Fake News detection

1. Introduction

Fake news is a treat to the stability of the society as a whole. Currently, it has become a menace and had the capacity to sway elections, like the case of Donald Trump

2. Problem Definition

People often share information without verifying it. This can include political agendas or to sway opinions. Media outlets are fighting fake news so that it does not impact their credibility and revenue. Hence , it is important to detect fake news as it impacts the society as a whole.

3. Experimental Evaluation

3.1 Methodology

- Data cleaning and removing the unwanted columns from the data
- Combining the true and Fake into a single Label column and concating the 2 data sets with pandas and numpy
- Checking the length and count of the messages
- Used label encoding to add values to spam and ham messages
- Used stop words and word net lemmatizer to clean the data
- Used word cloud to analyse the most common words used in fake and True categories in label column
- Convert text into vectors using TF-IDF
- Instantiate MultinomiaLNB classifier
- Split feature and Label
- Train and split on the data set
- Train and predict
- Plot confusion matrix heatmap

3.2 Results

1. Found out the shape and size of the data

In [274]:
```
import pandas as pd
import numpy as np
```

In [281]:
```
#for date - https://stackoverflow.com/questions/38067704/how-to-change-the-datetime-format-in-pandas
```

In [282]:
```
Fake = pd.read_csv('Fake.csv')
Fake.head()
```

Out[282]:

	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

In [283]:
```
Fake.shape
```

Out[283]: (23481, 4)

2. Added a Fake category under label to the Fake.csv data set

In [284]:
```
Fake['Label']="Fake"
```

In [285]:
```
Fake
```

Out[285]:

	title	text	subject	date	Label
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	Fake
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	Fake
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	Fake
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	Fake
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	Fake
...	...	...	...	...	...
23476	McPain: John McCain Furious That Iran Treated ...	21st Century Wire says As 21WIRE reported earl...	Middle-east	January 16, 2016	Fake
23477	JUSTICE? Yahoo Settles E-mail Privacy Class-ac...	21st Century Wire says It s a familiar theme. ...	Middle-east	January 16, 2016	Fake
23478	Sunnistan: US and Allied 'Safe Zone' Plan to T...	Patrick Henningsen 21st Century WireRemember ...	Middle-east	January 15, 2016	Fake
23479	How to Blow $700 Million: Al Jazeera America F...	21st Century Wire says Al Jazeera America will...	Middle-east	January 14, 2016	Fake
23480	10 U.S. Navy Sailors Held by Iranian Military ...	21st Century Wire says As 21WIRE predicted in ...	Middle-east	January 12, 2016	Fake

23481 rows × 5 columns

3. Added a true label to the true.csv data set

In [288]:
```
true['Label']="true"
```

In [289]:
```
true
```

Out[289]:

	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	true
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	true
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	true
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017	true
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017	true
...	...	...	...	...	...
21412	'Fully committed' NATO backs new U.S. approach...	BRUSSELS (Reuters) - NATO allies on Tuesday we...	worldnews	August 22, 2017	true
21413	LexisNexis withdrew two products from Chinese ...	LONDON (Reuters) - LexisNexis, a provider of l...	worldnews	August 22, 2017	true
21414	Minsk cultural hub becomes haven from authorities	MINSK (Reuters) - In the shadow of disused Sov...	worldnews	August 22, 2017	true
21415	Vatican upbeat on possibility of Pope Francis ...	MOSCOW (Reuters) - Vatican Secretary of State ...	worldnews	August 22, 2017	true
21416	Indonesia to buy $1.14 billion worth of Russia...	JAKARTA (Reuters) - Indonesia will buy 11 Sukh...	worldnews	August 22, 2017	true

21417 rows × 5 columns

4. Used concat function to combine two data sets. And two labels: True and Fake

21417 rows × 5 columns

```
In [290]: #concatenation
          df_cat1 = pd.concat([Fake, true], axis=0, ignore_index=True)
          print (df_cat1)
```

```
                                                   title  \
0       Donald Trump Sends Out Embarrassing New Year'...
1       Drunk Bragging Trump Staffer Started Russian ...
2       Sheriff David Clarke Becomes An Internet Joke...
3       Trump Is So Obsessed He Even Has Obama's Name...
4       Pope Francis Just Called Out Donald Trump Dur...
...                                                   ...
44893  'Fully committed' NATO backs new U.S. approach...
44894  LexisNexis withdrew two products from Chinese ...
44895  Minsk cultural hub becomes haven from authorities
44896  Vatican upbeat on possibility of Pope Francis ...
44897  Indonesia to buy $1.14 billion worth of Russia...

                                                    text    subject  \
0      Donald Trump just couldn t wish all Americans ...       News
1      House Intelligence Committee Chairman Devin Nu...       News
2      On Friday, it was revealed that former Milwauk...       News
3      On Christmas day, Donald Trump announced that ...       News
4      Pope Francis used his annual Christmas Day mes...       News
...                                                  ...        ...
44893  BRUSSELS (Reuters) - NATO allies on Tuesday we...  worldnews
44894  LONDON (Reuters) - LexisNexis, a provider of l...  worldnews
44895  MINSK (Reuters) - In the shadow of disused Sov...  worldnews
44896  MOSCOW (Reuters) - Vatican Secretary of State ...  worldnews
44897  JAKARTA (Reuters) - Indonesia will buy 11 Sukh...  worldnews

                    date Label
0       December 31, 2017  Fake
1       December 31, 2017  Fake
2       December 30, 2017  Fake
3       December 29, 2017  Fake
4       December 25, 2017  Fake
...                   ...   ...
44893   August 22, 2017   true
44894   August 22, 2017   true
44895   August 22, 2017   true
44896   August 22, 2017   true
```

