Fake News Detection

Dataset: https://www.kaggle.com/c/fake-news/data# (https://www.kaggle.com/c/fake-news/data)

In [1]:
```
1 import pandas as pd
```

In [2]:
```
1 df=pd.read_csv("D:/fakenews_data/train.csv")
```

In [3]:
```
df
```

Out[3]:

	id	title	author	text	label
0	0	House Dem Aide: We Didn’t Even See Comey’s Let...	Darrell Lucus	House Dem Aide: We Didn’t Even See Comey’s Let...	1
1	1	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...	0
2	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, ...	1
3	3	15 Civilians Killed In Single US Airstrike Hav...	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstr...	1
4	4	Iranian woman jailed for fictional unpublished...	Howard Portnoy	Print \nAn Iranian woman has been sentenced to...	1
...	...	...	...	...	...
20795	20795	Rapper T.I.: Trump a ’Poster Child For White S...	Jerome Hudson	Rapper T. I. unloaded on black celebrities who...	0
20796	20796	N.F.L. Playoffs: Schedule, Matchups and Odds -...	Benjamin Hoffman	When the Green Bay Packers lost to the Washing...	0
20797	20797	Macy’s Is Said to Receive Takeover Approach by...	Michael J. de la Merced and Rachel Abrams	The Macy’s of today grew from the union of sev...	0
20798	20798	NATO, Russia To Hold Parallel Exercises In Bal...	Alex Ansary	NATO, Russia To Hold Parallel Exercises In Bal...	1
20799	20799	What Keeps the F-35 Alive	David Swanson	David Swanson is an author, activist, journa...	1

20800 rows × 5 columns

In [4]:
```
# we will consider title column
# to get the Independent Features

X=df.drop('label',axis=1)
```

In [5]:
```
1 X.head()
```

Out[5]:

	id	title	author	text
0	0	House Dem Aide: We Didn’t Even See Comey’s Let...	Darrell Lucus	House Dem Aide: We Didn’t Even See Comey’s Let...
1	1	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...
2	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, ...
3	3	15 Civilians Killed In Single US Airstrike Hav...	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstr...
4	4	Iranian woman jailed for fictional unpublished...	Howard Portnoy	Print \nAn Iranian woman has been sentenced to...

In [6]:
```
1 ## Get the Dependent features
2 y=df['label']
```

In [7]:
```
1 y.head()
```

Out[7]:
```
0    1
1    0
2    1
3    1
4    1
Name: label, dtype: int64
```

In [8]:
```
1 df.shape
```

Out[8]:
```
(20800, 5)
```

In [9]:
```
1 #CountVectorizer for bags of words
2 from sklearn.feature_extraction.text import CountVectorizer
```

In [10]:
```
1 df.isna().sum()
```

Out[10]:
```
id          0
title     558
author   1957
text       39
label       0
dtype: int64
```

In [11]:
```
1 df=df.dropna()
```

In [12]:
```
1 df.shape
```

Out[12]:
```
(18285, 5)
```

In [13]:
```
1 df.head(10)
```

Out[13]:

	id	title	author	text	label
0	0	House Dem Aide: We Didn’t Even See Comey’s Let...	Darrell Lucus	House Dem Aide: We Didn’t Even See Comey’s Let...	1
1	1	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...	0
2	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, ...	1
3	3	15 Civilians Killed In Single US Airstrike Hav...	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstr...	1
4	4	Iranian woman jailed for fictional unpublished...	Howard Portnoy	Print \nAn Iranian woman has been sentenced to...	1
5	5	Jackie Mason: Hollywood Would Love Trump if He...	Daniel Nussbaum	In these trying times, Jackie Mason is the Voi...	0
7	7	Benoît Hamon Wins French Socialist Party’s Pre...	Alissa J. Rubin	PARIS — France chose an idealistic, traditi...	0
9	9	A Back-Channel Plan for Ukraine and Russia, Co...	Megan Twohey and Scott Shane	A week before Michael T. Flynn resigned as nat...	0
10	10	Obama’s Organizing for Action Partners with So...	Aaron Klein	Organizing for Action, the activist group that...	0
11	11	BBC Comedy Sketch "Real Housewives of ISIS" Ca...	Chris Tomlinson	The BBC produced spoof on the “Real Housewives...	0

In [14]:
```
#making copy of data in varaible messages
messages=df.copy()
```

In [15]:
```
#reset indexing because after droping null value
#some of the index will drop
messages.reset_index(inplace=True)
```

In [16]:
```
messages.head(10)
```

Out[16]:

	index	id	title	author	text	label
0	0	0	House Dem Aide: We Didn’t Even See Comey’s Let...	Darrell Lucus	House Dem Aide: We Didn’t Even See Comey’s Let...	1
1	1	1	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...	0
2	2	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, ...	1
3	3	3	15 Civilians Killed In Single US Airstrike Hav...	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstr...	1
4	4	4	Iranian woman jailed for fictional unpublished...	Howard Portnoy	Print \nAn Iranian woman has been sentenced to...	1
5	5	5	Jackie Mason: Hollywood Would Love Trump if He...	Daniel Nussbaum	In these trying times, Jackie Mason is the Voi...	0
6	7	7	Benoît Hamon Wins French Socialist Party’s Pre...	Alissa J. Rubin	PARIS — France chose an idealistic, traditi...	0
7	9	9	A Back-Channel Plan for Ukraine and Russia, Co...	Megan Twohey and Scott Shane	A week before Michael T. Flynn resigned as nat...	0
8	10	10	Obama’s Organizing for Action Partners with So...	Aaron Klein	Organizing for Action, the activist group that...	0
9	11	11	BBC Comedy Sketch "Real Housewives of ISIS" Ca...	Chris Tomlinson	The BBC produced spoof on the “Real Housewives...	0

