FAKE NEWS DETECTION USING NLP

TEAM MEMBER

YUVA SREE.U

Phase 2 Submission Document

Project: Fake news detection

Introduction:

- Fake news detection is the process of identifying and verifying the accuracy of news or information that is intentionally false, misleading, or fabricated. It has become a critical concern in today's digital age, where misinformation can spread rapidly through various media channels. Here's an introduction to the topic:

- **Definition of Fake News:** Fake news encompasses various types of misinformation, including fabricated stories, manipulated images or videos, and misleading headlines. It can be spread through websites, social media, or traditional media outlets.

- **Motivations for Fake News:** Fake news can be created for various reasons, such as political manipulation, financial gain, or simply for entertainment. It often seeks to exploit emotions, biases, or controversy to gain attention and traction.

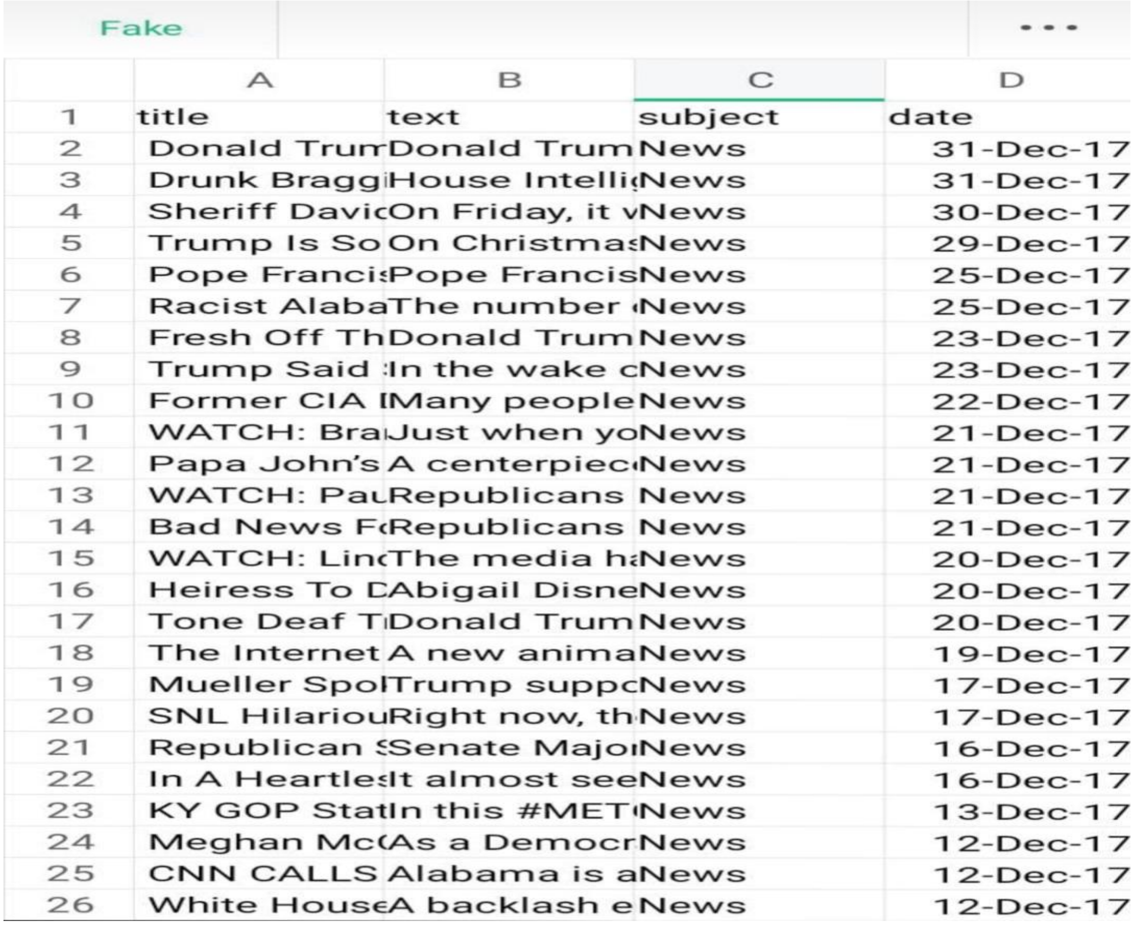
- **Impact of Fake News:** Fake news can have serious consequences, including influencing public opinion, swaying elections, causing panic, or harming individuals' reputations. It can erode trust in journalism and democratic processes.

- **Challenges in Fake News Detection:** Detecting fake news is a complex task due to its constantly evolving nature. Some challenges include the speed at which fake news spreads, the use of sophisticated techniques to make it appear legitimate, and the fine line between satire and actual misinformation.

Data Source:

A good data source for Fake news detection using NLP should be Accurate, Complete, Covering the geographic area of interest, Accessible.

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



	A	B	C	D
1	title	text	subject	date
2	Donald Trur	Donald Trum	News	31-Dec-17
3	Drunk Braggi	House Intelli(	News	31-Dec-17
4	Sheriff Davic	On Friday, it v	News	30-Dec-17
5	Trump Is So	On Christma:	News	29-Dec-17
6	Pope Franci:	Pope Francis	News	25-Dec-17
7	Racist Alaba	The number (	News	25-Dec-17
8	Fresh Off Th	Donald Trum	News	23-Dec-17
9	Trump Said :	In the wake c	News	23-Dec-17
10	Former CIA I	Many people	News	22-Dec-17
11	WATCH: Bra	Just when yo	News	21-Dec-17
12	Papa John's	A centerpiec	News	21-Dec-17
13	WATCH: Pau	Republicans	News	21-Dec-17
14	Bad News F(	Republicans	News	21-Dec-17
15	WATCH: Linc	The media ha	News	20-Dec-17
16	Heiress To D	Abigail Disne	News	20-Dec-17
17	Tone Deaf T	Donald Trum	News	20-Dec-17
18	The Internet	A new anima	News	19-Dec-17
19	Mueller Spol	Trump suppc	News	17-Dec-17
20	SNL Hilariou	Right now, th	News	17-Dec-17
21	Republican S	Senate Majo	News	16-Dec-17
22	In A Heartle:	It almost see	News	16-Dec-17
23	KY GOP Stat	In this #MET(	News	13-Dec-17
24	Meghan Mc(	As a Democr	News	12-Dec-17
25	CNN CALLS	Alabama is a	News	12-Dec-17
26	White House	A backlash e	News	12-Dec-17

DATA COLLECTION AND PREPARATION :

- Gather a diverse dataset of news articles or social media posts, including both real and fake examples. These articles should cover a wide range of topics and sources.