5. Used label encoder to add 0 and 1 values to label column, (0 being Fake and 1 is true)

```
In [292]: #Encoder and Imputers  #0 is Fake and 1 is true
          from sklearn.preprocessing import LabelEncoder
          lab_enc=LabelEncoder() #Label econder is used to convert categorical data into numbers, like 0 and 1
          df= lab_enc.fit_transform(df_cat1['Label'])
          pd.Series(df)
          df_cat1['Label'] = df
          df_cat1
```

Out[292]:

	title	text	subject	date	Label
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
...	...	...	...	...	...
44893	'Fully committed' NATO backs new U.S. approach...	BRUSSELS (Reuters) - NATO allies on Tuesday we...	worldnews	August 22, 2017	1
44894	LexisNexis withdrew two products from Chinese ...	LONDON (Reuters) - LexisNexis, a provider of l...	worldnews	August 22, 2017	1
44895	Minsk cultural hub becomes haven from authorities	MINSK (Reuters) - In the shadow of disused Sov...	worldnews	August 22, 2017	1
44896	Vatican upbeat on possibility of Pope Francis ...	MOSCOW (Reuters) - Vatican Secretary of State ...	worldnews	August 22, 2017	1
44897	Indonesia to buy $1.14 billion worth of Russia...	JAKARTA (Reuters) - Indonesia will buy 11 Sukh...	worldnews	August 22, 2017	1

44898 rows × 5 columns

6. Finding out any null values

```
In [293]: df_cat1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 44898 entries, 0 to 44897
Data columns (total 5 columns):
 #   Column   Non-Null Count  Dtype 
---  ------   --------------  ----- 
 0   title    44898 non-null  object
 1   text     44898 non-null  object
 2   subject  44898 non-null  object
 3   date     44898 non-null  object
 4   Label    44898 non-null  int32 
dtypes: int32(1), object(4)
memory usage: 1.5+ MB
```

```
In [294]:
```

7. Finding out if there are any special characters in the object data type

```
In [294]: #finding unique values in object data types
          def explore_object_type(df_cat1,feature_name):
              if df_cat1[feature_name].dtype == 'object':
                  print(df_cat1[feature_name].value_counts())
```

```
In [295]: for featureName in df_cat1:
              if df_cat1[featureName].dtype == 'object':
                  print('\n' + str(featureName) + '\'s" Values with count are :')
                  explore_object_type(df_cat1, str(featureName))
```

```
1
OBAMA STARES DOWN AMERICAN SNIPER Widow Taya Kyle, As CNN Gives Her Time To Confront Him About Gun Control On Live TV [Vide
o]    1
BREAKING: “The Real Donald Trump Story” [VIDEO]
1
Indonesia to buy $1.14 billion worth of Russian jets
1
Name: title, Length: 38729, dtype: int64

text's" Values with count are :

627
(Reuters) - Highlights for U.S. President Donald Trump’s administration on Thursday: The United States drops a massive GBU-4
3 bomb, the largest non-nuclear bomb it has ever used in combat, in Afghanistan against a series of caves used by Islamic St
ate militants, the Pentagon says. Trump says Pyongyang is a problem that “will be taken care of” amid speculation that North
Korea is on the verge of a sixth nuclear test. Military force cannot resolve tension over North Korea, China warns, while an
influential Chinese newspaper urges Pyongyang to halt its nuclear program in exchange for Beijing’s protection. The Trump ad
ministration is focusing its North Korea strategy on tougher economic sanctions, possibly including intercepting cargo ships
and punishing Chinese banks doing business with Pyongyang, U.S. officials say. Trump says “things will work out fine” betwee
```

8.I have dropped the date column as it is a challenge to deal with and there are many special character and unique values to calculate

```
In [298]: # dropping the date column
          df_cat1.drop('date',axis=1,inplace=True)
```

```
In [299]: df_cat1.shape
```

```
Out[299]: (44898, 4)
```

9.replaced any missing values with blank values

```
In [301]: df_cat1 = df_cat1.fillna(' ')

In [302]: df_cat1.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 44898 entries, 0 to 44897
Data columns (total 4 columns):
 #   Column   Non-Null Count  Dtype 
---  ------   --------------  ----- 
 0   title    44898 non-null  object
 1   text     44898 non-null  object
 2   subject  44898 non-null  object
 3   Label    44898 non-null  int32 
dtypes: int32(1), object(3)
memory usage: 1.2+ MB
```

10. Counts in the label column

```
In [303]: df_cat1["Label"].value_counts()

Out[303]: 0    23481
          1    21417
          Name: Label, dtype: int64
```