In [17]:
```
1 messages['title'][0:6]
```

Out[17]:
```
0    House Dem Aide: We Didn’t Even See Comey’s Let...
1    FLYNN: Hillary Clinton, Big Woman on Campus - ...
2                    Why the Truth Might Get You Fired
3    15 Civilians Killed In Single US Airstrike Hav...
4    Iranian woman jailed for fictional unpublished...
5    Jackie Mason: Hollywood Would Love Trump if He...
Name: title, dtype: object
```

Text Preprocessing

In [18]:
```
1  #To remove stopwords,special characters
2  from nltk.corpus import stopwords
3  from nltk.stem.porter import PorterStemmer
4  #porter stemmer is used to convert wording like learning into learn
5  import re
6  ps = PorterStemmer()
7  corpus = []
8  for i in range(0, len(messages)):
9      #remove all special character except a to z,A to Z
10     review = re.sub('[^a-zA-Z]', ' ', messages['title'][i])
11     #making all text to lower
12     review = review.lower()
13     #we are doing split to apply stop keywords and stemming
14     review = review.split()
15     #to check whether word belongs to stop word of english or not,
16     #if not present we will be doing stemming of that word
17     #stop word such as i,can she etc this are not so much impotant
18     review = [ps.stem(word) for word in review if not word in stopwords.words('english')]
19     #combine all space
20     review = ' '.join(review)
21     corpus.append(review)
```

In [19]:
```
1 corpus[0]
```

Out[19]:
```
'hous dem aid even see comey letter jason chaffetz tweet'
```

In [20]:
```
## Applying Countvectorizer
# Creating the Bag of Words model
from sklearn.feature_extraction.text import CountVectorizer

cv = CountVectorizer(max_features=5000,ngram_range=(1,3))
#ngram range mean it will take combination of one word,two word,three word as feature then apply bag of word
#fitting and then converting it into array
X = cv.fit_transform(corpus).toarray()
```

In [21]:
```
X.shape
```

Out[21]: (18285, 5000)

In [22]:
```
#output feature
y=messages['label']
```

In [23]:
```
## Divide the dataset into Train and Test
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=0)
```

In [24]:
```
#feature name by the combination of ngram
cv.get_feature_names()[:20]
```

Out[24]:
```
['abandon',
 'abc',
 'abc news',
 'abduct',
 'abe',
 'abedin',
 'abl',
 'abort',
 'abroad',
 'absolut',
 'abstain',
 'absurd',
 'abus',
 'abus new',
 'abus new york',
 'academi',
 'accept',
 'access',
 'access pipelin',
 'access pipelin protest']
```

In [25]:
```
#Get parameters
cv.get_params()
```

Out[25]:
```
{'analyzer': 'word',
 'binary': False,
 'decode_error': 'strict',
 'dtype': numpy.int64,
 'encoding': 'utf-8',
 'input': 'content',
 'lowercase': True,
 'max_df': 1.0,
 'max_features': 5000,
 'min_df': 1,
 'ngram_range': (1, 3),
 'preprocessor': None,
 'stop_words': None,
 'strip_accents': None,
 'token_pattern': '(?u)\\b\\w\\w+\\b',
 'tokenizer': None,
 'vocabulary': None}
```

In [26]:
```
#X Dataset converting dataframe
count_df = pd.DataFrame(X_train, columns=cv.get_feature_names())
```

In [27]:
```
count_df.head()
```

Out[27]:

	abandon	abc	abc news	abduct	abe	abedin	abl	abort	abroad	absolut	...	zero	zika	zika viru	zionist	zone	zone new	zone new york	zoo	zu	zuc
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	0	0	1	...	0	0	0	0	0	0	0	0	0	

5 rows × 5000 columns

In [28]:
```
import matplotlib.pyplot as plt
```

In [29]:
```
#create function for plotting
def plot_confusion_matrix(cm, classes,
                          normalize=False,
                          title='Confusion matrix',
                          #colour for plot
                          cmap=plt.cm.Greens):

    #This function prints and plots the confusion matrix.
    #Normalization can be applied by setting `normalize=True`.

    plt.imshow(cm, interpolation='nearest', cmap=cmap)
    plt.title(title)
    plt.colorbar()
    tick_marks = np.arange(len(classes))
    plt.xticks(tick_marks, classes, rotation=45)
    plt.yticks(tick_marks, classes)

    if normalize:
        cm = cm.astype('float') / cm.sum(axis=1)[:, np.newaxis]
        print("Normalized confusion matrix")
    else:
        print('Confusion matrix, without normalization')

    thresh = cm.max() / 2.
    for i, j in itertools.product(range(cm.shape[0]), range(cm.shape[1])):
        plt.text(j, i, cm[i, j],
                 horizontalalignment="center",
                 color="white" if cm[i, j] > thresh else "black")

    plt.tight_layout()
    plt.ylabel('True label')
    plt.xlabel('Predicted label')
```

MultinomialNB Algorithm

In [30]:
```
from sklearn.naive_bayes import MultinomialNB
#creating classifier for multinomail
classifier=MultinomialNB()
```

In [33]:
```
from sklearn import metrics
import numpy as np
import itertools
```

In [34]:
```
classifier.fit(X_train, y_train)
pred = classifier.predict(X_test)
score = metrics.accuracy_score(y_test, pred)
print("accuracy:   %0.3f" % (score*100))
#getting confusion matrix and giving to function
cm = metrics.confusion_matrix(y_test, pred)
plot_confusion_matrix(cm, classes=['FAKE', 'REAL'])
```

```
accuracy:   90.108
Confusion matrix, without normalization
```