- Prepare the text data by removing stop words, punctuation, and converting text to lowercase. Tokenization and stemming or lemmatization may also be applied to standardize the text.

FEATURES EXTRACTION AND LABELLING:

- Convert the textual content into numerical features that machine learning algorithms can understand. Common techniques include TF-IDF (Term Frequency-Inverse Document Frequency) and word embeddings like Word2Vec or GloVe.

- Annotate your dataset to indicate which articles are real and which are fake. This labeled data will be used for training and testing your model.

MODEL SELECTION AND TRAINING:

- Choose an appropriate NLP model or algorithm. Common choices include logistic regression, random forests, or more advanced methods like recurrent neural networks (RNNs) or transformer-based models like BERT.

- Use the labelled dataset to train your NLP model. The model learns to recognize patterns and features that distinguish real news from fake news.

TRAINING AND EVALUATION:

- Use the labelled dataset to train your NLP model. The model learns to recognize patterns and features that distinguish real news from fake news.

- Assess the performance of your model using metrics such as accuracy, precision, recall, and F1-score on a separate validation or test dataset. Fine-tune your model to improve its performance.

FEATURE ENGINEERING AND BIAS DETECTION:

- Experiment with different features or techniques, such as n-grams, to enhance your model’s ability to detect fake news.

- Be aware of potential biases in your dataset and model. Ensure that your model doesn't unfairly label certain sources or topics as fake news.

DEPLOYMENT :

- Once satisfied with the model's performance, deploy it to analyze and classify news articles or social media content in real-time.

- Continuously monitor your model's performance and update it as needed to adapt to evolving fake news tactics.

ETHICAL CONSIDERATION:

- Be mindful of ethical considerations, such as privacy and freedom of speech, when developing and deploying fake news detection systems.

- Remember that fake detection is a challenging task, and achieving high accuracy can be difficult due to the evolving nature of fake news. It often requires ongoing research and adaptation to stay effective in identifying misinformation and disinformation online.

- It's important to balance the detection of fake news with respect for free speech and privacy. Striking this balance can be challenging and requires careful consideration.

CONTINUOUS EVOLUTION:

- Fake news detection methods must continually adapt to new tactics used by purveyors of misinformation. Ongoing research and collaboration are crucial in this ever-changing landscape.

TEXT ANALYSIS:

- NLP techniques are used to analyze the content of news articles or social media posts. This includes sentiment analysis, identifying unusual language patterns, and examining the tone of the text.
- Creating meaningful features from the text data is crucial. Features might include word frequency, readability scores, or linguistic features that can help distinguish fake from real news.

SUPERVISED LEARNING:

- Most fake news detection models are trained using supervised learning. They learn from labeled datasets that contain examples of both fake and real news to make predictions on new, unlabeled data.

ENSEMBLE METHODS:

- Combining the predictions of multiple machine learning models can enhance accuracy. Techniques like Random Forests or Gradient Boosting are commonly used.

PROGRAM:

FAKE NEWS DETECTION

IMPORT LIBRARIES:

In[1]:

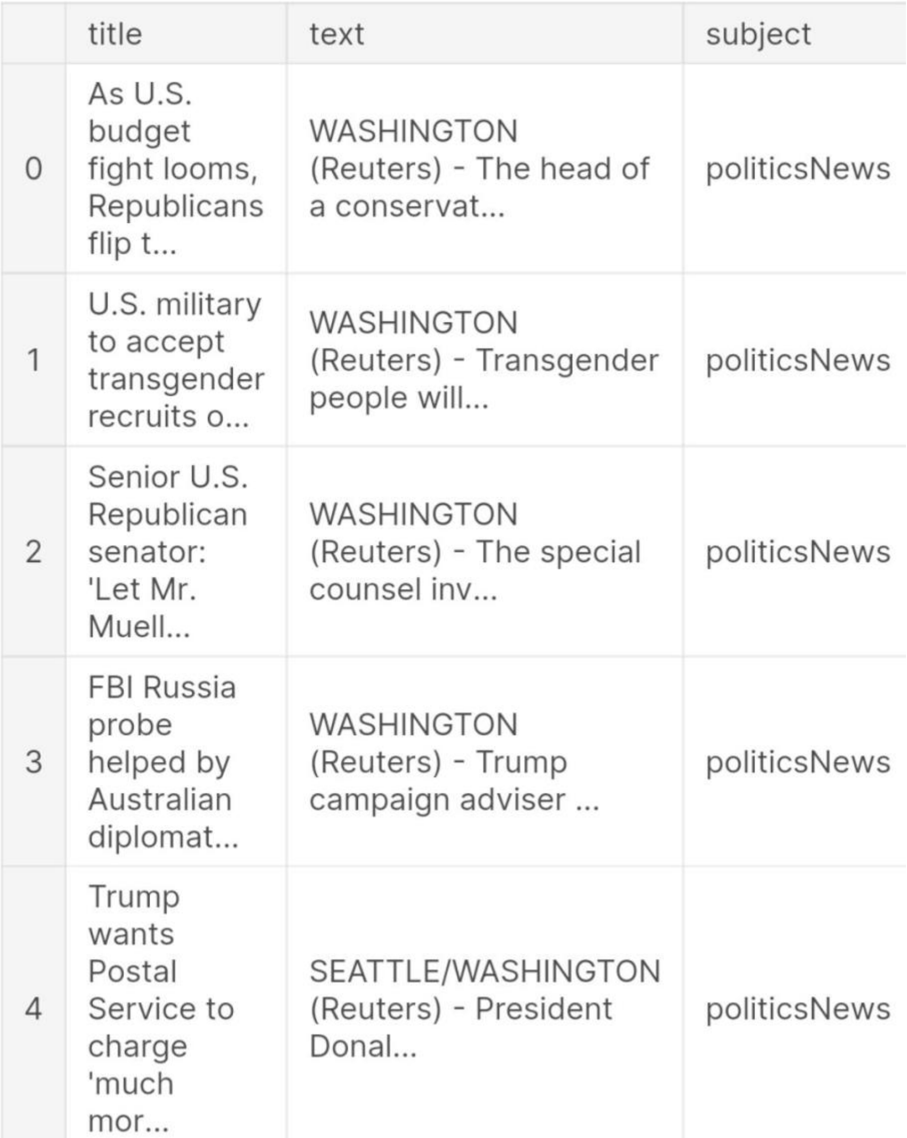
```
Import numpy as np
Import pandas as pd
Import matplotlib.pyplot as plt
Import seaborn as sns

Import nltk
Import re
Import string

From sklearn.model_selection import train_test_split
From sklearn.metrics import classification_report

Import keras
From keras.preprocessing import text,sequence
From keras.models import Sequential
From keras.layers import Dense,Embedding,LSTM,Dropout

Import warnings
Warnings.filterwarnings(‘ignore’)
```

```
Import os
For dirname, _, filenames in os.walk('/kaggle/input'):
    For filename in filenames:
        Print(os.path.join(dirname, filename))
```

