In [1]:
```
#importind libraries required
import pandas as pd
import numpy as np
```

In [2]:
```
#impporting the data
data = pd.read_csv("twitter.csv")
```

In [3]:
```
#checking the data
data
```

Out[3]:

	Unnamed: 0	count	hate_speech	offensive_language	neither	class	tweet
0	0	3	0	0	3	2	!!! RT @mayasolovely: As a woman you shouldn't...
1	1	3	0	3	0	1	!!!!! RT @mleew17: boy dats cold...tyga dwn ba...
2	2	3	0	3	0	1	!!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby...
3	3	3	0	2	1	1	!!!!!!!!! RT @C_G_Anderson: @viva_based she lo...
4	4	6	0	6	0	1	!!!!!!!!!!!! RT @ShenikaRoberts: The shit you...
...	...	...	...	...	...	...	...
24778	25291	3	0	2	1	1	you's a muthaf***in lie “@LifeAsKing: @2...
24779	25292	3	0	1	2	2	you've gone and broke the wrong heart baby, an...
24780	25294	3	0	3	0	1	young buck wanna eat!!.. dat nigguh like I ain...
24781	25295	6	0	6	0	1	youu got wild bitches tellin you lies
24782	25296	3	0	0	3	2	~~Ruffled | Ntac Eileen Dahlia - Beautiful col...

24783 rows × 7 columns

In [4]:
```
data.describe()
```

Out[4]:

	Unnamed: 0	count	hate_speech	offensive_language	neither	class
count	24783.000000	24783.000000	24783.000000	24783.000000	24783.000000	24783.000000
mean	12681.192027	3.243473	0.280515	2.413711	0.549247	1.110277
std	7299.553863	0.883060	0.631851	1.399459	1.113299	0.462089
min	0.000000	3.000000	0.000000	0.000000	0.000000	0.000000
25%	6372.500000	3.000000	0.000000	2.000000	0.000000	1.000000
50%	12703.000000	3.000000	0.000000	3.000000	0.000000	1.000000
75%	18995.500000	3.000000	0.000000	3.000000	0.000000	1.000000
max	25296.000000	9.000000	7.000000	9.000000	9.000000	2.000000

In [5]:
```
#cchecking whether any null entries are present or not
data.isnull().sum()
```

Out[5]:
```
Unnamed: 0            0
count                 0
hate_speech           0
offensive_language    0
neither               0
class                 0
tweet                 0
dtype: int64
```

In [6]:
```
#Labeling the tweets
data["labels"]  = data["class"].map({0: "Hate speech",
                                     1:"Offensive speech",
                                     2:"Neither hate nor offensive"})
```

In [7]:
```
data
```

Out[7]:

	Unnamed: 0	count	hate_speech	offensive_language	neither	class	tweet	labels
0	0	3	0	0	3	2	!!! RT @mayasolovely: As a woman you shouldn't...	Neither hate nor offensive
1	1	3	0	3	0	1	!!!!! RT @mleew17: boy dats cold...tyga dwn ba...	Offensive speech
2	2	3	0	3	0	1	!!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby...	Offensive speech
3	3	3	0	2	1	1	!!!!!!!! RT @C_G_Anderson: @viva_based she lo...	Offensive speech
4	4	6	0	6	0	1	!!!!!!!!!!!! RT @ShenikaRoberts: The shit you...	Offensive speech
...	...	...	...	...	...	...	...	...
24778	25291	3	0	2	1	1	you's a muthaf***in lie “@LifeAsKing: @2...	Offensive speech
24779	25292	3	0	1	2	2	you've gone and broke the wrong heart baby, an...	Neither hate nor offensive
24780	25294	3	0	3	0	1	young buck wanna eat!!.. dat nigguh like I ain...	Offensive speech
24781	25295	6	0	6	0	1	youu got wild bitches tellin you lies	Offensive speech
24782	25296	3	0	0	3	2	~~Ruffled | Ntac Eileen Dahlia - Beautiful col...	Neither hate nor offensive

24783 rows × 8 columns

In [8]:
```
data = data[["tweet","labels"]]
```

In [9]:
```
data
```

Out[9]:

	tweet	labels
0	!!! RT @mayasolovely: As a woman you shouldn't...	Neither hate nor offensive
1	!!!!! RT @mleew17: boy dats cold...tyga dwn ba...	Offensive speech
2	!!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby...	Offensive speech
3	!!!!!!!!! RT @C_G_Anderson: @viva_based she lo...	Offensive speech
4	!!!!!!!!!!!!! RT @ShenikaRoberts: The shit you...	Offensive speech
...	...	...
24778	you's a muthaf***in lie “@LifeAsKing: @2...	Offensive speech
24779	you've gone and broke the wrong heart baby, an...	Neither hate nor offensive
24780	young buck wanna eat!!.. dat nigguh like I ain...	Offensive speech
24781	youu got wild bitches tellin you lies	Offensive speech
24782	~~Ruffled | Ntac Eileen Dahlia - Beautiful col...	Neither hate nor offensive

24783 rows × 2 columns

In [10]:
```
#Data preprocessing-(removing ! & @)
import re
import nltk
import string
```

