Fake News Classifier

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]:   1  df.head()
```

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4	4	Iranian woman jailed for fictional unpublished...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

```
In [4]:   1  ## Get the Independent Features
          2
          3  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4	4	Iranian woman jailed for fictional unpublished...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  from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer, HashingVectorizer
```

```
In [10]:   1  df=df.dropna()
```

In [11]:

```
1 df.head(10)
```

Out[11]:

	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 [12]:

```
1 messages=df.copy()
```

In [13]:

```
1 messages.reset_index(inplace=True)
```

In [14]:
```
1 messages.head(10)
```

Out[14]:

	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 [15]:
```python
1  messages['text'][6]
```

Out[15]: 'PARIS — France chose an idealistic, traditional candidate in Sunday's primary to represent the Sociali st and parties in the presidential election this spring. The candidate, Benoît Hamon, 49, who ran on the s logan that he would "make France's heart beat," bested Manuel Valls, the former prime minister, whose campai gn has promoted more policies and who has a strong background. Mr. Hamon appeared to have won by a wide margin, with incomplete returns showing him with an estimated 58 percent of the vote to Mr. Valls's 41 perce nt. "Tonight the left holds its head up high again it is looking to the future," Mr. Hamon said, addressing his supporters. "Our country needs the left, but a modern, innovative left," he said. Mr. Hamon's victory wa s the clearest sign yet that voters on the left want a break with the policies of President François Holland e, who in December announced that he would not seek . However, Mr. Hamon's strong showing is unlikely to ch ange widespread assessments that candidates have little chance of making it into the second round of votin g in the general election. The first round of the general election is set for April 23 and the runoff for Ma y 7. The Socialist Party is deeply divided, and one measure of its lack of popular enthusiasm was the relati vely low number of people voting. About two million people voted in the second round of the primary on Sunda y, in contrast with about 2. 9 million in the second round of the last presidential primary on the left, in 2011. However, much of the conventional wisdom over how the elections will go has been thrown into question over the past week, because the leading candidate, François Fillon, who represents the main party, the Rep ublicans, was accused of paying his wife large sums of money to work as his parliamentary aide. While nepoti sm is legal in the French political system, it is not clear that she actually did any work. Prosecutors who specialize in financial malfeasance are reviewing the case. France's electoral system allows multiple candid ates to run for president in the first round of voting, but only the top two go on to a second round. Mr. Hamon is entering a race that is already crowded on the left, with candidates who include Mélenchon on the far left, and Emmanuel Macron, an independent who served as economy minister in Mr. Hollande's government an d who embraces more policies. Unless he decides to withdraw, Mr. Fillon, the mainstream right candidate, w ill also run, as will the extreme right candidate Marine Le Pen. The two have been expected to go to the run off. Mr. Hamon's victory can be attributed at least in part to his image as an idealist and traditional left ist candidate who appeals to union voters as well as more environmentally concerned and socially liberal you ng people. Unlike Mr. Valls, he also clearly distanced himself from some of Mr. Hollande's more unpopular po licies, especially the economic ones. Thomas Kekenbosch, 22, a student and one of the leaders of the group t he Youth With Benoît Hamon, said Mr. Hamon embodied a new hope for those on the left. "We have a perspective we have something to do, to build," Mr. Kekenbosch said. Mr. Hollande had disappointed many young people bec ause under him the party abandoned ideals, such as support for workers, that many voters believe in, accor ding to Mr. Kekenbosch. Mr. Hollande's government, under pressure from the European Union to meet budget res traints, struggled to pass labor code reforms to make the market more attractive to foreign investors and al so to encourage French businesses to expand in France. The measures ultimately passed after weeks of strike s, but they were watered down and generated little concrete progress in improving France's roughly 10 percen t unemployment rate and its nearly 25 percent youth joblessness rate. Mr. Hamon strongly endorses a stimulus approach to improving the economy and has promised to phase in a universal income, which would especially he lp young people looking for work, but would also supplement the livelihood of French workers. The end goal would be to have everyone receive 750 euros per month (about $840). "We have someone that trusts us," Mr. Ke kenbosch said, "who says: 'I give you enough to pay for your studies. You can have a scholarship which spare

s you from working at McDonald's on provisional contracts for 4 years. " Mr. Hamon advocates phasing out die
sel fuel and encouraging drivers to replace vehicles that use petroleum products with electrical ones. His l
eftist pedigree began early. His father worked at an arsenal in Brest, a city in the far west of Brittany, a
nd his mother worked off and on as a secretary. He was an early member of the Movement of Young Socialists,
and he has continued to work closely with them through his political life. He also worked for Martine Aubry,
now the mayor of Lille and a former Socialist Party leader.'

In [16]:
```python
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
import re
ps = PorterStemmer()
corpus = []
for i in range(0, len(messages)):
    review = re.sub('[^a-zA-Z]', ' ', messages['text'][i])
    review = review.lower()
    review = review.split()

