Importing necessary libraries.

In [ ]:

```
import  pandas as pd
import numpy as np
import statsmodels.api as sm
import matplotlib.pyplot as plt
import seaborn as sns
sns.set()
import sklearn
```

Loading data set into raw_data variable.

In [38]:

```
raw_data=pd.read_csv('Desktop/news.csv')
```

Displaying the content of the data set 'news.csv'.

In [39]:

```
raw_data
```

Out[39]:

	Unnamed: 0	title	text	label
0	8476	You Can Smell Hillary’s Fear	Daniel Greenfield, a Shillman Journalism Fello...	FAKE
1	10294	Watch The Exact Moment Paul Ryan Committed Pol...	Google Pinterest Digg Linkedin Reddit Stumbleu...	FAKE
2	3608	Kerry to go to Paris in gesture of sympathy	U.S. Secretary of State John F. Kerry said Mon...	REAL
3	10142	Bernie supporters on Twitter erupt in anger ag...	— Kaydee King (@KaydeeKing) November 9, 2016 T...	FAKE
4	875	The Battle of New York: Why This Primary Matters	It's primary day in New York and front-runners...	REAL
...	...	...	...	...
6330	4490	State Department says it can't find emails fro...	The State Department told the Republican Natio...	REAL
6331	8062	The ‘P’ in PBS Should Stand for ‘Plutocratic’ ...	The ‘P’ in PBS Should Stand for ‘Plutocratic’ ...	FAKE
6332	8622	Anti-Trump Protesters Are Tools of the Oligarc...	Anti-Trump Protesters Are Tools of the Oligar...	FAKE
6333	4021	In Ethiopia, Obama seeks progress on peace, se...	ADDIS ABABA, Ethiopia —President Obama convene...	REAL
6334	4330	Jeb Bush Is Suddenly Attacking Trump. Here's W...	Jeb Bush Is Suddenly Attacking Trump. Here's W...	REAL

6335 rows × 4 columns

Displaying the shape of the data set.

In [4]:

```
raw_data.shape
```

Out[4]:

```
(6335, 4)
```

Describing the data set like its count, mean, min, max, etc.

In [5]:

```
raw_data.describe()
```

Out[5]:

	Unnamed: 0
count	6335.000000
mean	5280.415627
std	3038.503953
min	2.000000
25%	2674.500000
50%	5271.000000
75%	7901.000000
max	10557.000000

Displaying only the top 5 row using .head command.

In [6]:

```
raw_data.head()
```

Out[6]:

	Unnamed: 0	title	text	label
0	8476	You Can Smell Hillary's Fear	Daniel Greenfield, a Shillman Journalism Fello...	FAKE
1	10294	Watch The Exact Moment Paul Ryan Committed Pol...	Google Pinterest Digg Linkedin Reddit Stumbleu...	FAKE
2	3608	Kerry to go to Paris in gesture of sympathy	U.S. Secretary of State John F. Kerry said Mon...	REAL
3	10142	Bernie supporters on Twitter erupt in anger ag...	— Kaydee King (@KaydeeKing) November 9, 2016 T...	FAKE
4	875	The Battle of New York: Why This Primary Matters	It's primary day in New York and front-runners...	REAL

Displaying the last 5 rows using .tail.

In [7]:

```
raw_data.tail()
```

Out[7]:

	Unnamed: 0	title	text	label
6330	4490	State Department says it can't find emails fro...	The State Department told the Republican Natio...	REAL
6331	8062	The ‘P’ in PBS Should Stand for ‘Plutocratic’ ...	The ‘P’ in PBS Should Stand for ‘Plutocratic’ ...	FAKE
6332	8622	Anti-Trump Protesters Are Tools of the Oligarc...	Anti-Trump Protesters Are Tools of the Oligar...	FAKE
6333	4021	In Ethiopia, Obama seeks progress on peace, se...	ADDIS ABABA, Ethiopia —President Obama convene...	REAL
6334	4330	Jeb Bush Is Suddenly Attacking Trump. Here's W...	Jeb Bush Is Suddenly Attacking Trump. Here's W...	REAL

Dropping the field 'Unnamed:0' as its irrelevant index field.