In [11]:
```
import nltk
nltk.download('stopwords')
```

```
[nltk_data] Error loading stopwords: <urlopen error [WinError 10060] A
[nltk_data]     connection attempt failed because the connected party
[nltk_data]     did not properly respond after a period of time, or
[nltk_data]     established connection failed because connected host
[nltk_data]     has failed to respond>
```

Out[11]:
```
False
```

In [12]:
```
from nltk.corpus import stopwords
stopwords = stopwords.words("english")
```

In [13]:
```
stemmer = nltk.SnowballStemmer("english")
```

In [14]:
```
#data cleaning
def clean(text):
    text = str(text).lower()
    text = re.sub('https?://\+-|www.Smh.', ' ', text)
    text = re.sub('\.img[#*&]',' ',text)
    text = re.sub('[%s]'%re.escape(string.punctuation),' ',text)
    text = re.sub('\n','',text)
    #stopwords removal
    text={word for word in text.split(' ') if word not in stopwords}
    text = ' '.join(text)
    #stemming
    text = {stemmer.stem(word) for word in text.split(' ')}
    text= ' '.join(text)
    return text
```

In [15]:
```
 data["tweet"] = data["tweet"].apply(clean)
```

```
C:\Users\Saketh Nandan\AppData\Local\Temp\ipykernel_21356\2479938855.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  data["tweet"] = data["tweet"].apply(clean)
```

In [16]:
```
data
```

Out[16]:

	tweet	labels
0	man rt woman take trash amp clean mayasolov h...	Neither hate nor offensive
1	mleew17 boy rt cuffin dat place hoe bad 1st t...	Offensive speech
2	fuck confus rt cri dawg bitch ever 80sbaby4li...	Offensive speech
3	c viva g rt like base anderson look tranni	Offensive speech
4	faker might rt ya bitch true hear told shenik...	Offensive speech
...	...	...
24778	right tl lie 8221 muthaf hymn corey trash 20 ...	Offensive speech
24779	broke redneck heart drove babi gone wrong crazi	Neither hate nor offensive
24780	like fuckin young dat wanna nigguh eat dis bu...	Offensive speech
24781	lie wild bitch youu tellin got	Offensive speech
24782	white coll h0dyebvnzb ntac color combin ruffl...	Neither hate nor offensive

24783 rows × 2 columns

In [17]:
```
data.tail(10)
```

Out[17]:

	tweet	labels
24773	smh ya cheat gf nigger	Offensive speech
24774	care realli bitch bout dis dick feel yo	Offensive speech
24775	need bitch bout worri	Offensive speech
24776	nigger	Hate speech
24777	faggot fuck 2 sugar retard type hope get dare...	Hate speech
24778	right tl lie 8221 muthaf hymn corey trash 20 ...	Offensive speech
24779	broke redneck heart drove babi gone wrong crazi	Neither hate nor offensive
24780	like fuckin young dat wanna nigguh eat dis bu...	Offensive speech
24781	lie wild bitch youu tellin got	Offensive speech
24782	white coll h0dyebvnzb ntac color combin ruffl...	Neither hate nor offensive

In [18]:
```
data
```

Out[18]:

	tweet	labels
0	man rt woman take trash amp clean mayasolov h...	Neither hate nor offensive
1	mleew17 boy rt cuffin dat place hoe bad 1st t...	Offensive speech
2	fuck confus rt cri dawg bitch ever 80sbaby4li...	Offensive speech
3	c viva g rt like base anderson look tranni	Offensive speech
4	faker might rt ya bitch true hear told shenik...	Offensive speech
...	...	...
24778	right tl lie 8221 muthaf hymn corey trash 20 ...	Offensive speech
24779	broke redneck heart drove babi gone wrong crazi	Neither hate nor offensive
24780	like fuckin young dat wanna nigguh eat dis bu...	Offensive speech
24781	lie wild bitch youu tellin got	Offensive speech
24782	white coll h0dyebvnzb ntac color combin ruffl...	Neither hate nor offensive

24783 rows × 2 columns

In [19]:
```
#storing the clean data in np arrays
X = np.array(data["tweet"])
Y = np.array(data["labels"])
```

In [20]:
```
Y
```

Out[20]:
```
array(['Neither hate nor offensive', 'Offensive speech',
       'Offensive speech', ..., 'Offensive speech', 'Offensive speech',
       'Neither hate nor offensive'], dtype=object)
```

In [21]:
```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import train_test_split
```

In [22]:
```
cv = CountVectorizer()
X = cv.fit_transform(X)
```

In [23]:
```
X
```

Out[23]:
```
<24783x31370 sparse matrix of type '<class 'numpy.int64'>'
        with 224308 stored elements in Compressed Sparse Row format>
```

In [24]:
```
X_train, X_test, Y_train, Y_test =train_test_split(X,Y,test_size=0.33,random_state=42)
```

In [25]:
```
Y_train
```

Out[25]:
```
array(['Offensive speech', 'Offensive speech', 'Offensive speech', ...,
       'Offensive speech', 'Offensive speech', 'Offensive speech'],
      dtype=object)
```