    review = [ps.stem(word) for word in review if not word in stopwords.words('english')]
    review = ' '.join(review)
    corpus.append(review)
```

In [17]:
```python
corpus[3]
```

Out[17]: 'video civilian kill singl us airstrik identifi rate civilian kill american airstrik afghanistan higher us eng
ag activ combat oper photo hellfir missil load onto us militari reaper drone afghanistan staff sgt brian fergu
son u air forc bureau abl identifi civilian kill singl us drone strike afghanistan last month biggest loss civ
ilian life one strike sinc attack medecin san frontier hospit msf last octob us claim conduct counter terror s
trike islam state fighter hit nangarhar provinc missil septemb next day unit nation issu unusu rapid strong st
atement say strike kill civilian injur other gather hous celebr tribal elder return pilgrimag mecca bureau spo
ke man name haji rai said owner hous target said peopl kill other injur provid name list bureau abl independ v
erifi ident die rai son headmast local school among anoth man abdul hakim lost three son attack rai said invol
v deni us claim member visit hous strike said even speak sort peopl phone let alon receiv hous death amount bi
ggest confirm loss civilian life singl american strike afghanistan sinc attack msf hospit kunduz last octob ki
ll least peopl nangarhar strike us attack kill civilian septemb bureau data indic mani civilian alli soldier k
ill four american strike afghanistan somalia month septemb pair strike kill eight afghan policemen tarinkot ca
pit urozgan provic us jet reportedli hit polic checkpoint kill one offic return target first respond use tacti
c known doubl tap strike controversi often hit civilian rescuer us told bureau conduct strike individu fire po
se threat afghan forc email directli address alleg afghan policemen kill end month somalia citizen burnt us fl
ag street north central citi galcayo emerg drone attack may unintent kill somali soldier civilian strike occur
day one nangarhar somali afghan incid us first deni non combat kill investig strike nangarhar galcayo rate civ
ilian kill american airstrik afghanistan higher us engag activ combat oper name'

In [18]:
```python
## TFidf Vectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
tfidf_v=TfidfVectorizer(max_features=5000,ngram_range=(1,3))
X=tfidf_v.fit_transform(corpus).toarray()

```

In [19]:
```python
X.shape
```

Out[19]: (18285, 5000)

In [20]:
```python
y=messages['label']
```

In [21]:
```python
## 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 [22]:
```python
tfidf_v.get_feature_names()[:20]
```

Out[22]: ['aaron',
 'abandon',
 'abc',
 'abe',
 'abedin',
 'abil',
 'abl',
 'abort',
 'abroad',
 'absenc',
 'absolut',
 'absorb',
 'absurd',
 'abu',
 'abus',
 'academ',
 'academi',
 'acceler',
 'accept',
 'access']

In [23]:
```python
tfidf_v.get_params()
```

Out[23]: {'analyzer': 'word',
 'binary': False,
 'decode_error': 'strict',
 'dtype': numpy.float64,
 'encoding': 'utf-8',
 'input': 'content',
 'lowercase': True,
 'max_df': 1.0,
 'max_features': 5000,
 'min_df': 1,
 'ngram_range': (1, 3),
 'norm': 'l2',
 'preprocessor': None,
 'smooth_idf': True,
 'stop_words': None,
 'strip_accents': None,
 'sublinear_tf': False,
 'token_pattern': '(?u)\\b\\w\\w+\\b',
 'tokenizer': None,
 'use_idf': True,
 'vocabulary': None}