11. Finding the value counts in text

```
In [304]: df_cat1['text'][5]

Out[304]: 'The number of cases of cops brutalizing and killing people of color seems to see no end. Now, we have another case that needs
          to be shared far and wide. An Alabama woman by the name of Angela Williams shared a graphic photo of her son, lying in a hospit
          al bed with a beaten and fractured face, on Facebook. It needs to be shared far and wide, because this is unacceptable.It is un
          clear why Williams  son was in police custody or what sort of altercation resulted in his arrest, but when you see the photo yo
          u will realize that these details matter not. Cops are not supposed to beat and brutalize those in their custody. In the post y
          ou are about to see, Ms. Williams expresses her hope that the cops had their body cameras on while they were beating her son, b
          ut I think we all know that there will be some kind of convenient  malfunction  to explain away the lack of existence of dash o
          r body camera footage of what was clearly a brutal beating. Hell, it could even be described as attempted murder. Something tel
          ls me that this young man will never be the same. Without further ado, here is what Troy, Alabama s finest decided was appropri
          ate treatment of Angela Williams  son:No matter what the perceived crime of this young man might be, this is completely unaccep
          table. The cops who did this need to rot in jail for a long, long time  but what you wanna bet they get a paid vacation while
          the force  investigates  itself, only to have the officers returned to duty posthaste?This, folks, is why we say BLACK LIVES MA
          TTER. No way in hell would this have happened if Angela Williams  son had been white. Please share far and wide, and stay tuned
          to Addicting Info for further updates.Featured image via David McNew/Stringer/Getty Images'

In [305]: df_cat1.value_counts()

          ge, according to an email seen by Reuters. Trump will seek to rebuild the U.S. relationship with Egypt at a meeting on Monda
          y with Egyptian President Abdel Fattah al-Sisi focused on security issues and military aid, a senior White House official sa
          ys. Trump will host Jordan's King Abdullah at the White House next week to discuss the fight against Islamic State militant
          s, the Syria crisis and advancing peace between Israelis and Palestinians, the White House says. A U.S. judge approves a $25
          million settlement to resolve a class action lawsuit that claimed fraud against Trump and his Trump University real estate s
          eminars.
          politicsNews  1       5
          Highlights: The Trump presidency on April 26 at 9:12 P.M. EDT/0112 GMT on April 27                                    (Reut
          ers) - Highlights for U.S. President Donald Trump's administration on Wednesday: Trump proposes slashing tax rates for busin
          esses and on overseas corporate profits returned to the country in a plan greeted as an opening gambit by his fellow Republi
          cans in Congress. Trump's plan could shift the U.S. economy into higher gear but could have one effect the White House would
          not welcome – interest rates ratcheted higher than expected by a wary central bank. The Trump tax cut will generate growth b
          ut not nearly enough to replace trillions of dollars in lost revenues, while rising deficits could even take back some of th
          e economic gains, fiscal experts say. Congress inches toward a deal to fund the government through September but is preparin
          g to possibly extend a midnight Friday deadline in order to wrap up negotiations and avoid an imminent government shutdown.
          Trump is considering issuing an executive order to pull the United States from the North American Free Trade Agreement, an a
          dministration official says, a move that could unravel one of the world's biggest trading blocs. Trump and Canadian Prime Mi
          nister Justin Trudeau discuss bilateral trade in their second conversation in as many days  amid strains over softwood lumbe
          r and dairy. The Trump administration says it aims to push North Korea into dismantling its nuclear and missile programs thr
```

12. Checked for 209 duplicate value and removed them

```
In [306]: #Checking for Duplicates and Dropping them
          df_cat1.duplicated().sum() #there are 209 duplicate rows

Out[306]: 213

In [307]: #dropping duplicate rows
          df_cat1.drop_duplicates(inplace=True)
          df_cat1.duplicated().sum()

Out[307]: 0
```

13. Working with regular expressions and removing characters

```
dtype: int64

In [309]: import re

In [310]: # expanding English Language contractions: https://stackoverflow.com/a/47091490/4084039
          def decontracted(phrase):
              # specific
              phrase = re.sub(r"won\'t", "will not", phrase)
              phrase = re.sub(r"can\'t", "can not", phrase)

              # general
              phrase = re.sub(r"n\'t", " not", phrase)
              phrase = re.sub(r"\'re", " are", phrase)
              phrase = re.sub(r"\'s", " is", phrase)
              phrase = re.sub(r"\'d", " would", phrase)
              phrase = re.sub(r"\'ll", " will", phrase)
              phrase = re.sub(r"\'t", " not", phrase)
              phrase = re.sub(r"\'ve", " have", phrase)
              phrase = re.sub(r"\'m", " am", phrase)
              return phrase
```

14. Importing stopwords and word net lemmatizer from nltk

```
311]: import nltk
      from nltk.corpus import stopwords
      from nltk.stem import WordNetLemmatizer

312]: nltk.download('stopwords')

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

312]: True

313]: nltk.download('punkt')

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

313]: True

314]: stop_words=stopwords.words('english')

315]: lemmatizer=WordNetLemmatizer()
```

15. Import tqm library

```
In [316]: #https://stackoverflow.com/a/5843547/4084039
          #Title
          from tqdm import tqdm
          preprocessed_titles = []
          #tqdm is for printing the status bar
          for sentence in tqdm(df_cat1['title'].values):
              sent = decontracted(sentence)
              sent = re.sub(r'https?:\/\/.*[\r\n]*', '', sent) # remove hyperlinks
              sent=re.sub('[^A-Za-z]+',' ', sent) #remove spacial character, numbers:
              sent = ' '.join(e for e in sent.split() if e not in stop_words) #removing stop words
              sent=' '.join(lemmatizer.lemmatize(e) for e in sent.split()) # Lemmatization
              preprocessed_titles.append(sent.lower().strip())

          100%|██████████| 44685/44685 [00:15<00:00, 2812.51it/s]

In [317]: df_cat1['titles']=preprocessed_titles

In [318]: df_cat1['titles']

Out[318]: 0        donald trump sends out embarrassing new year e...
          1        drunk bragging trump staffer started russian c...
          2        sheriff david clarke becomes an internet joke ...
          3        trump is so obsessed he even has obama name co...
          4        pope francis just called out donald trump duri...
                                       ...
          44893    fully committed nato back new u s approach afg...
          44894       lexisnexis withdrew two product chinese market
          44895                  minsk cultural hub becomes authority
          44896    vatican upbeat possibility pope francis visiti...
          44897              indonesia buy billion worth russian jet
          Name: titles, Length: 44685, dtype: object
```

16. Working with title column

```
In [317]: df_cat1['titles']=preprocessed_titles

In [318]: df_cat1['titles']

Out[318]: 0        donald trump sends out embarrassing new year e...
          1        drunk bragging trump staffer started russian c...
          2        sheriff david clarke becomes an internet joke ...
          3        trump is so obsessed he even has obama name co...
          4        pope francis just called out donald trump duri...
                                       ...
          44893    fully committed nato back new u s approach afg...
          44894       lexisnexis withdrew two product chinese market
          44895                  minsk cultural hub becomes authority
          44896    vatican upbeat possibility pope francis visiti...
          44897              indonesia buy billion worth russian jet
          Name: titles, Length: 44685, dtype: object

In [319]: #TEXT
          preprocessed_texts = []
          #tqdm is for printing the status bar
          for sentence in tqdm (df_cat1['text'].values):
              sent = decontracted(sentence)
              sent = re.sub(r'https?:\/\/.*[\r\n]*', '', sent) # remove hyperlinks
              sent=re.sub('[^A-Za-z]+',' ', sent) #remove spacial character, numbers:
              sent = ' '.join(e for e in sent.split() if e not in stop_words) #removing stop words
              sent=' '.join(lemmatizer.lemmatize(e) for e in sent.split()) # Lemmatization
              preprocessed_texts.append(sent.lower().strip())

          100%|██████████| 44685/44685 [03:40<00:00, 202.89it/s]
```

17. Working with text column

```
In [319]: #TEXT
preprocessed_texts = []
#tqdm is for printing the status bar
for sentence in tqdm (df_cat1['text'].values):
    sent = decontracted(sentence)
    sent = re.sub(r'https?:\/\/.*[\r\n]*', '', sent) # remove hyperlinks
    sent=re.sub('[^A-Za-z]+',' ', sent) #remove spacial character, numbers:
    sent = ' '.join(e for e in sent.split() if e not in stop_words) #removing stop words
    sent=' '.join(lemmatizer.lemmatize(e) for e in sent.split()) # Lemmatization
    preprocessed_texts.append(sent.lower().strip())

100%|██████████| 44685/44685 [03:40<00:00, 202.89it/s]

In [320]: df_cat1['text']= preprocessed_texts

In [321]: df_cat1['text']

Out[321]: 0        donald trump wish americans happy new year lea...
          1        house intelligence committee chairman devin nu...
          2        on friday revealed former milwaukee sheriff da...
          3        on christmas day donald trump announced would ...
          4        pope francis used annual christmas day message...
                                       ...
          44893    brussels reuters nato ally tuesday welcomed pr...
          44894    london reuters lexisnexis provider legal regul...
          44895    minsk reuters in shadow disused soviet era fac...
          44896    moscow reuters vatican secretary state cardina...
          44897    jakarta reuters indonesia buy sukhoi fighter j...
          Name: text, Length: 44685, dtype: object
```

18. Processing for subject column

```
[223]: #subject
preprocessed_subject = []
#tqdm is for printing the status bar
for sentence in tqdm (df_cat1['subject'].values):
    sent = decontracted(sentence)
    sent = re.sub(r'https?:\/\/.*[\r\n]*', '', sent) # remove hyperlinks
    sent=re.sub('[^A-Za-z]+',' ', sent) #remove spacial character, numbers:
    sent = ' '.join(e for e in sent.split() if e not in stop_words) #removing stop words
    sent=' '.join(lemmatizer.lemmatize(e) for e in sent.split()) # Lemmatization
    preprocessed_subject.append(sent.lower().strip())

100%|██████████| 44685/44685 [00:02<00:00, 15252.64it/s]

[224]: df_cat1['subject']= preprocessed_subject

[225]: df_cat1['subject']

t[225]: 0            news
        1            news
        2            news
        3            news
        4            news
                    ...
        44893    worldnews
        44894    worldnews
        44895    worldnews
        44896    worldnews
        44897    worldnews
        Name: subject, Length: 44685, dtype: object
```

19. Splitting the data into train and test

In [226]:
```
#Splitting data into train, cv, test
from sklearn.model_selection import train_test_split
y=df_cat1["Label"]
x = df_cat1.drop(columns={'Label'})
```

In [227]:
```
train, test, train_output, test_output= train_test_split(x,
                                                         y,test_size=0.25,
                                                         stratify=y,
                                                         random_state=0)
train_modified, cv, train_output_modified, cv_output = train_test_split(train,
                                                                        train_output,
                                                                        test_size=0.25,
                                                                        stratify=train_output,
                                                                        random_state=0)
```

In [228]:
```
train.shape, cv.shape, test.shape
```

Out[228]:
```
((33513, 4), (8379, 4), (11172, 4))
```

In [229]:
```
train_output.shape, cv_output.shape, test_output.shape
```

Out[229]:
```
((33513,), (8379,), (11172,))
```

20. Using TDF vectorization

]0]:
```
#data encoding
#title-TFIDF Vectorization
from sklearn.feature_extraction.text import TfidfVectorizer
title_tfidf_vectorizer = TfidfVectorizer(min_df=5)
```

]1]:
```
train_title_tfidf=title_tfidf_vectorizer.fit_transform(train['title'].values)
```

]2]:
```
cv_title_tfidf=title_tfidf_vectorizer.transform(cv['title'].values)
test_title_tfidf=title_tfidf_vectorizer.transform(test['title'].values)
```