In [26]:
```
X_train
```

Out[26]:
```
<16604x31370 sparse matrix of type '<class 'numpy.int64'>'
        with 150226 stored elements in Compressed Sparse Row format>
```

In [27]:
```
X_test
```

Out[27]:
```
<8179x31370 sparse matrix of type '<class 'numpy.int64'>'
        with 74082 stored elements in Compressed Sparse Row format>
```

In [28]:
```
#machine learning model intiation
from sklearn.tree import DecisionTreeClassifier
```

In [29]:
```
dt = DecisionTreeClassifier()
dt.fit(X_train,Y_train)
```

Out[29]:
```
DecisionTreeClassifier()
```
In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [30]:
```
Y_pred = dt.predict(X_test)
```

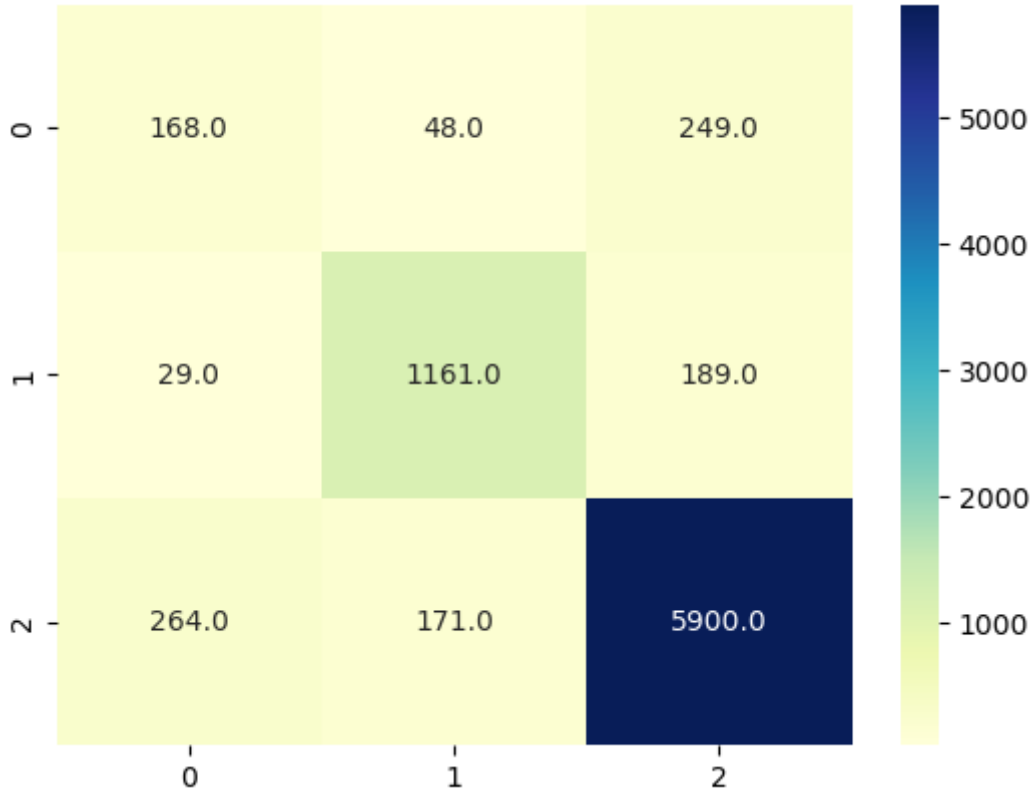
In [31]:
```
#confusion matrix and accuracy
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(Y_test,Y_pred)
cm
```

Out[31]:
```
array([[ 168,   48,  249],
       [  29, 1161,  189],
       [ 264,  171, 5900]], dtype=int64)
```

In [32]:
```
import seaborn as sns
import matplotlib.pyplot as ply
%matplotlib inline
```

In [33]:
```
sns.heatmap(cm, annot = True, fmt = ".1f",cmap="YlGnBu" )
```

Out[33]: <Axes: >



In [34]:
```
from sklearn.metrics import accuracy_score
accuracy_score(Y_test,Y_pred)
```

Out[34]: 0.8838488812813302

In [35]:
```
sample = " bitch i am stylist black shirt pink tishirt mother"
sample = clean(sample)
sample
```

Out[35]:
```
' black stylist mother shirt bitch tishirt pink'
```

In [36]:
```
data1 = cv.transform([sample]).toarray()
```

In [37]:
```
data1
```

Out[37]:
```
array([[0, 0, 0, ..., 0, 0, 0]], dtype=int64)
```

In [38]:
```
dt.predict(data1)
```

Out[38]:
```
array(['Offensive speech'], dtype=object)
```

In [41]:
```
s2= "hello i am saketh %&^$:"
s2=clean(s2)
s2
```

Out[41]:
```
' hello saketh'
```

In [44]:
```
data2=cv.transform([s2]).toarray()
data2
```

Out[44]:
```
array([[0, 0, 0, ..., 0, 0, 0]], dtype=int64)
```

In [46]:
```
dt.predict(data2)
```

Out[46]:
```
array(['Neither hate nor offensive'], dtype=object)
```