In [8]:

```
data=raw_data.drop(['Unnamed: 0'],axis=1)
```

In [9]:

```
data
```

Out[9]:

	title	text	label
0	You Can Smell Hillary's Fear	Daniel Greenfield, a Shillman Journalism Fello...	FAKE
1	Watch The Exact Moment Paul Ryan Committed Pol...	Google Pinterest Digg Linkedin Reddit Stumbleu...	FAKE
2	Kerry to go to Paris in gesture of sympathy	U.S. Secretary of State John F. Kerry said Mon...	REAL
3	Bernie supporters on Twitter erupt in anger ag...	— Kaydee King (@KaydeeKing) November 9, 2016 T...	FAKE
4	The Battle of New York: Why This Primary Matters	It's primary day in New York and front-runners...	REAL
...	...	...	...
6330	State Department says it can't find emails fro...	The State Department told the Republican Natio...	REAL
6331	The 'P' in PBS Should Stand for 'Plutocratic' ...	The 'P' in PBS Should Stand for 'Plutocratic' ...	FAKE
6332	Anti-Trump Protesters Are Tools of the Oligarc...	Anti-Trump Protesters Are Tools of the Oligar...	FAKE
6333	In Ethiopia, Obama seeks progress on peace, se...	ADDIS ABABA, Ethiopia —President Obama convene...	REAL
6334	Jeb Bush Is Suddenly Attacking Trump. Here's W...	Jeb Bush Is Suddenly Attacking Trump. Here's W...	REAL

6335 rows × 3 columns

Declaring a variable y with 'label' field, on the basis of y we will build our model. As the field y depicts whether a news is fake or real.

In [10]:

```
y=data['label']
y
```

Out[10]:

```
0       FAKE
1       FAKE
2       REAL
3       FAKE
4       REAL
        ...
6330    REAL
6331    FAKE
6332    FAKE
6333    REAL
6334    REAL
Name: label, Length: 6335, dtype: object
```

Dropping the column 'label' as it is no longer required in data variable as we have saved 'label' as 'y' variable.

In [11]:

```
data.drop(['label'],axis=1)
```

Out[11]:

	title	text
0	You Can Smell Hillary’s Fear	Daniel Greenfield, a Shillman Journalism Fello...
1	Watch The Exact Moment Paul Ryan Committed Pol...	Google Pinterest Digg Linkedin Reddit Stumbleu...
2	Kerry to go to Paris in gesture of sympathy	U.S. Secretary of State John F. Kerry said Mon...
3	Bernie supporters on Twitter erupt in anger ag...	— Kaydee King (@KaydeeKing) November 9, 2016 T...
4	The Battle of New York: Why This Primary Matters	It's primary day in New York and front-runners...
...	...	...
6330	State Department says it can't find emails fro...	The State Department told the Republican Natio...
6331	The ‘P’ in PBS Should Stand for ‘Plutocratic’ ...	The ‘P’ in PBS Should Stand for ‘Plutocratic’ ...
6332	Anti-Trump Protesters Are Tools of the Oligarc...	Anti-Trump Protesters Are Tools of the Oligar...
6333	In Ethiopia, Obama seeks progress on peace, se...	ADDIS ABABA, Ethiopia —President Obama convene...
6334	Jeb Bush Is Suddenly Attacking Trump. Here's W...	Jeb Bush Is Suddenly Attacking Trump. Here's W...

6335 rows × 2 columns

Importing relevant libraries from sklearn. Spliting the data sets into 'train' and test' helps us to split our data set into two sets in a specific ratio, so that we can train our model on the basis of 'train' set and then apply the model on 'test' set to predict the accuracy of our model we have built. Making two unique sets each time we run the model is assured by giving it a random state value. So everytime the model is run, 'train' and 'test' sets are mutually exclusive to each other every time also the data within them is randomly arrange but the data between two sets are never mixed, no matter how many times we run the model. Random state takes care of that. Usually its recmmended that the split is 80-20, 80%of original data set is given to train set and 20% to test data set. But here we have make it 3:1, to build more robust model.

In [12]:

```
from sklearn.model_selection import train_test_split
```

In [13]:

```
x_train, x_test, y_train, y_test = train_test_split(data['text'], y, test_size=0.33, ra
ndom_state=53)
```

In [14]:

```
x_train
```

Out[14]:

```
2576
1539    Report Copyright Violation Do you think there ...
5163    The election in 232 photos, 43 numbers and 131...
2615    Email Ever wonder what’s on the mind of today’...
4270    Wells Fargo is Rotting from the Top Down Wells...
                              ...
662     —Debby Borza stood before a wall of photos of ...
3261    Presumptive Republican nominee Donald Trump ha...
5883    December's job growth numbers are in, and they...
2933    In a wide-ranging discussion, Trump also said ...
797     Top officials of the Cruz campaign are convinc...
Name: text, Length: 4244, dtype: object
```