LOAD AND CHECK DATA:

In[2]:

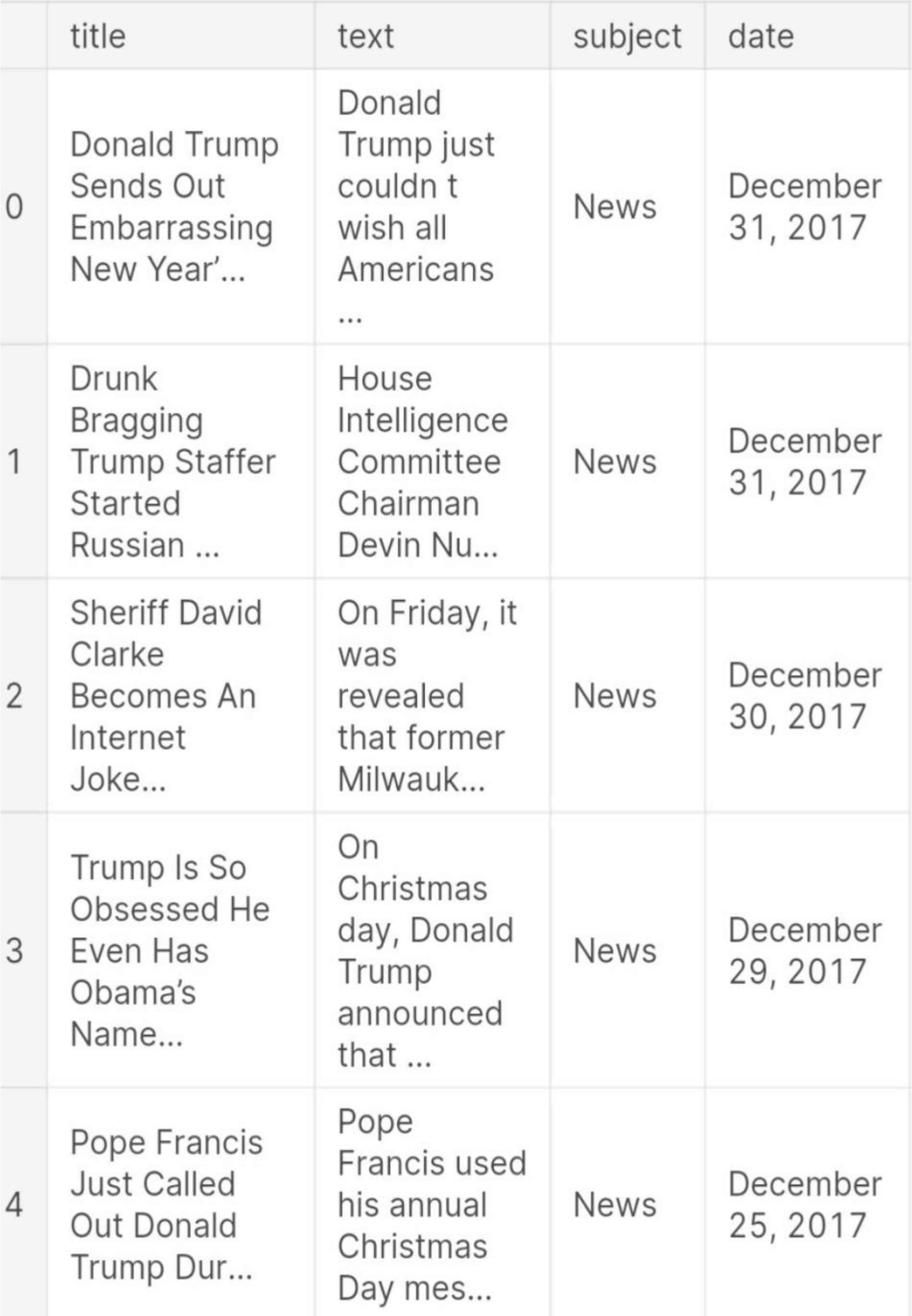
```
Real_data = pd.read_csv('/kaggle/input/fake-and-real-news-dataset/True.csv')
Fake_data = pd.read_csv('/kaggle/input/fake-and-real-news-dataset/Fake.csv')
```



