Fake News Detection ¶

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

In [1]:

```
import pandas as pd
```

In [2]:

```
df=pd.read_csv("D:/fakenews_data/train.csv")
```

In [3]:

```
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4	4	Iranian woman jailed for fictional unpublished...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]:

```
## Get the Independent Features

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

In [5]:

```
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4	4	Iranian woman jailed for fictional unpublished...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]:

```
## Get the Dependent features
y=df['label']
```

In [7]:

```
y.head()
```

Out[7]:

```
0    1
1    0
2    1
3    1
4    1
Name: label, dtype: int64
```

In [8]:

```
df.shape
```

Out[8]:

```
(20800, 5)
```

In [9]:

```
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer, HashingVectorizer
```

In [10]:

```
df=df.dropna()
```

In [11]:

```
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]:

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

In [13]:

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

In [14]:

```
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]:
```
messages['text'][0]
```

Out[15]:
```
'House Dem Aide: We Didn’t Even See Comey’s Letter Until Jason Chaffetz Tweeted It By Darrell Lucus on Octob
er 30, 2016 Subscribe Jason Chaffetz on the stump in American Fork, Utah ( image courtesy Michael Jolley, av
ailable under a Creative Commons-BY license) \nWith apologies to Keith Olbermann, there is no doubt who the
Worst Person in The World is this week–FBI Director James Comey. But according to a House Democratic aide, i
t looks like we also know who the second-worst person is as well. It turns out that when Comey sent his now-
infamous letter announcing that the FBI was looking into emails that may be related to Hillary Clinton’s ema
il server, the ranking Democrats on the relevant committees didn’t hear about it from Comey. They found out
via a tweet from one of the Republican committee chairmen. \nAs we now know, Comey notified the Republican c
hairmen and Democratic ranking members of the House Intelligence, Judiciary, and Oversight committees that h
is agency was reviewing emails it had recently discovered in order to see if they contained classified infor
mation. Not long after this letter went out, Oversight Committee Chairman Jason Chaffetz set the political w
orld ablaze with this tweet. FBI Dir just informed me, "The FBI has learned of the existence of emails that
appear to be pertinent to the investigation." Case reopened \n— Jason Chaffetz (@jasoninthehouse) October 2
8, 2016 \nOf course, we now know that this was not the case . Comey was actually saying that it was reviewin
g the emails in light of “an unrelated case”–which we now know to be Anthony Weiner’s sexting with a teenage
r. But apparently such little things as facts didn’t matter to Chaffetz. The Utah Republican had already vow
ed to initiate a raft of investigations if Hillary wins–at least two years’ worth, and possibly an entire te
rm’s worth of them. Apparently Chaffetz thought the FBI was already doing his work for him–resulting in a tw
eet that briefly roiled the nation before cooler heads realized it was a dud. \nBut according to a senior Ho
use Democratic aide, misreading that letter may have been the least of Chaffetz’ sins. That aide told Shareb
lue that his boss and other Democrats didn’t even know about Comey’s letter at the time–and only found out w
hen they checked Twitter. “Democratic Ranking Members on the relevant committees didn’t receive Comey’s lett
er until after the Republican Chairmen. In fact, the Democratic Ranking Members didn’ receive it until after
the Chairman of the Oversight and Government Reform Committee, Jason Chaffetz, tweeted it out and made it pu
blic.” \nSo let’s see if we’ve got this right. The FBI director tells Chaffetz and other GOP committee chair
men about a major development in a potentially politically explosive investigation, and neither Chaffetz nor
his other colleagues had the courtesy to let their Democratic counterparts know about it. Instead, according
to this aide, he made them find out about it on Twitter. \nThere has already been talk on Daily Kos that Com
ey himself provided advance notice of this letter to Chaffetz and other Republicans, giving them time to tur
n on the spin machine. That may make for good theater, but there is nothing so far that even suggests this i
s the case. After all, there is nothing so far that suggests that Comey was anything other than grossly inco
mpetent and tone-deaf. \nWhat it does suggest, however, is that Chaffetz is acting in a way that makes Dan B
urton and Darrell Issa look like models of responsibility and bipartisanship. He didn’t even have the decenc
y to notify ranking member Elijah Cummings about something this explosive. If that doesn’t trample on basic
standards of fairness, I don’t know what does. \nGranted, it’s not likely that Chaffetz will have to answer
for this. He sits in a ridiculously Republican district anchored in Provo and Orem; it has a Cook Partisan V
oting Index of R+25, and gave Mitt Romney a punishing 78 percent of the vote in 2012. Moreover, the Republic
an House leadership has given its full support to Chaffetz’ planned fishing expedition. But that doesn’t mea
n we can’t turn the hot lights on him. After all, he is a textbook example of what the House has become unde
r Republican control. And he is also the Second Worst Person in the World. About Darrell Lucus \nDarrell is
```

