and frameworks for deployment.

Loading the dataset:

Loading a dataset for fake news detection in NLP involves obtaining it in a suitable format, often CSV. Python, via libraries like pandas, is used to load and explore the data. Data is split into training and testing sets. Preprocessing tasks such as cleaning, tokenization, and addressing class imbalances are essential. Finally, data is prepared for NLP models, and saving preprocessed data is a good practice to streamline subsequent model training.

```
Import numpy as np # linear algebra

Import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

Import os

For dirname, _, filenames in os.walk(‘/kaggle/input’):

 For filename in filenames:

 Print(os.path.join(dirname, filename))
```

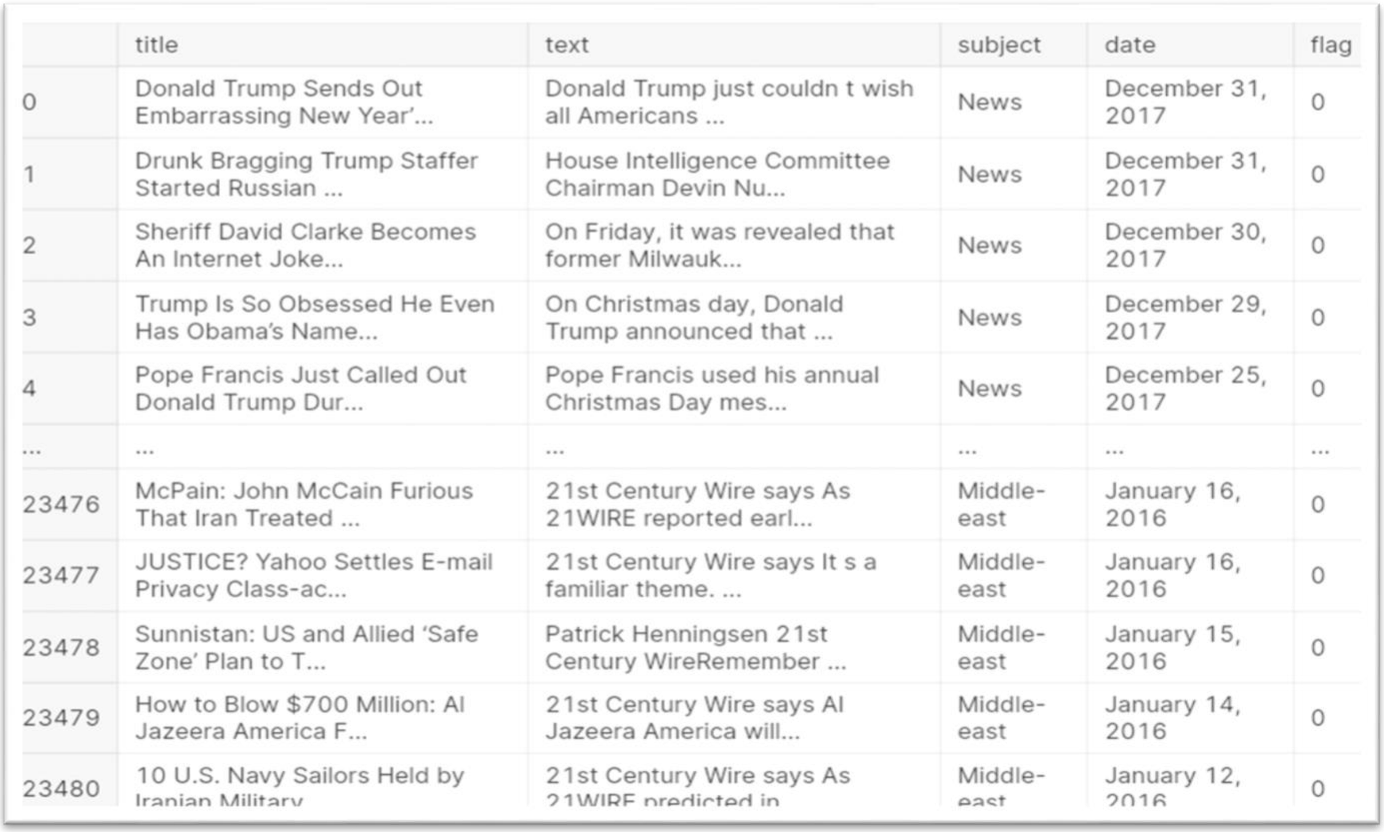
In:

```
Fake = pd.read_csv(‘/kaggle/input/fake-and-real-news-dataset/Fake.csv’)
```