]3]:
```
#saving the tfidf vectorizer
import pickle
with open("title_tfidf_vectorizer.pickle", "wb") as fp:
    pickle.dump(title_tfidf_vectorizer, fp, protocol=pickle.HIGHEST_PROTOCOL)
    title_tfidf_vectorizer.get_feature_names() [:10]
```

```
C:\Users\Office\anaconda3\lib\site-packages\sklearn\utils\deprecation.py:87: FutureWarning: Function get_feature_names is depre
cated; get_feature_names is deprecated in 1.0 and will be removed in 1.2. Please use get_feature_names_out instead.
  warnings.warn(msg, category=FutureWarning)
```

]4]:
```
title_tfidf_vectorizer.get_feature_names()[:10]
```

]4]:
```
['000', '04', '10', '100', '100k', '101', '11', '12', '120', '13']
```

21. Data encoding for text column with TDF

```
#data encoding
#title-TFIDF Vectorization
from sklearn.feature_extraction.text import TfidfVectorizer
text_tfidf_vectorizer = TfidfVectorizer(min_df=10)
```

```
train_text_tfidf=text_tfidf_vectorizer.fit_transform(train['text'].values)
```

```
cv_text_tfidf=text_tfidf_vectorizer.transform(cv['text'].values)
test_text_tfidf=text_tfidf_vectorizer.transform(test['text'].values)
```

```
#saving the tfidf vectorizer
import pickle
with open("text_tfidf_vectorizer.pickle", "wb") as fp:
    pickle.dump(text_tfidf_vectorizer, fp, protocol=pickle.HIGHEST_PROTOCOL)
    text_tfidf_vectorizer.get_feature_names() [:10]
```

```
text_tfidf_vectorizer.get_feature_names()[:10]
```

```
['00', '000', '0000', '001', '002', '003', '005', '005380', '00pm', '01']
```

22. Using TDF use multinomial classifier. The accuracy is 93%

```
#1. Convert text into vectors using TF-IDF
#2. Instantiate MultinomialNB classifier
#3. Split feature and Label
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
tf_vec = TfidfVectorizer()
naive=MultinomialNB()
features=tf_vec.fit_transform(df_cat1['text'])
x = features
y = df_cat1['Label']
```

```
# Train and predict
x_train,x_test, y_train,y_test=train_test_split(x,y,random_state=42)
naive.fit(x_train,y_train)
y_pred=naive.predict(x_test)
print('Final score = > ', accuracy_score(y_test,y_pred))
```

```
Final score = >  0.9327219901942981
```

```
print (classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.93      0.94      0.94      5702
           1       0.94      0.92      0.93      5312

    accuracy                           0.93     11014
   macro avg       0.93      0.93      0.93     11014
weighted avg       0.93      0.93      0.93     11014
```

Conclusion:

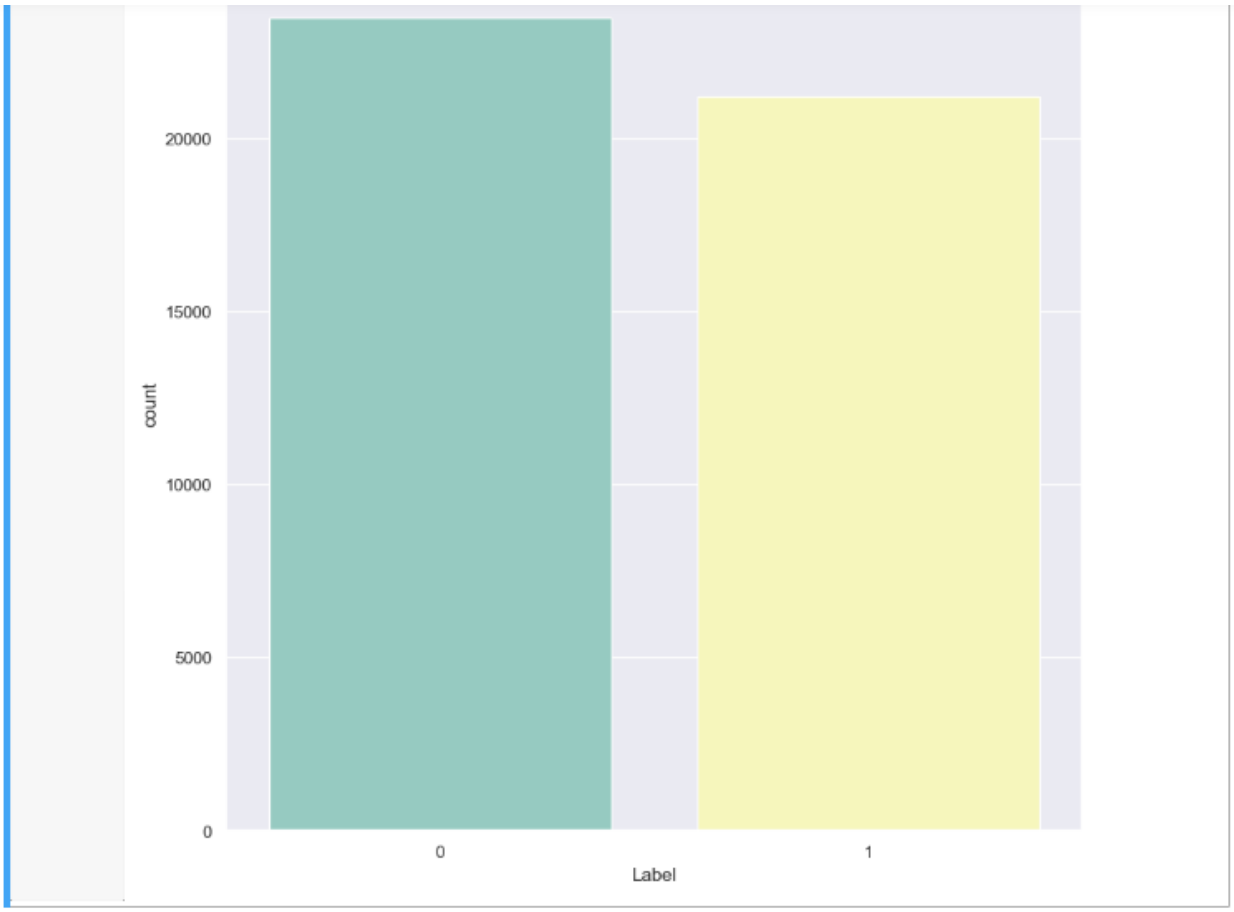
1. Finding out the unique value in object type column

```
df_cat1.describe(include=['object','datetime']).transpose()
```