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

In[3]:

```
real_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

<u>In[5]:</u>

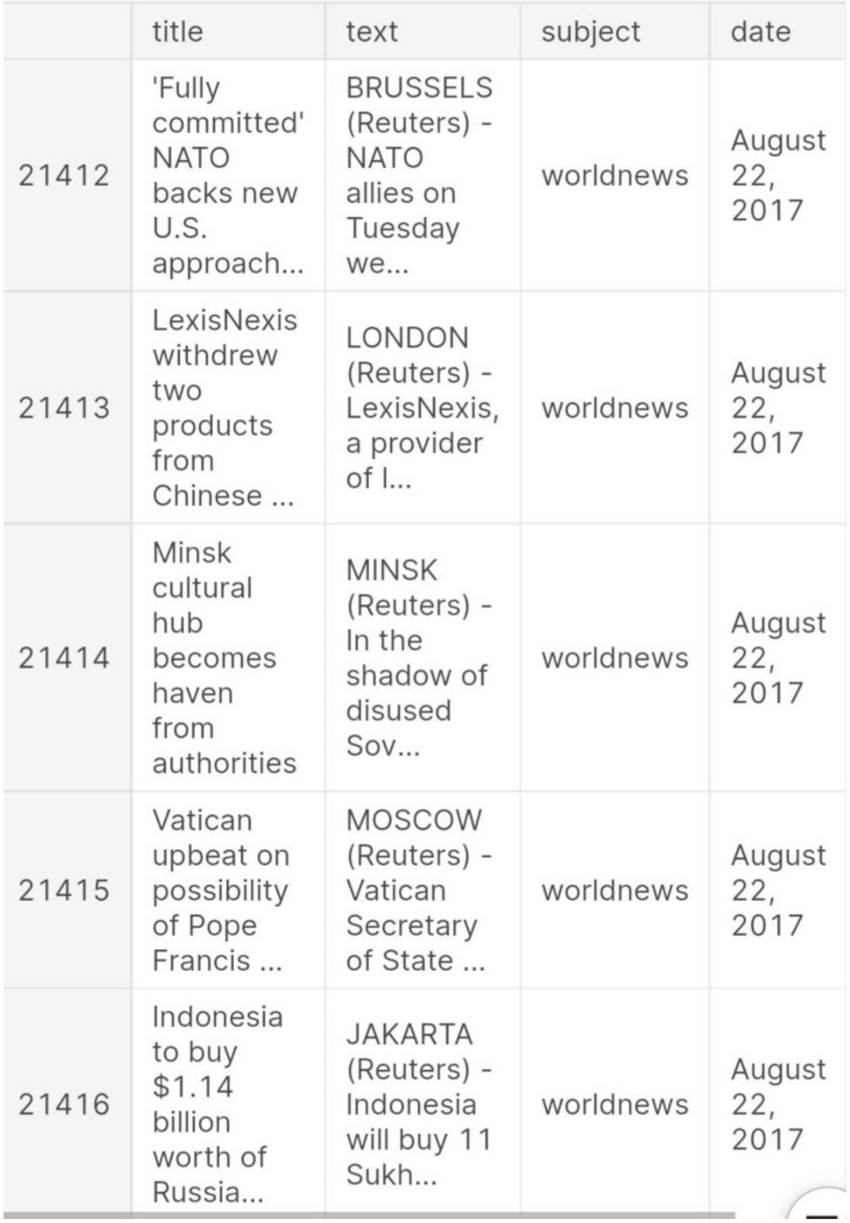
```
real_data[‘target’] = 1
```

```
fake_data[‘target’] = 0
```

In[6]:

```
real_data.tail()
```

Out[6]:



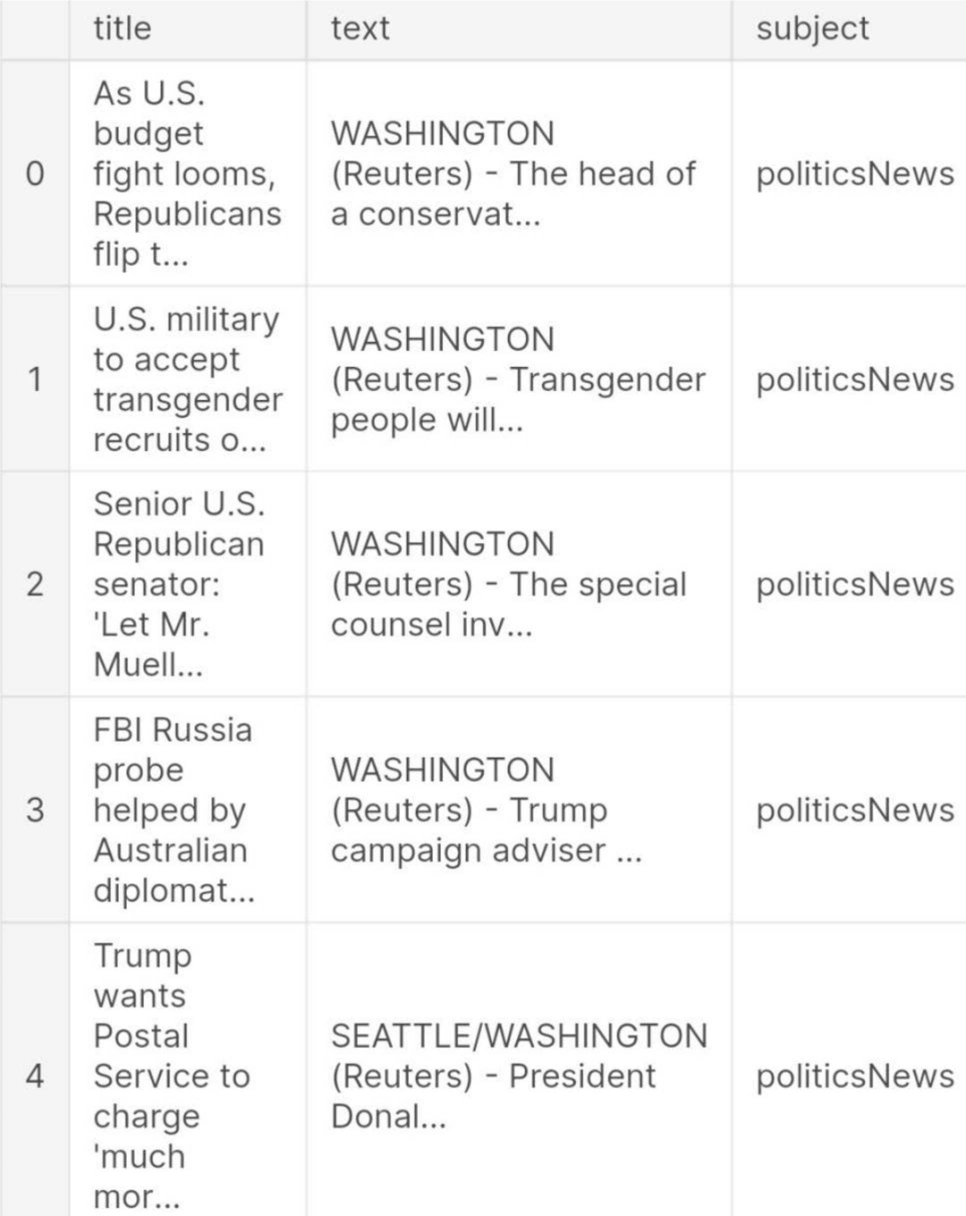
	title	text	subject	date
21412	'Fully committed' NATO backs new U.S. approach...	BRUSSELS (Reuters) - NATO allies on Tuesday we...	worldnews	August 22, 2017
21413	LexisNexis withdrew two products from Chinese ...	LONDON (Reuters) - LexisNexis, a provider of l...	worldnews	August 22, 2017
21414	Minsk cultural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

In[7]:

```
Data = pd.concat([real_data, fake_data], ignore_index=True, sort=False)
```

```
Data.head()
```

Out[7]:



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

In[8]:

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

Out[8]:

```
Title      0
Text       0
Subject    0
Date       0
Target     0
Dtype: int64
```

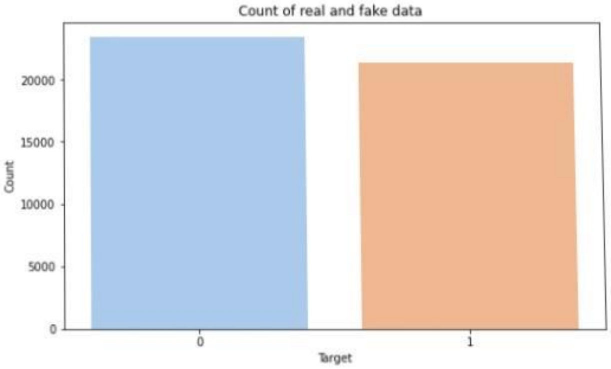
VISUALIZATION

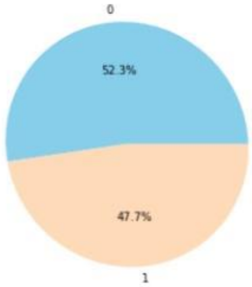
Count of Fake and Real Data

In[9]:

```
print(data[“target”].value_counts())
fig, ax = plt.subplots(1,2, figsize=(19, 5))
G1 = sns.countplot(data.target,ax=ax[0],palette=”pastel”);
G1.set_title(“Count of real and fake data”)
G1.set_ylabel(“Count”)
G1.set_xlabel(“Target”)
G2 =
plt.pie(data[“target”].value_counts().values,explode=[0,0],labels=data.target.value_counts().index, autopct=’%1.1f%%’,colors=[‘SkyBlue’,’PeachPuff’])
fig.show()
```

```
    0    1    21417
Name: target,  dtype: int64
```





Distribution of The Subject According to Real and Fake Data

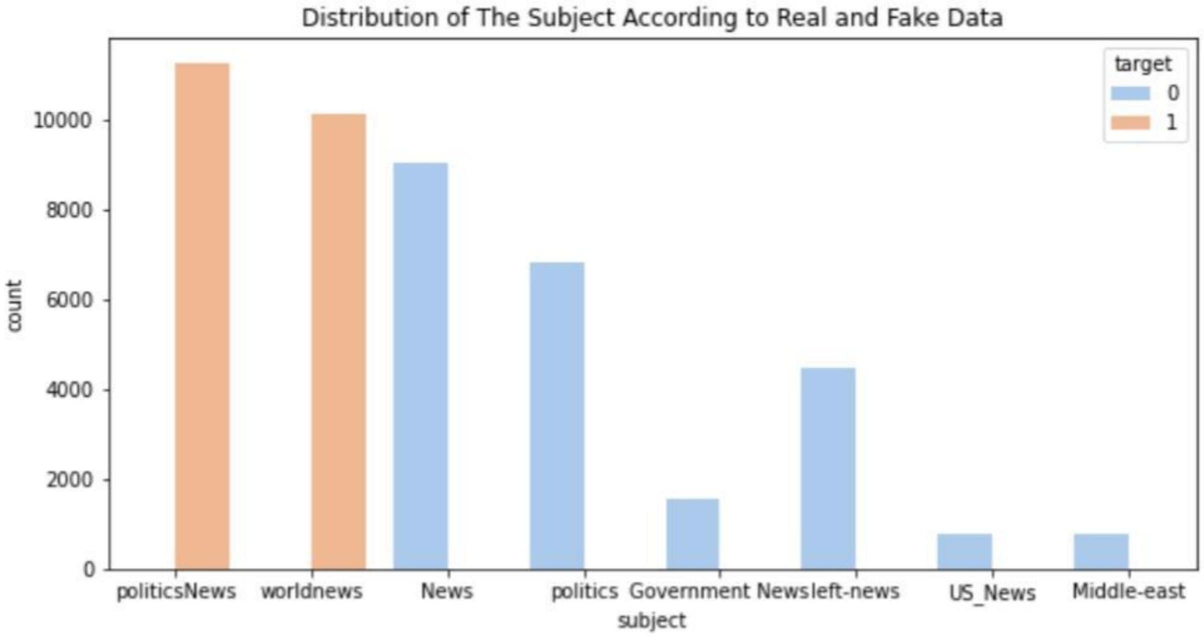
In[9]:

```
print(data.subject.value_counts())
plt.figure(figsize=(10, 5))

ax = sns.countplot(x=”subject”,  hue=’target’, data=data, palette=”pastel”)
plt.title(“Distribution of The Subject According to Real and Fake Data”)
```

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

Out[10]:

```
Text(0.5, 1.0, ‘Distribution of The Subject According to Real and Fake Data’)
```

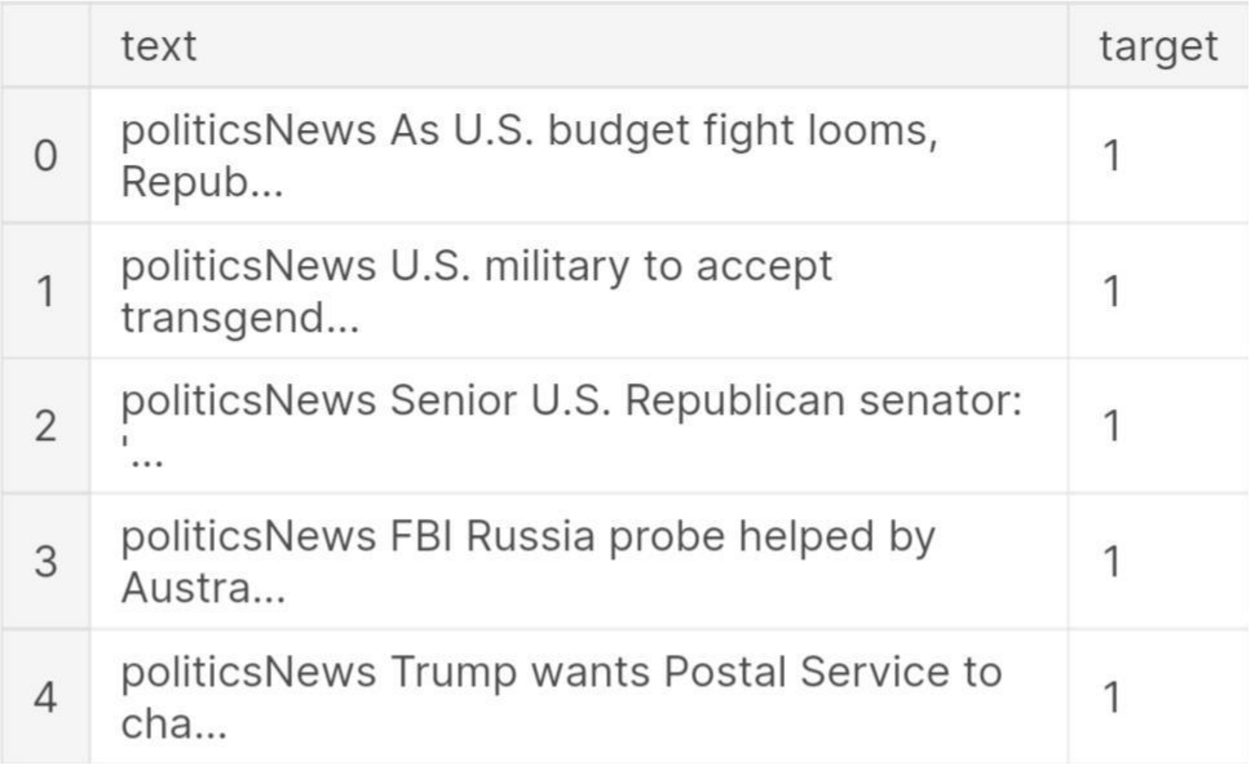


DATA CLEANING

In[11]:

```
data[‘text’]= data[‘subject’] + “ “ + data[‘title’] + “ “ + data[‘text’]
del data[‘title’]
del data[‘subject’]
del data[‘date’]
data.head()
```

Out[11]:



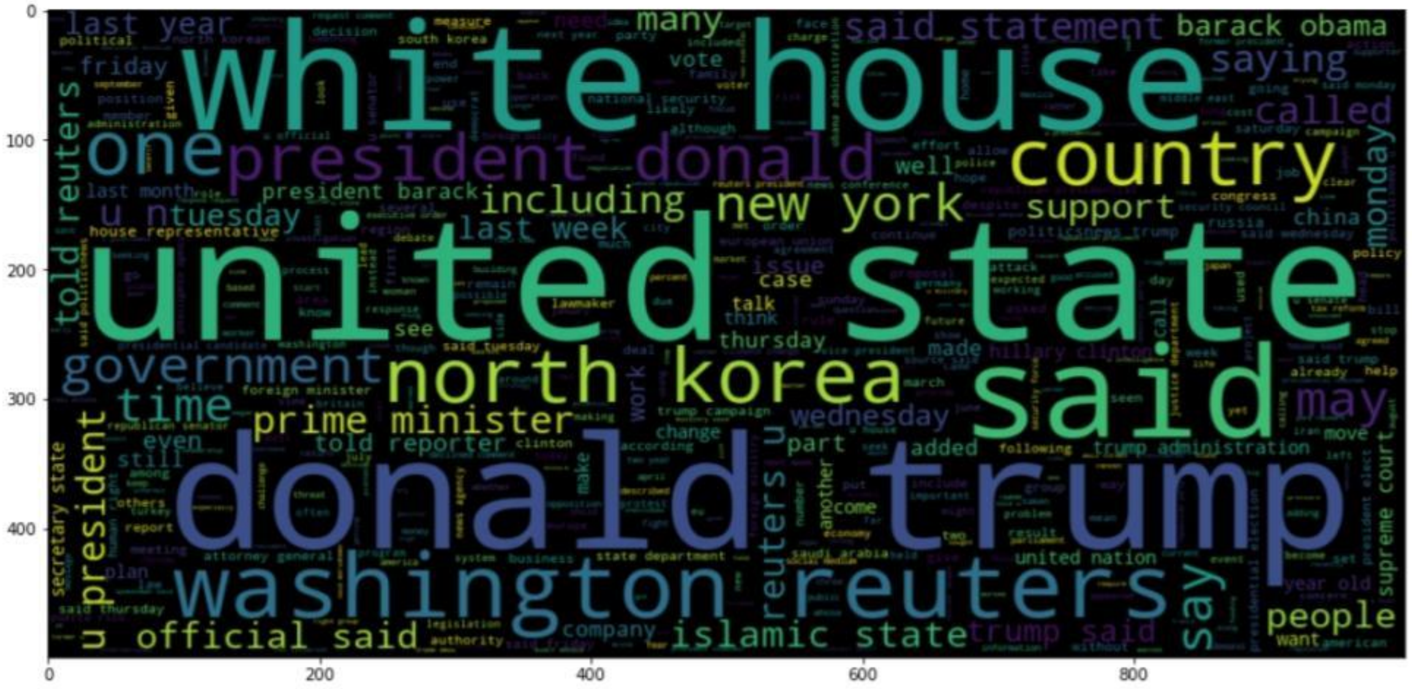
	text	target
0	politicsNews As U.S. budget fight looms, Repub...	1
1	politicsNews U.S. military to accept transgend...	1
2	politicsNews Senior U.S. Republican senator: '...	1
3	politicsNews FBI Russia probe helped by Austra...	1
4	politicsNews Trump wants Postal Service to cha...	1

Int[12]:

```
from wordcloud import WordCloud,STOPWORDS

plt.figure(figsize = (15,15))

Wc = WordCloud(max_words = 500 , width = 1000 , height = 500 , stopwords =
STOPWORDS).generate(" ".join(data[data.target == 1].text))

Plt.imshow(wc , interpolation = 'bilinear')
```

Out[12]:

```
<matplotlib.image.AxesImage at 0x7f6934fd2750>
```

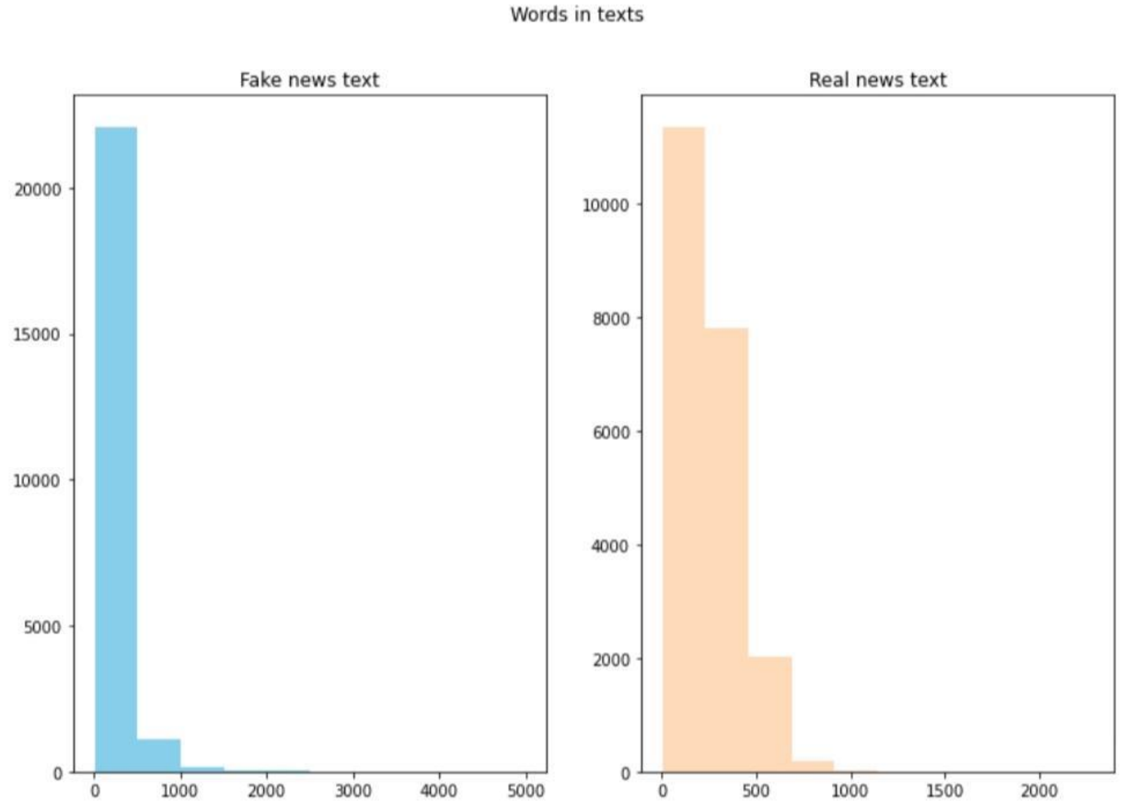


Int[13]:

Number of words in each text

```
fig,(ax1,ax2)=plt.subplots(1,2,figsize=(12,8))
```

```
text_len=data[data[‘target’]==0][‘text’].str.split().map(lambda x: len(x))
ax1.hist(text_len,color=’SkyBlue’)
ax1.set_title(‘Fake news text’)
text_len=data[data[‘target’]==1][‘text’].str.split().map(lambda x: len(x))
ax2.hist(text_len,color=’PeachPuff’)
ax2.set_title(‘Real news text’)
fig.suptitle(‘Words in texts’)
plt.show()
```



The number of words seems to be a bit different. 500 words are most common in real news category while around 250 words are most common in fake news category.

N-Gram Analysis

Int[14]:

```
Texts = ‘ ‘.join(data[‘text’]
```

Int[15]:

```
String = texts.split(“ “)
```

Int[16]:

```
  def draw_n_gram(string,i):
   N_gram = (pd.Series(nltk.ngrams(string, i)).value_counts())[:15]
   N_gram_df=pd.DataFrame(n_gram)
   N_gram_df = n_gram_df.reset_index()
   N_gram_df = n_gram_df.rename(columns={“index”: “word”, 0: “count”})
   Print(n_gram_df.head())
   Plt.figure(figsize = (16,9))
   Return sns.barplot(x=’count’,y=’word’, data=n_gram_df)
```

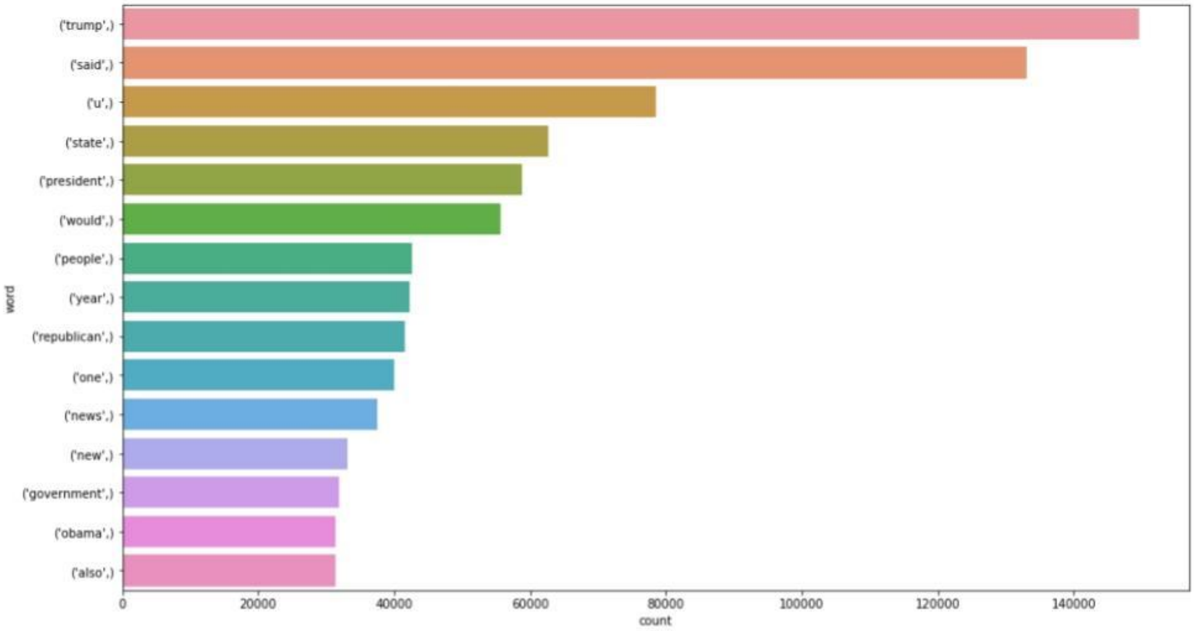
Unigram Analysis

Int[17]:

```
Draw_n_gram(string )
           word   count
0     (trump,)  149603
1      (said,)  133030
2         (u,)   78516
3     (state,)   62726
4  (president,)   58790
```

Out[17]:

```
<AxesSubplot:xlabel=’count’, ylabel=’word ’>
```



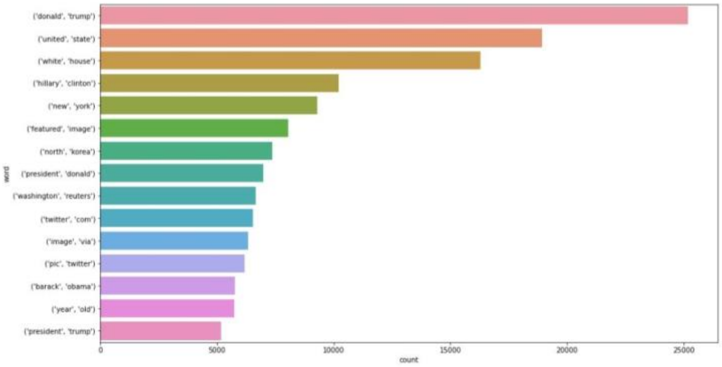
Bigram Analysis

Int[18]:

```
Draw_n_gram(string,2)
```

```
0     (donald, trump)  25203
1     (united, state)  18943
2      (white, house)  16296
3  (hillary, clinton)  10217
4         (new, york)   9305
```

Out[18]:



```
<AxesSubplot:xlabel=’count’, ylabel=’word’>
```

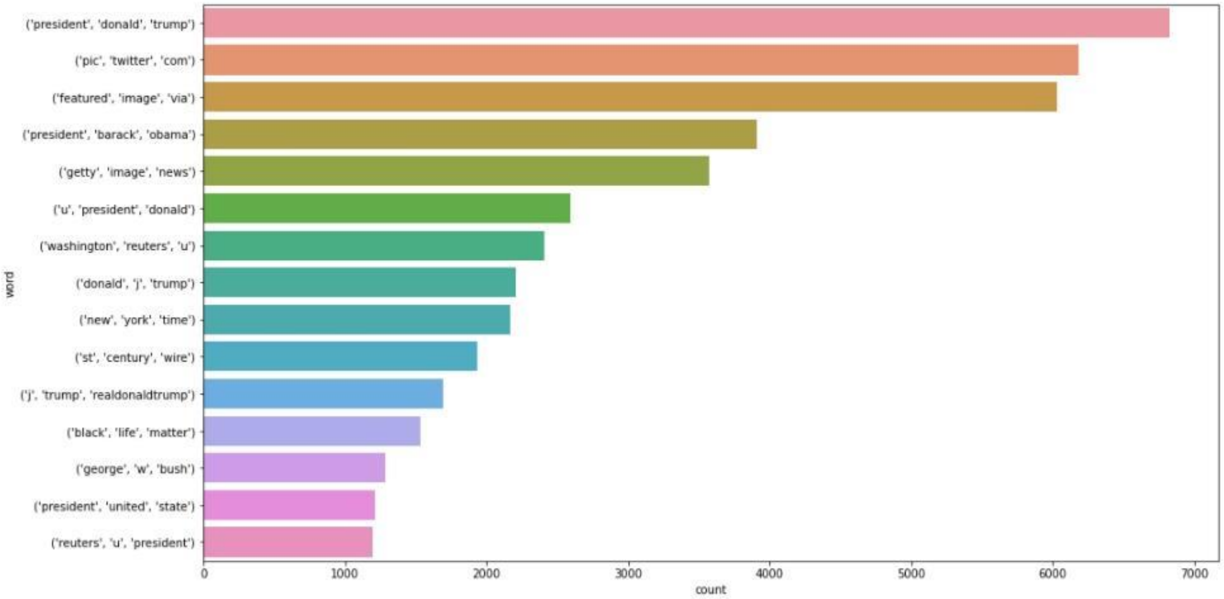
Trigram Analysis

Int[19]:

```
Draw_n_gram(string,3)
```

Out[19]

```
<AxesSubplot:xlabel=’count’, ylabel=’word’>
```



Train Test Split

Int[20]:

```
X_train, X_test, y_train, y_test = train_test_split(data[‘text’], data[‘target’], random_state=0)
```

Tokenizing

Tokenizing Text -> Repsesenting each word by a number

Mapping of orginal word to number is preserved in word_index property of tokenizer

CONCLUSION AND FUTURE WORK(Phase2):

Project Conclusion:

In conclusion, fake news detection using Natural Language Processing (NLP) is a vital and evolving field in the fight against misinformation. NLP techniques have shown promise in identifying and flagging potentially deceptive content by analyzing linguistic patterns, sources, and context. However, it is essential to acknowledge that no single method is foolproof, and ongoing research and development are necessary to stay ahead of increasingly sophisticated fake news tactics. Collaborative efforts between researchers, technology companies, and fact-checkers are crucial in building more robust and accurate fake news detection systems to promote trustworthy information in the digital age.