Fake['flag'] = 0

Fake

Out:



	title	text	subject	date	flag
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	0
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	0
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	0
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	0
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	0
...	...	...	...	...	...
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

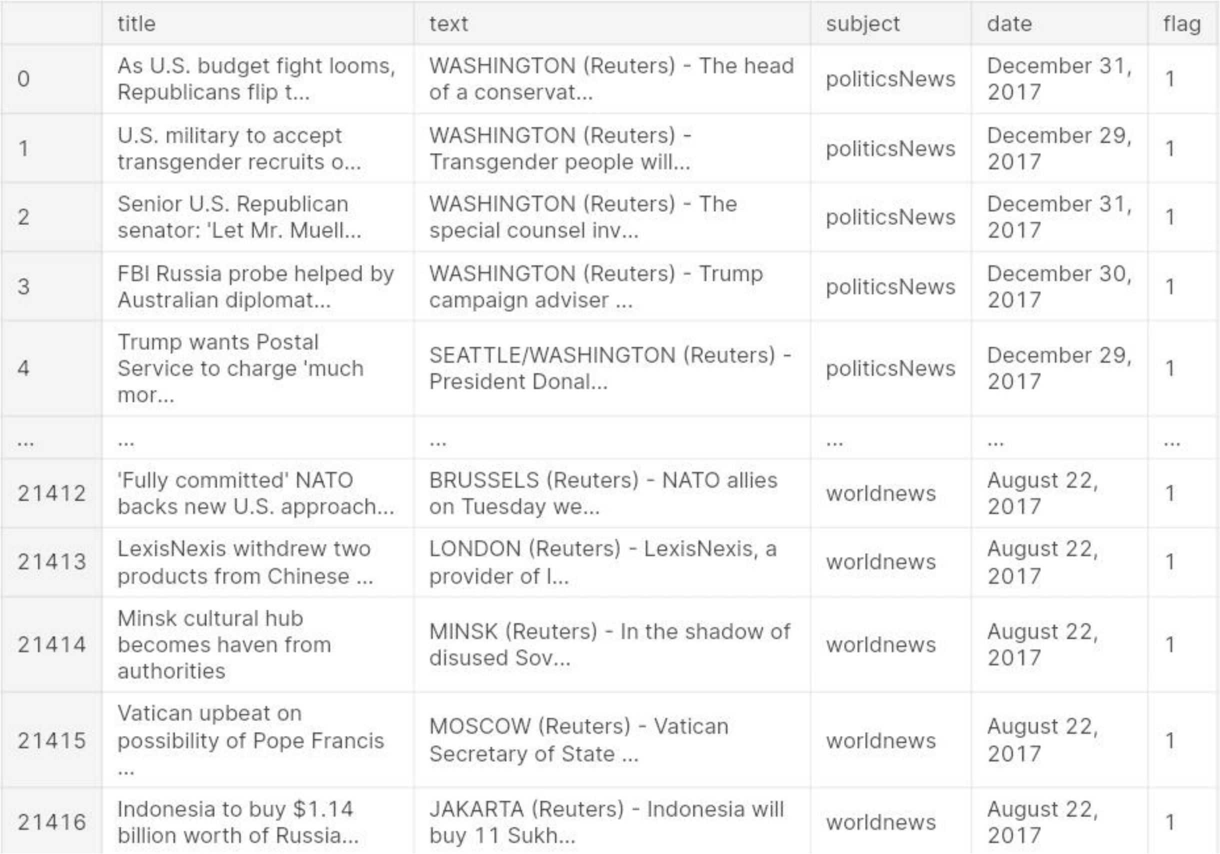
In:

True = pd.read_csv('/kaggle/input/fake-and-real-news-dataset/True.csv')

True['flag'] = 1

true

out:



	title	text	subject	date	flag
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
...	...	...	...	...	...
21412	'Fully committed' NATO backs new U.S. approach...	BRUSSELS (Reuters) - NATO allies on Tuesday we...	worldnews	August 22, 2017	1
21413	LexisNexis withdrew two products from Chinese ...	LONDON (Reuters) - LexisNexis, a provider of l...	worldnews	August 22, 2017	1
21414	Minsk cultural hub becomes haven from authorities	MINSK (Reuters) - In the shadow of disused Sov...	worldnews	August 22, 2017	1
21415	Vatican upbeat on possibility of Pope Francis ...	MOSCOW (Reuters) - Vatican Secretary of State ...	worldnews	August 22, 2017	1
21416	Indonesia to buy $1.14 billion worth of Russia...	JAKARTA (Reuters) - Indonesia will buy 11 Sukh...	worldnews	August 22, 2017	1

Preprocessing the dataset:

Preprocessing the dataset for fake news detection using NLP involves cleaning, tokenization, stopword removal, lemmatization or stemming, removing numeric and non-alphabetic characters, handling imbalanced data, text vectorization, feature selection, data splitting, handling missing data, text length normalization, label encoding, addressing outliers if needed, implementing cross-validation, feature scaling (if necessary), and conducting exploratory data analysis to prepare the data for accurate model training and evaluation.

In:

```
Data['Category'].value_counts()
```

Out:

```
Category

Fake    23481

True    21417

Name: count, dtype: int64
```

In:

```
#Transforming category values to numerical

From sklearn.preprocessing import LabelEncoder

Encoder = LabelEncoder()

Data[‘Category’] = encoder.fit_transform(data[‘Category’])
```

In:

```
Data[‘Category’]
```

Out:

```
    0       0
    1       1       0
    2       0
    3       3       0
    4       0
    5               ..

44893   1

44894   1

44895   1

44896   1

44897   1

Name: Category, Length: 44898, dtype: int64
```

In:

```
Vectorizer = TfidfVectorizer()

Title = vectorizer.fit_transform(data[‘title’])

Title
```

Out:

```
<44898x20896 sparse matrix of type ‘<class ‘numpy.float64’>’

With 546512 stored elements in Compressed Sparse Row format>
```

Some common data preprocessing tasks include:

Data cleaning:

This involves identifying and correcting errors andInconsistencies in the data. For example, this may involveRemoving duplicate records, correcting typos, and filling in missing values.

Data transformation:

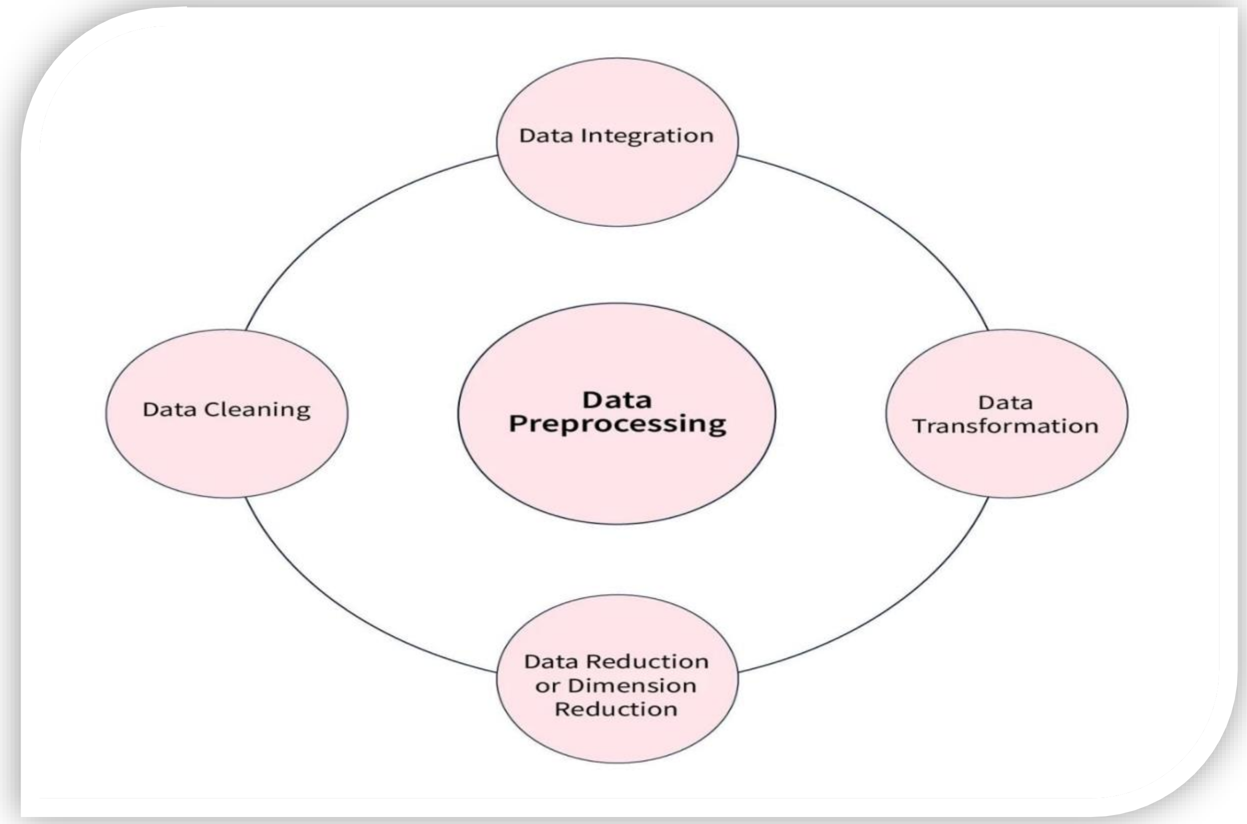
This involves converting the data into a Format that is suitable for the analysis task. For example, this may Involve converting categorical data to numerical data, or scaling The data to a suitable range.

Feature engineering:

This involves creating new features from The existing data. For example, this may involve creating features That represent interactions between variables, or features that Represent summary statistics of the data.

Data integration:

This involves combining data from multiple Sources into a single dataset. This may involve resolving Inconsistencies in the data, such as different data formats or Different variable names.Data preprocessing is an essential step in many data Science projects. By carefully preprocessing the data, data scientists can Improve the accuracy and reliability of their results.



Program:

In:

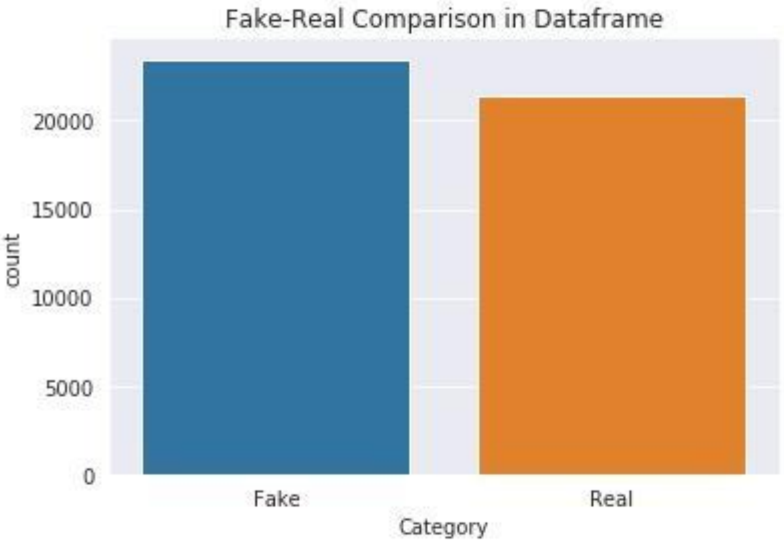
```
 # Data Preprocessing
Import pandas as pd
Import numpy as np
Import matplotlib.pyplot as plt
Import seaborn as sns
```