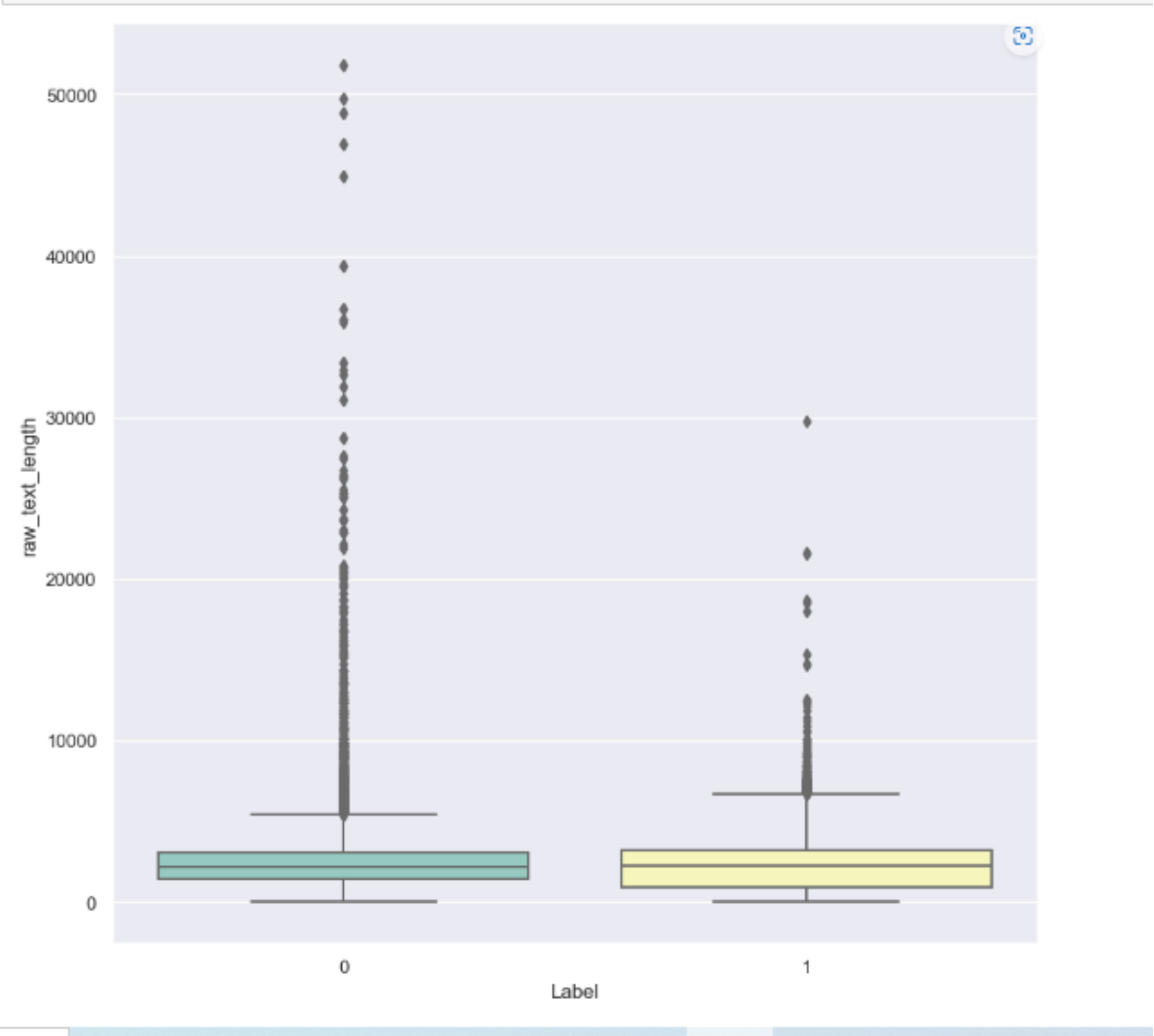
	count	unique	top	freq
title	44898	38729	Factbox: Trump fills top jobs for his administ...	14
text	44898	38646		627
subject	44898	8	politicsNews	11272

There are 38729 unique values and they occur in frequency of 14 for title
There are v