In [15]:

```
x_train.shape
```

Out[15]:

```
(4244,)
```

In [16]:

```
x_test.shape
```

Out[16]:

```
(2091,)
```

In [17]:

```
x_test
```

Out[17]:

```
4221    Donald Trump threatened to sue the New York Ti...
1685    Planned Parenthood: Abortion pill usage now ri...
3348    In a last dash, final "hail mary" attempt to e...
2633    Washington (CNN) Donald Trump and Ben Carson n...
975     The Obama administration announced Friday it w...
                              ...
3888    In a marketing fiasco that could rank right up...
2015    Email \nThe Politico/Morning Consult Poll find...
5860    The Maryland Democrat made the announcement Mo...
3071    Prev post Page 1 of 4 Next \nWhen most people ...
4284    The Bushes are burning as they consume the new...
Name: text, Length: 2091, dtype: object
```

In [18]:

```
y_train.shape
```

Out[18]:

```
(4244,)
```

In [19]:

```
y_train
```

Out[19]:

```
2576    FAKE
1539    FAKE
5163    REAL
2615    FAKE
4270    FAKE
        ...
662     REAL
3261    REAL
5883    REAL
2933    REAL
797     REAL
Name: label, Length: 4244, dtype: object
```

In [20]:

```
y_test.shape
```

Out[20]:

```
(2091,)
```

In [21]:

```
y_test
```

Out[21]:

```
4221    REAL
1685    FAKE
3348    REAL
2633    REAL
975     REAL
        ...
3888    REAL
2015    FAKE
5860    REAL
3071    FAKE
4284    REAL
Name: label, Length: 2091, dtype: object
```

Importing relevant libraries from sklearn ,'TfidfVectorizer' by using it we are fiting our model's train set and transforming into 'tfidf_train' set. Similarly fiting test data into 'tfidf_test'.

In [22]:

```
from sklearn.feature_extraction.text import TfidfVectorizer
```

In [23]:

```
tfidf_vectorizer = TfidfVectorizer(stop_words='english', max_df=0.7)
```

In [24]:

```
tfidf_train = tfidf_vectorizer.fit_transform(x_train)
```

In [25]:

```
tfidf_train
```

Out[25]:

```
<4244x56922 sparse matrix of type '<class 'numpy.float64'>'
        with 1119820 stored elements in Compressed Sparse Row format>
```

In [26]:

```
tfidf_test = tfidf_vectorizer.transform(x_test)
```

In [27]:

```
tfidf_test
```

Out[27]:

```
<2091x56922 sparse matrix of type '<class 'numpy.float64'>'
        with 533697 stored elements in Compressed Sparse Row format>
```

In [28]:

```
print(tfidf_vectorizer.get_feature_names()[-10:])
```

['والمرضى', 'هذا', 'من', 'محاولات', 'ما', 'لم', 'عن', 'عربي', 'حلب', 'ﻮade']

Importing relevant libraries from sklearn so that we can use 'Passive Aggressive Classifier', so that we can finally test our model on 'TfidfVectorizer' and at the end calling the confusion matrix and accuracy_score to finally check how well our model is, by evaluating the confusion_matrix and the accuracy being predicted.

In [36]:

```
from sklearn.linear_model import PassiveAggressiveClassifier as pc
from sklearn.metrics import accuracy_score, confusion_matrix
```

In [30]:

```
clf=pc()
```

In [31]:

```
clf.fit(tfidf_train,y_train)
```

Out[31]:

```
PassiveAggressiveClassifier(C=1.0, average=False, class_weight=None,
                            early_stopping=False, fit_intercept=True,
                            loss='hinge', max_iter=1000, n_iter_no_change=
5,
                            n_jobs=None, random_state=None, shuffle=True,
                            tol=0.001, validation_fraction=0.1, verbose=0,
                            warm_start=False)
```

In [40]:

```
pred = clf.predict(tfidf_test)
score =accuracy_score(y_test, pred)
score
```

Out[40]:

```
0.9354375896700143
```

Above is the Accuracy and Below is the confusion matrix of our model.

In [42]:

```
confusion_matrix(y_test,pred)
```

Out[42]:

```
array([[ 952,   56],
       [  79, 1004]], dtype=int64)
```

CONCLUSION:

Accuracy of our model is 93.78%, means approx 94% of the time the model predicted the news fake/real. Which we can consider it to be a successful model but needs more variation of methods. Stll, it was able to predict 93.78% of the time based on our train data set tested on our test data set.

Confusion matrix shows us:

1) 952 times it predicted 'FAKE' news while it was a 'FAKE' news, though it predicted wrong 56 times and predicted 'REAL' although the news was 'FAKE'.

2) 1004 times it predicted 'REAL' news while it was a 'REAL' news, though it predicted wrong 79 times and predicted 'FAKE' although the news was 'REAL'.