In:

```
# Natural Language Processing
Import nltk
From nltk.corpus import stopwords
From nltk.stem.porter import PorterStemmer
From nltk.stem import WordNetLemmatizer
From nltk.tokenize import word_tokenize, sent_tokenize
From nltk import pos_tag
From nltk.corpus import wordnet
```

In:

```
Sns.set_style(“darkgrid”)
Sns.countplot(x=”category”, Data=df)
Plt.title(“Fake-Real Comparison in Dataframe”)
Plt.xlabel(“Category”)
Plt.xticks([0, 1], [“Fake”, “Real”])
Plt.show();
```



In:

```
Df.isna().sum()
```

Out:

```
Title      0

Text       0

Subject    0

Date       0

Category   0

Dtype: int64
```

In:

```
Print(“Number of rows:”, len(df))

Print(“Number of columns:”, len(df.columns))

Number of rows: 44898

Number of columns: 5
```

In:

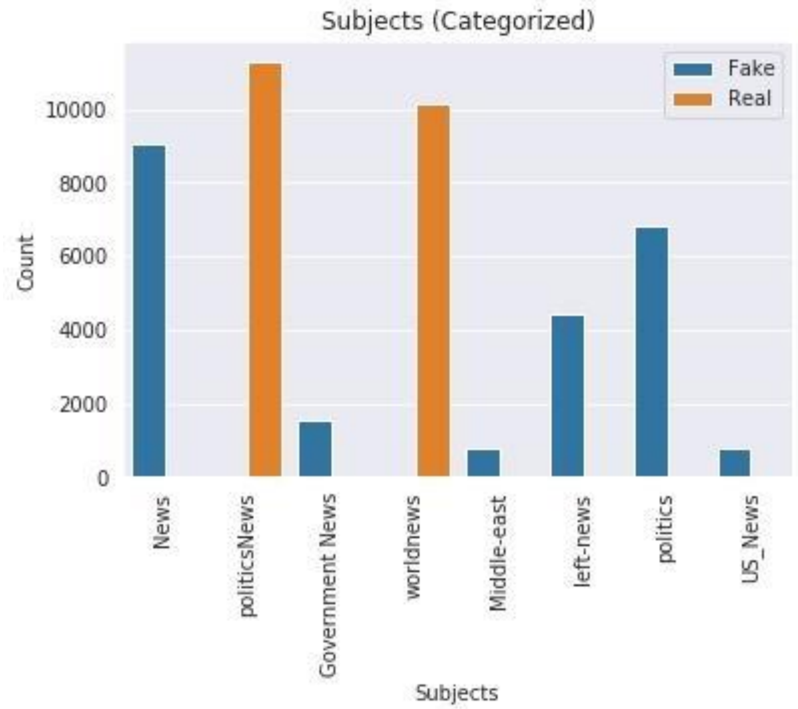
```
Df[“subject”].value_counts()
```

Out:

```
politicsNews      11272

worldnews         10145

News              9050

Politics          6841

Left-news         4459

Government News   1570

US_News           783

Middle-east       778

Name: subject, dtype: int64
```

In:

```
Sns.countplot(x=”subject”,
            Hue=”category”,
            Data=df)


Plt.title(“Subjects (Categorized)”)
Plt.xlabel(“Subjects”)
Plt.ylabel(“Count”)
Plt.xticks(rotation=90)
Plt.legend([“Fake”, “Real”], loc=”best”)
Plt.show();
```



In:

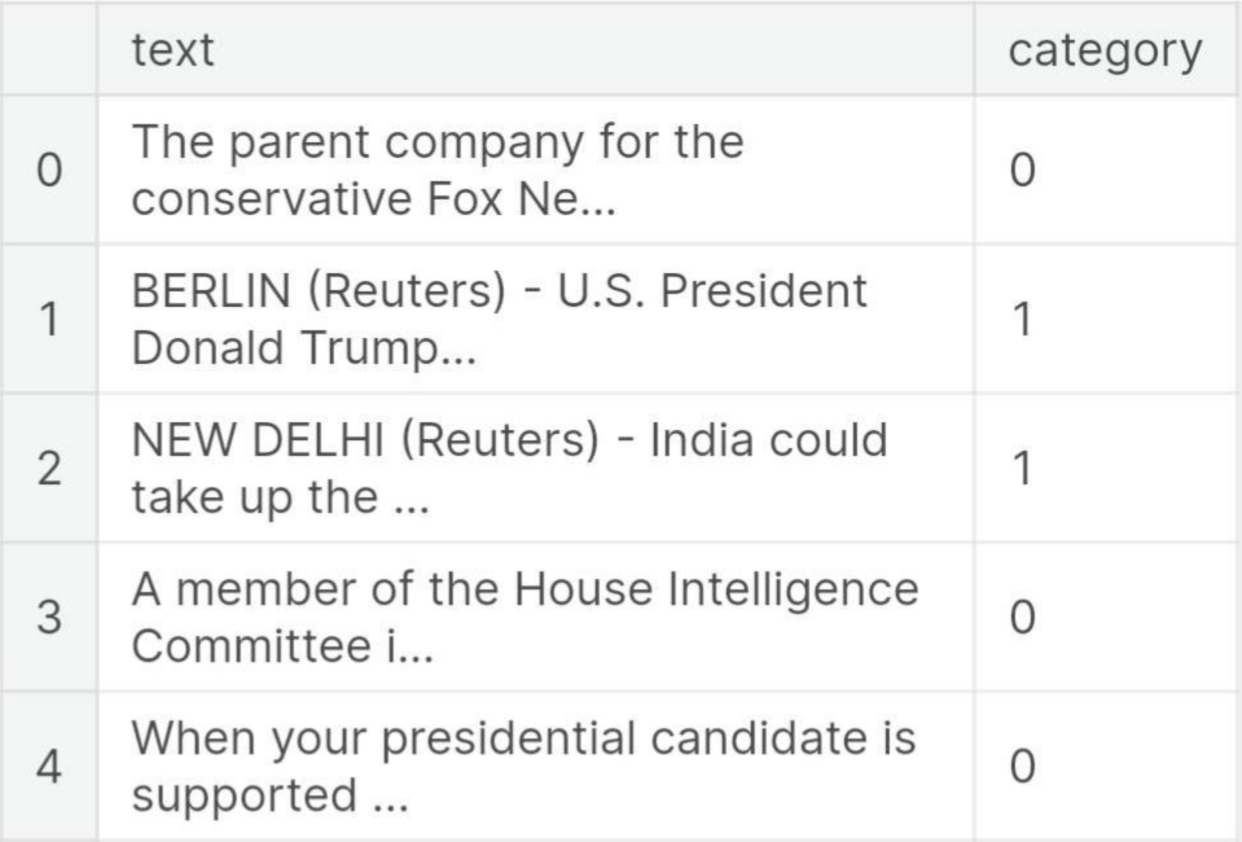
```
Df[“text”] = df[“text”] + “ “ + df[“title”]
```

```
Del df[“title”]
Del df[“subject”]
Del df[“date”]
```

In:

```
Df.head()
```

Out:



	text	category
0	The parent company for the conservative Fox Ne...	0
1	BERLIN (Reuters) - U.S. President Donald Trump...	1
2	NEW DELHI (Reuters) - India could take up the ...	1
3	A member of the House Intelligence Committee i...	0
4	When your presidential candidate is supported ...	0