2. Dividing labels into 0 and 1. (0 being Fake and 1 is true). Around 2800 for Fake and 2300 for true

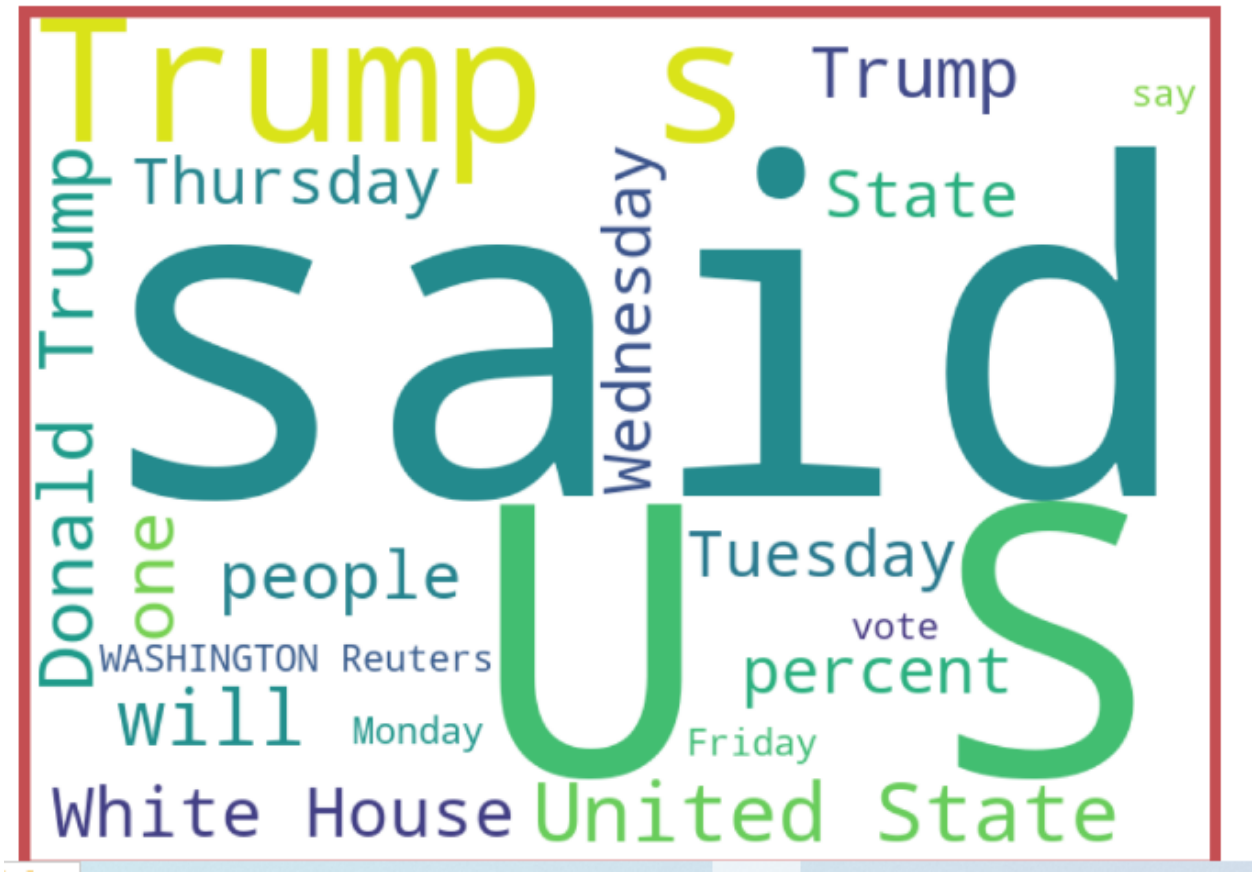


3. Raw text length for both labels 0 and 1 . (0 being Fake and 1 is true)
   5500 for Fake and around 3000 for true

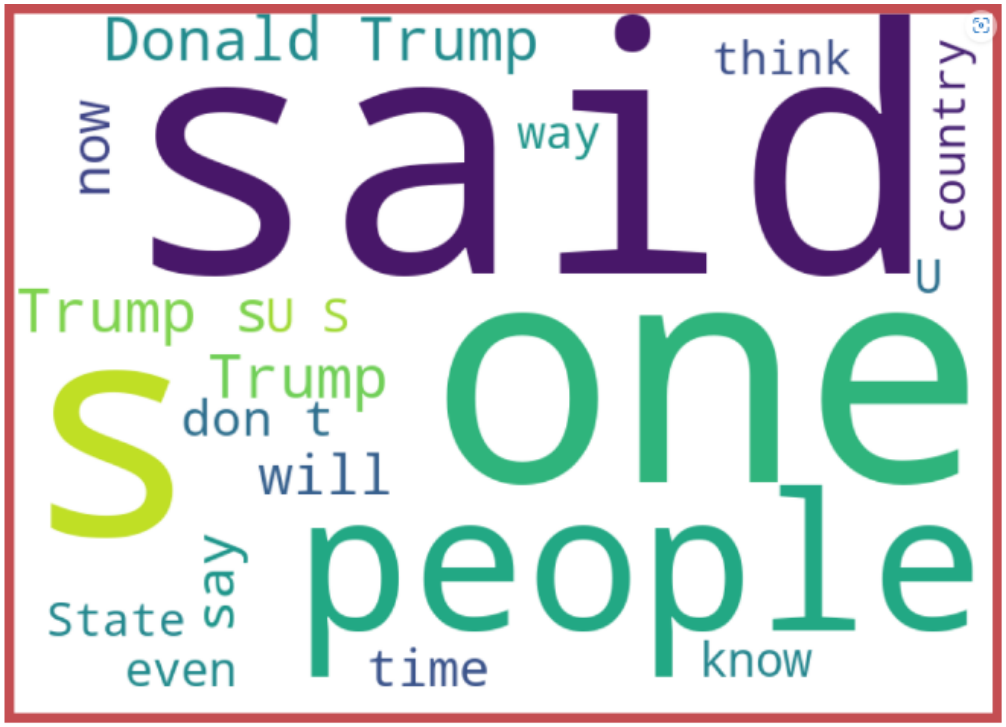


4. Using word cloud, some common word cloud words for true label ie 1 include:
   Trump, said, Thursday, White House, United states, Vote, Donald Trump, State

```
#Getting sense of word cloud in spam
from wordcloud import WordCloud
spams=df_cat1['text'][df_cat1['Label']==1]
spam_cloud = WordCloud(width=700,height=500, background_color='white', max_words=20).generate(' '.join(spams))
plt.figure(figsize=(10,8), facecolor='r')
plt.imshow(spam_cloud)
plt.axis('off')
plt.tight_layout(pad=0)
plt.show()
```