```
a 30-something graduate of the University of North Carolina who considers himself a journalist of the old sc
hool. An attempt to turn him into a member of the religious right in college only succeeded in turning him i
nto the religious right\'s worst nightmare--a charismatic Christian who is an unapologetic liberal. His desi
re to stand up for those who have been scared into silence only increased when he survived an abusive three-
year marriage. You may know him on Daily Kos as Christian Dem in NC . Follow him on Twitter @DarrellLucus or
connect with him on Facebook . Click here to buy Darrell a Mello Yello. Connect'
```

In [16]:

```python
#To remove stopwords,special characters
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
#porter stemmer is used to convert wording like learning into learn
import re
ps = PorterStemmer()
corpus = []
for i in range(0, len(messages)):
    #remove all special character except a to z,A to Z
    review = re.sub('[^a-zA-Z]', ' ', messages['text'][i])
    #making all text to lower
    review = review.lower()
    #we are doing split to apply stop keywords and stemming
    review = review.split()
    #to check whether word belongs to stop word of english or not,
    #if not present we will be doing stemming of that word
    #stop word such as i,can she etc this are not so much impotant
    review = [ps.stem(word) for word in review if not word in stopwords.words('english')]
    #combine all space
    review = ' '.join(review)
    corpus.append(review)
```

In [17]:

```
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]:

```
## Applying Countvectorizer
# Creating the Bag of Words model
from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer(max_features=5000,ngram_range=(1,3))
X = cv.fit_transform(corpus).toarray()
```

In [19]:

```
X.shape
```

Out[19]:

```
(18285, 5000)
```

In [20]:

```
y=messages['label']
```

In [21]:

```
## 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]:

```
cv.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]:
```
cv.get_params()
```

Out[23]:
```
{'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 [24]:
```
count_df = pd.DataFrame(X_train, columns=cv.get_feature_names())
```

In [25]:
```
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	...	3	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
2	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
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0

5 rows × 5000 columns

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

In [27]:

```python
def plot_confusion_matrix(cm, classes,
                          normalize=False,
                          title='Confusion matrix',
                          cmap=plt.cm.Blues):


#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 [28]:

```python

from sklearn.naive_bayes import MultinomialNB
classifier=MultinomialNB()
```

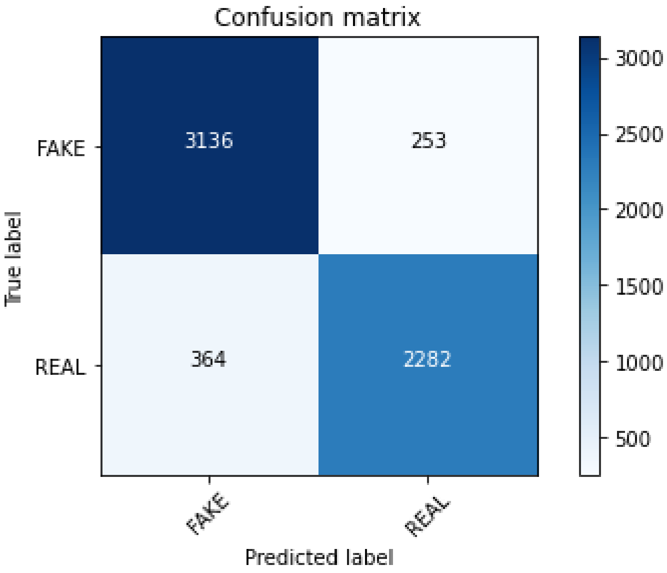
In [29]:

```python
from sklearn import metrics
import numpy as np
import itertools
```

In [30]:

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

```
accuracy:   0.898
Confusion matrix, without normalization
```



Confusion matrix

True label \ Predicted label	FAKE	REAL
FAKE	3136	253
REAL	364	2282