In:

```
Stop = set(stopwords.words(“english”))
```

```
Punc = list(string.punctuation)

Stop.update(punc)
```

DATA CLEANING

```
Def remove_html(text):

    Soup = BeautifulSoup(text, “html.parser”)

    Return soup.get_text()


#Remove the square brackets

Def remove_between_square_brackets(text):

    Return re.sub(“\[[^]]*\]”, “”, text)


# Remove URLs

Def remove_between_square_brackets(text):

Return re.sub(r”http\S+”, “”, text)


#Remove the stopwords

Def remove_stopwords(text):

 Final_text = []

 For I in text.split():

 If i.strip().lower() not in stop:

 Final_text.append(i.strip())

 Return “ “.join(final_text)


#Remove the noisy text
```

```
Def denoise_text(text):
Text = remove_html(text)
Text = remove_between_square_brackets(text)
Text = remove_stopwords(text)
Return text
```

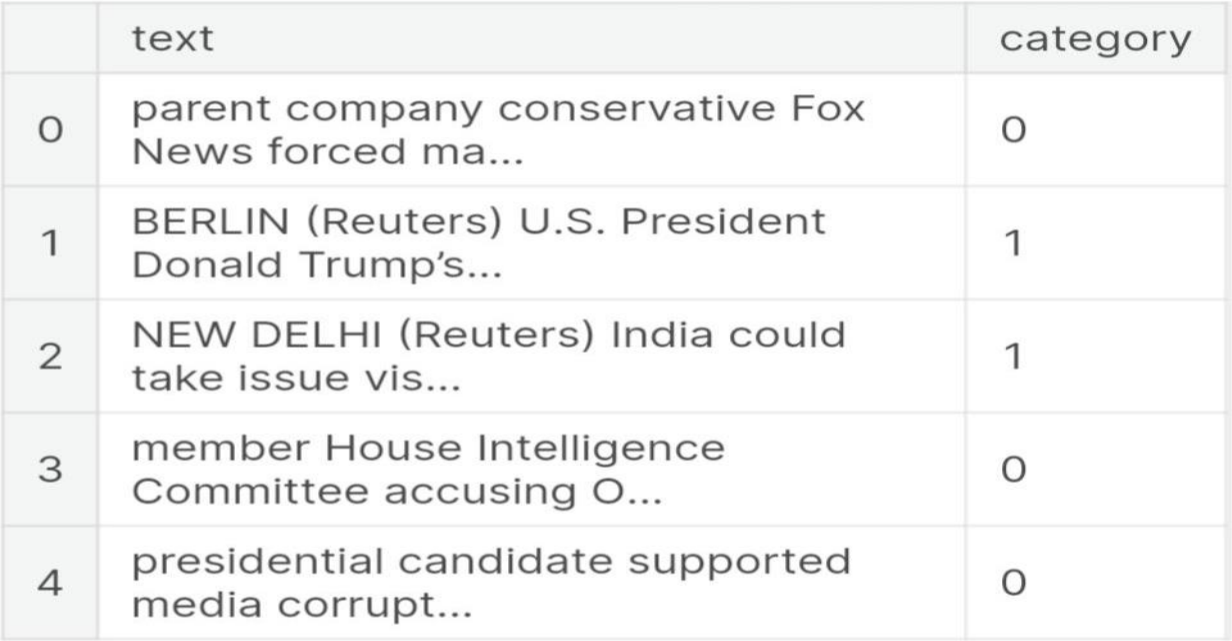
In:

```
#Apply functions
Df["text"] = df["text"].apply(denoise_text)
```

In:

```
Df.head()
```

Out:



	text	category
0	parent company conservative Fox News forced ma...	0
1	BERLIN (Reuters) U.S. President Donald Trump's...	1
2	NEW DELHI (Reuters) India could take issue vis...	1
3	member House Intelligence Committee accusing O...	0
4	presidential candidate supported media corrupt...	0

Conclusion:

In the quest to build a fake news detection using nlp, 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.