5. Using word cloud, some common word cloud words for Fake label ie 0 include: Said, One, people Trump, Donald Trump, dont, will, state, say, time , know



5]:

6. Accuracy with multinomial classifier. Accuracy is 93%

5]:
```
#1. Convert text into vectors using TF-IDF
#2. Instantiate MultinomialNB classifier
#3. Split feature and Label
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
tf_vec = TfidfVectorizer()
naive=MultinomialNB()
features=tf_vec.fit_transform(df_cat1['text'])
x = features
y = df_cat1['Label']
```

5]:
```
# Train and predict
x_train,x_test, y_train,y_test=train_test_split(x,y,random_state=42)
naive.fit(x_train,y_train)
y_pred=naive.predict(x_test)
print('Final score = > ', accuracy_score(y_test,y_pred))
```

```
Final score = >  0.9327219901942981
```

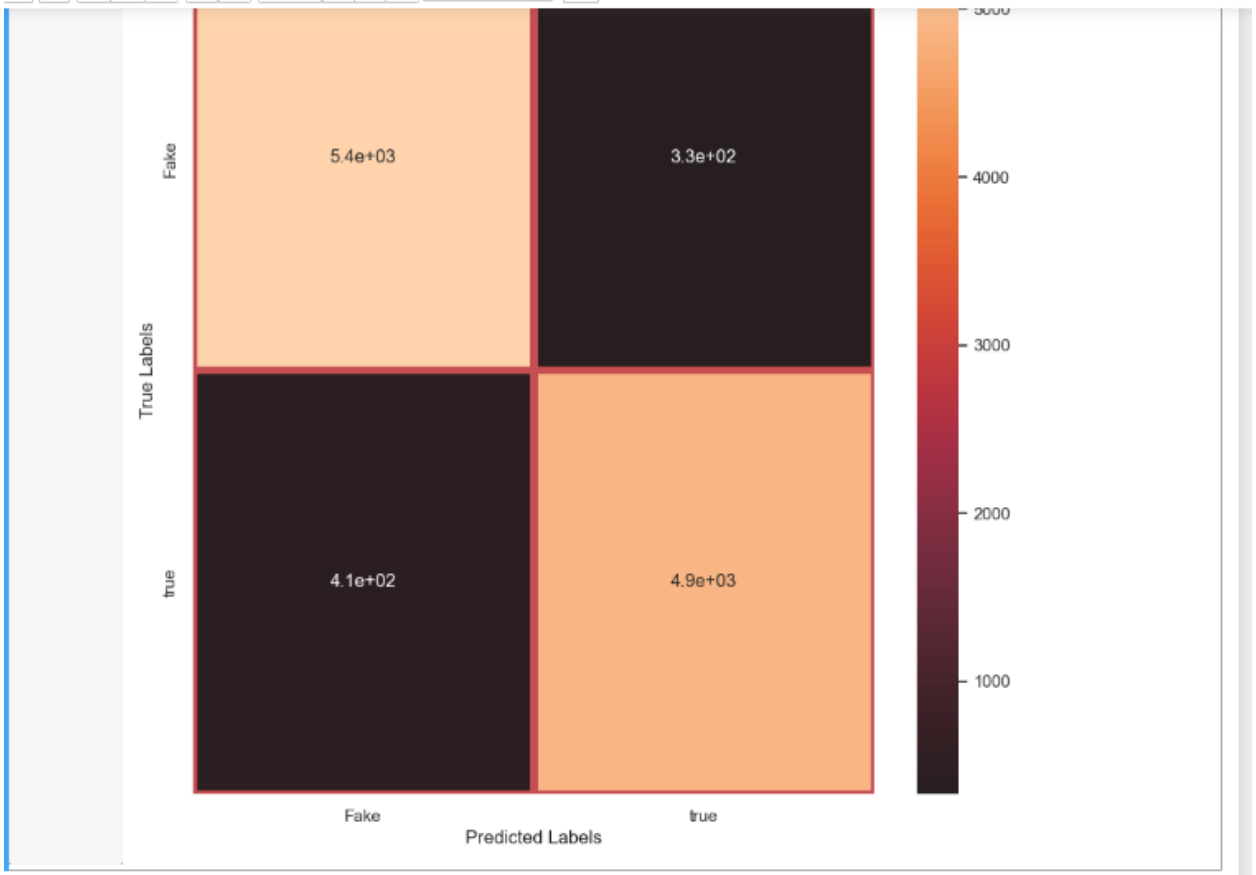
7]:
```
print (classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.93      0.94      0.94      5702
           1       0.94      0.92      0.93      5312

    accuracy                           0.93     11014
   macro avg       0.93      0.93      0.93     11014
weighted avg       0.93      0.93      0.93     11014
```

7. Using the correlation matrix, plotting the True and predicted values




```
In [269]: conf_mat

Out[269]: array([[5375,  327],
                 [ 414, 4898]], dtype=int64)
```