In [24]:
```python
count_df = pd.DataFrame(X_train, columns=tfidf_v.get_feature_names())
```

In [25]:
```
1  count_df.head()
```

Out[25]:

	aaron	abandon	abc	abe	abedin	abil	abl	abort	abroad	absenc	...	young	young peopl	younger	youth	youtub	zero	zika	zionist
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.200698	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	0.0	...	0.000000	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	0.0	...	0.000000	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	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0

5 rows × 5000 columns

In [26]:
```
1  import matplotlib.pyplot as plt
```

```
In [28]:    1   #Normalization can be applied by setting `normalize=True.
            2   def plot_confusion_matrix(cm, classes,
            3                             normalize=False,
            4                             title='Confusion matrix',
            5                             cmap=plt.cm.Blues):
            6
            7
            8
            9       plt.imshow(cm, interpolation='nearest', cmap=cmap)
           10       plt.title(title)
           11       plt.colorbar()
           12       tick_marks = np.arange(len(classes))
           13       plt.xticks(tick_marks, classes, rotation=45)
           14       plt.yticks(tick_marks, classes)
           15
           16       if normalize:
           17           cm = cm.astype('float') / cm.sum(axis=1)[:, np.newaxis]
           18           print("Normalized confusion matrix")
           19       else:
           20           print('Confusion matrix, without normalization')
           21
           22       thresh = cm.max() / 2.
           23       for i, j in itertools.product(range(cm.shape[0]), range(cm.shape[1])):
           24           plt.text(j, i, cm[i, j],
           25                    horizontalalignment="center",
           26                    color="white" if cm[i, j] > thresh else "black")
           27
           28       plt.tight_layout()
           29       plt.ylabel('True label')
           30       plt.xlabel('Predicted label')
```
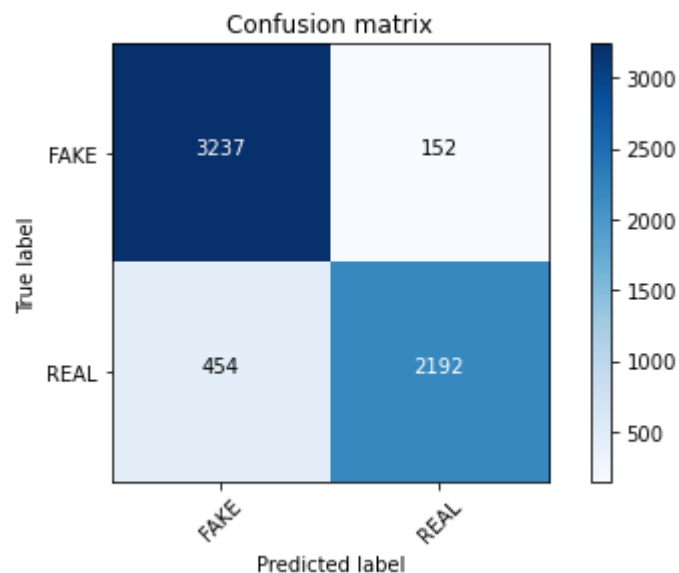
MultinomialNB Algorithm

```
In [29]:    1   from sklearn.naive_bayes import MultinomialNB
            2   classifier=MultinomialNB()
```

In [30]:
```python
from sklearn import metrics
import numpy as np
import itertools
```

In [31]:
```python

classifier.fit(X_train, y_train)
pred = classifier.predict(X_test)
score = metrics.accuracy_score(y_test, pred)
print("accuracy:   %0.3f" % (score*100))
cm = metrics.confusion_matrix(y_test, pred)
plot_confusion_matrix(cm, classes=['FAKE', 'REAL'])
```

accuracy: 89.959
Confusion matrix, without normalization

In [ ]:
```python

```