In [1]:

```
import pandas as pd
```

In [2]:

```
true = pd.read_csv('True.csv')
fake = pd.read_csv('Fake.csv')
```

In [3]:

```
true.head(3)
```

Out[3]:

	title	text	subject	date
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017

In [4]:

```
fake.head(3)
```

Out[4]:

	title	text	subject	date
0	Donald Trump Sends Out Embarrassing New Year’...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017

In [5]:

```
true.shape
```

Out[5]:

```
(21417, 4)
```

In [6]:

```
fake.shape
```

Out[6]:

```
(23481, 4)
```

In [7]:

```
true['label'] = 1
fake['label'] = 0
```

In [8]:

```
#first 5000 data of true and fake for the model
frames = [true.loc[:5000][:], fake.loc[:5000][:]]
```

In [9]:

```
df = pd.concat(frames)
```

In [10]:

```
df.shape
```

Out[10]:

```
(10002, 5)
```

In [11]:
```
df.tail()
```

Out[11]:

	title	text	subject	date	label
4996	Justice Department Announces It Will No Longe...	Republicans are about to lose a huge source of...	News	August 18, 2016	0
4997	WATCH: S.E. Cupp Destroys Trump Adviser’s ‘Fa...	A pawn working for Donald Trump claimed that w...	News	August 18, 2016	0
4998	WATCH: Fox Hosts Claim Hillary Has Brain Dama...	Fox News is desperate to sabotage Hillary Clin...	News	August 18, 2016	0
4999	CNN Panelist LAUGHS In Corey Lewandowski’s Fa...	As Donald Trump s campaign continues to sink d...	News	August 18, 2016	0
5000	Trump Supporter Who Wants To Shoot Black Kids...	Hi folks, John Harper here, at least if you as...	News	August 18, 2016	0

In [12]:
```
X = df. drop('label', axis=1)
y = df['label']
```

In [13]:
```
df = df.dropna()
df2 = df.copy()
```

In [14]:
```
df2.head()
```

Out[14]:

	title	text	subject	date	label
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017	1
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	1
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	1
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017	1
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017	1

In [15]:
```
df2.reset_index(inplace=True)
df2.head()
```

Out[15]:

	index	title	text	subject	date	label
0	0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017	1
1	1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	1
2	2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	1
3	3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017	1
4	4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017	1

In [16]:

```
df2['title'][2]
```

Out[16]:

```
"Senior U.S. Republican senator: 'Let Mr. Mueller do his job'"
```

Data Preprocessing

In [17]:

```
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
ps = PorterStemmer()
```

In [18]:

```
import re
import nltk
nltk.download('stopwords')

corpus = []
for i in range(0, len(df2)):
    review = re.sub('[^a-zA-Z]', ' ', df2['text'][i])
    review = review.lower()
    review = review.split()

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

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\kevin\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

In [20]:

```
# TFidf Vectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
tfidf_v = TfidfVectorizer(max_features=5000, ngram_range=(1,3))
```

In [21]:

```
X = tfidf_v.fit_transform(corpus).toarray()
y = df2['label']
```

In [22]:

```
# 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=0.2, random_sta
```

Model building - Passive Aggresive Classifier

In [23]:

```
from sklearn.linear_model import PassiveAggressiveClassifier
classifier = PassiveAggressiveClassifier(max_iter=1000)
```

In [24]:

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

classifier.fit(X_train, y_train)

pred = classifier.predict(X_test)

score = metrics.accuracy_score(y_test, pred)
print("accuracy:   %0.3f" % score)
```

