# Fake News Classifier

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

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

In [2]:
```
df=pd.read_csv('fake-news/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| 4 | 4 | Iranian woman jailed for fictional unpublished...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| 4 | 4 | Iranian woman jailed for fictional unpublished...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 | Benoit 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 | Benoit 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'][6]
```

Out[15]:
```
'PARIS  —  France chose an idealistic, traditional  candidate in Sunday's primary to represent the Socialist and  parties in the presidential election this spring. The candidate, Benoît Hamon, 49, who ran on the slogan that he would "put some heart beat," bested Manuel Valls, the former prime minister, whose campaign has promoted more  policies and who has a strong  background. Mr. Hamon appeared to have won by a wide margin, with incomplete re
sults showing him with an estimated 58 percent of the vote to Mr. Valls's 41 percent. "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 was the clearest sign yet that voters on the left want a break with the policies of President François Hollande, who in December announced that he would not seek
. However, Mr. Hamon's strong showing is unlikely to change widespread assessments that  candidates have little chance of making it into the second round of voting in the general election. The first round of the general election
is set for April 23 and the runoff for May 7. The Socialist Party is deeply divided, and one measure of its lack of popular enthusiasm was the relatively low number of people voting. About two million people voted in the second r
ound of the primary on Sunday, 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 Republicans, was accused of paying his wife large sums of money to work as his parliamentary aide. While nepot
ism 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 candidates 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 M
acron, an independent who served as economy minister in Mr. Hollande's government and who embraces more  policies. Unless he decides to withdraw, Mr. Fillon, the mainstream right candidate, will also run, as will the extreme rig
ht candidate Marine Le Pen. The two would have been expected to go to the runoff. Mr. Hamon's victory can be attributed at least in part to his image as an idealist and traditional leftist candidate who appeals to union voters as well
as more environmentally concerned and socially liberal young people. Unlike Mr. Valls, he also clearly distanced himself from some of Mr. Hollande's more unpopular policies, especially the economic ones. Thomas Kekenbosch, 22, a
student and one of the leaders of the group the 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 d
isappointed many young people because under him the party abandoned ideals, such as support for workers, that many  voters believe in, according to Mr. Kekenbosch. Mr. Hollande's government, under pressure from the European Unio
n to meet deficit restraints, struggled to pass labor code reforms to make the market more attractive to foreign investors and also to encourage French businesses to expand in France. The measures ultimately passed after weeks of
strikes, but they were watered down and generated little concrete progress in improving France's roughly 10 percent unemployment rate and its nearly 25 percent youth joblessness rate. Mr. Hamon strongly endorses a stimulus approa
ch to improving the economy and has promised to phase in a universal income, which would especially help young people looking for work, but would also supplement the livelihood of  French workers. The end goal would be to have e
veryone receive 750 euros per month (about $840). "We have someone that trusts us," Mr. Kekenbosch said, "who says: 'I give you enough to pay for your studies. You can have a scholarship which spares you from working at McDonal
d's on provisional contracts for 4 years. " Mr. Hamon advocates phasing out diesel fuel and encouraging drivers to replace vehicles that use petroleum products with electrical ones. His leftist pedigree began early. His father wo
rked as an arsenal in Brest, a city in the far west of Brittany, and 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 [ ]:
```
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 [ ]:
```
corpus[3]
```

In [ ]:
```

```

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

In [ ]:
```

```

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

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

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

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

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

In [ ]:
```

```

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

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

    """
    See full source and example:
    http://scikit-learn.org/stable/auto_examples/model_selection/plot_confusion_matrix.html

    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 [ ]:
```
from sklearn.naive_bayes import MultinomialNB
classifier=MultinomialNB()
```

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

In [ ]:
```

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 Data', 'REAL Data'])
```

In [ ]:
```

classifier.fit(X_train, y_train)
pred = classifier.predict(X_test)
score = metrics.accuracy_score(y_test, pred)
score
```

In [ ]:
```
y_train.shape
```

## Passive Aggressive Classifier Algorithm

In [ ]:
```
from sklearn.linear_model import PassiveAggressiveClassifier
linear_clf = PassiveAggressiveClassifier(n_iter=50)
```

In [ ]:
```

linear_clf.fit(X_train, y_train)
pred = linear_clf.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 Data', 'REAL Data'])
```

## Multinomial Classifier with Hyperparameter

In [ ]:
```
classifier=MultinomialNB(alpha=0.1)
```

In [ ]:
```

previous_score=0
for alpha in np.arange(0,1,0.1):
    sub_classifier=MultinomialNB(alpha=alpha)
    sub_classifier.fit(X_train,y_train)
    y_pred=sub_classifier.predict(X_test)
    score = metrics.accuracy_score(y_test, y_pred)
    if score>previous_score:
        classifier=sub_classifier
    print("Alpha: {}, Score : {}".format(alpha,score))
```

In [ ]:
```
## Get Features names
feature_names = cv.get_feature_names()
```

In [ ]:
```

classifier.coef_[0]
```

In [ ]:
```

### Most real
sorted(zip(classifier.coef_[0], feature_names), reverse=True)[:20]
```

In [ ]:
```

### Most fake
sorted(zip(classifier.coef_[0], feature_names))[:5000]
```

## HashingVectorizer

In [ ]:
```
hs_vectorizer=HashingVectorizer(n_features=5000,non_negative=True)
X=hs_vectorizer.fit_transform(corpus).toarray()
```

In [ ]:
```

X.shape
```

In [ ]:
```
X
```

In [ ]:
```
## 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 [ ]:
```

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

In [ ]:
```

```