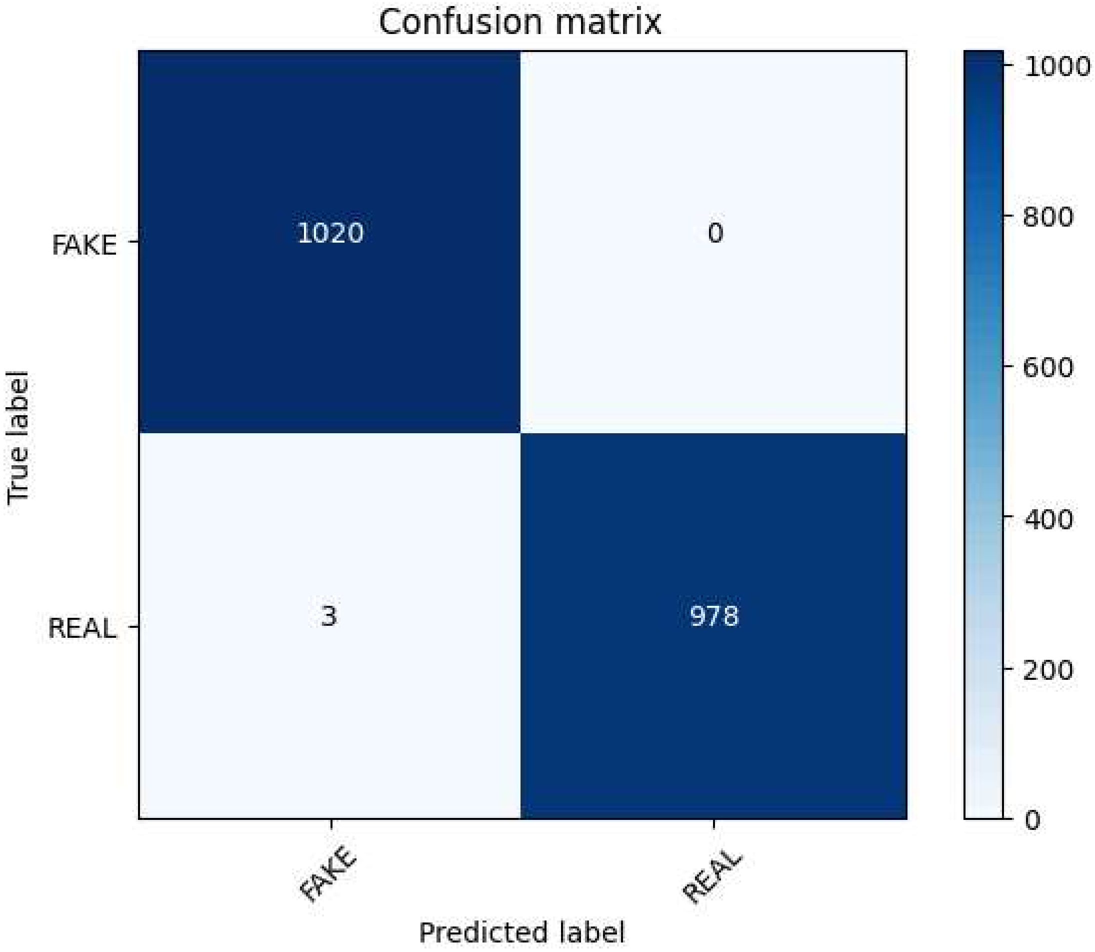
```
accuracy:   0.999
```

In [26]:

```
import matplotlib.pyplot as plt

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

    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')
```

In [27]:
```
cm = metrics.confusion_matrix(y_test, pred)
plot_confusion_matrix(cm, classes=['FAKE', 'REAL'])
```

```
Confusion matrix, without normalization
```



In [28]:
```
review = re.sub('[^a-zA-Z]', ' ', fake['text'][13070])
review = review.lower()
review = review.split()

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

Out[28]:
```
'mani c word hillari compet one ouch'
```

In [29]:
```
val = tfidf_v.transform([review]).toarray()
```

In [30]:
```
classifier.predict(val)
```

Out[30]:
```
array([0])
```

Save model and vectorizer

In [31]:
```
import pickle
```

In [32]:
```
pickle.dump(classifier, open('model2.pkl', 'wb'))
```

In [33]:
```
pickle.dump(tfidf_v, open('tfidfvect2.pkl', 'wb'))
```

Load model and vectorizer to predict the previous datapoint

In [34]:
```
joblib_model = pickle.load(open('model2.pkl', 'rb'))
```

In [35]:
```
joblib_vect = pickle.load(open('tfidfvect2.pkl', 'rb'))
```

In [36]:
```
val_pkl = joblib_vect.transform([review]).toarray()
```

In [37]:
```
joblib_model.predict(val_pkl)
```

Out[37]:
```
array([0])
```

In [ ]: