In [7]:

```
import pandas as pd
import numpy as np
```

In [8]:

```
dataset=pd.read_csv("labeled_data.csv")
```

In [9]:

```
dataset
```

Out[9]:

	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 &#8220;@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 [10]:

```
dataset.head()
```

Out[10]:

	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...

In [11]:

```
dataset.tail()
```

Loading [MathJax]/extensions/Safe.js

Out[11]:

	Unnamed: 0	count	hate_speech	offensive_language	neither	class	tweet
24778	25291	3	0	2	1	1	you's a muthaf***in lie &#8220;@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...

In [12]:

```
dataset.isnull().sum()
```

Out[12]:

```
Unnamed: 0            0
count                 0
hate_speech           0
offensive_language    0
neither               0
class                 0
tweet                 0
dtype: int64
```

In [13]:

```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 24783 entries, 0 to 24782
Data columns (total 7 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   Unnamed: 0          24783 non-null  int64 
 1   count               24783 non-null  int64 
 2   hate_speech         24783 non-null  int64 
 3   offensive_language  24783 non-null  int64 
 4   neither             24783 non-null  int64 
 5   class               24783 non-null  int64 
 6   tweet               24783 non-null  object
dtypes: int64(6), object(1)
memory usage: 1.2+ MB
```

In [14]:

```
dataset.describe()
```

Out[14]:

	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 [15]:

```
dataset["labels"]=dataset["class"].map({0:"Hate speech", 1:"Offensive language", 2:"No h
```

Loading [MathJax]/extensions/Safe.js

Out[16]:

	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...	No hate nor offensive
1	1	3	0	3	0	1	!!!!! RT @mleew17: boy dats cold...tyga dwn ba...	Offensive language
2	2	3	0	3	0	1	!!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby...	Offensive language
3	3	3	0	2	1	1	!!!!!!!!! RT @C_G_Anderson: @viva_based she lo...	Offensive language
4	4	6	0	6	0	1	!!!!!!!!!!!!! RT @ShenikaRoberts: The shit you...	Offensive language
...	...	...	...	...	...	...	...	...
24778	25291	3	0	2	1	1	you's a muthaf***in lie &#8220;@LifeAsKing: @2...	Offensive language
24779	25292	3	0	1	2	2	you've gone and broke the wrong heart baby, an...	No hate nor offensive
24780	25294	3	0	3	0	1	young buck wanna eat!!.. dat nigguh like I ain...	Offensive language
24781	25295	6	0	6	0	1	youu got wild bitches tellin you lies	Offensive language
24782	25296	3	0	0	3	2	~~Ruffled | Ntac Eileen Dahlia - Beautiful col...	No hate nor offensive

24783 rows × 8 columns

In [17]:

```
data = dataset[["tweet","labels"]]
```

In [18]:

```
data
```

Out[18]:

	tweet	labels
0	!!! RT @mayasolovely: As a woman you shouldn't...	No hate nor offensive
1	!!!!! RT @mleew17: boy dats cold...tyga dwn ba...	Offensive language
2	!!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby...	Offensive language
3	!!!!!!!!! RT @C_G_Anderson: @viva_based she lo...	Offensive language
4	!!!!!!!!!!!!! RT @ShenikaRoberts: The shit you...	Offensive language
...	...	...
24778	you's a muthaf***in lie &#8220;@LifeAsKing: @2...	Offensive language
24779	you've gone and broke the wrong heart baby, an...	No hate nor offensive
24780	young buck wanna eat!!.. dat nigguh like I ain...	Offensive language
24781	youu got wild bitches tellin you lies	Offensive language
24782	~~Ruffled | Ntac Eileen Dahlia - Beautiful col...	No hate nor offensive

24783 rows × 2 columns

In [19]:

```
import re
import nltk
import string
```

In [20]:

```
from nltk.corpus import stopwords
stopwords=set(stopwords.words('english'))
```

In [21]:

```
stemmer=nltk.SnowballStemmer('english')
```

In [22]:

```
def  clean_data(text):
    text=str(text).lower()
    text=re.sub('https ?://\S+|www\.S+', '',text)
    text=re.sub('\[.*?\]', '',text)
    text=re.sub('<,*?>+', '',text)
    text=re.sub('[%s]'%re.escape(string.punctuation), '',text)
    text=re.sub('\n', '',text)
    text=re.sub('\w*\d\w*', '',text)
    text=[word for word in text.split(' ') if word not in stopwords]
    text=" ".join(text)
    text=[stemmer.stem(word) for word in text.split(' ')]
    text=" ".join(text)
    return text
```

In [23]:

```
data["tweet"] = data["tweet"].apply(clean_data)
```

```
C:\Users\ASUS\AppData\Local\Temp\ipykernel_14804\1832165696.py:1: SettingWithCopyWarnin
g:
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
  data["tweet"] = data["tweet"].apply(clean_data)
```

In [24]:

```
data
```

Loading [MathJax]/extensions/Safe.js

Out[24]:

	tweet	labels
0	rt mayasolov woman shouldnt complain clean ho...	No hate nor offensive
1	rt boy dat coldtyga dwn bad cuffin dat hoe ...	Offensive language
2	rt urkindofbrand dawg rt ever fuck bitch sta...	Offensive language
3	rt cganderson vivabas look like tranni	Offensive language
4	rt shenikarobert shit hear might true might f...	Offensive language
...	...	...
24778	yous muthafin lie coreyemanuel right tl tras...	Offensive language
24779	youv gone broke wrong heart babi drove redneck...	No hate nor offensive
24780	young buck wanna eat dat nigguh like aint fuck...	Offensive language
24781	youu got wild bitch tellin lie	Offensive language
24782	ruffl ntac eileen dahlia beauti color combin...	No hate nor offensive

24783 rows × 2 columns

In [25]:

```
x= np.array(data["tweet"])
y= np.array(data["labels"])
```

In [26]:

```
x
```

Out[26]:

```
array([' rt mayasolov woman shouldnt complain clean hous amp man alway take trash',
       ' rt  boy dat coldtyga dwn bad cuffin dat hoe  place',
       ' rt urkindofbrand dawg rt  ever fuck bitch start cri confus shit',
       ..., 'young buck wanna eat dat nigguh like aint fuckin dis',
       'youu got wild bitch tellin lie',
       'ruffl  ntac eileen dahlia  beauti color combin pink orang yellow amp white coll
'],
      dtype=object)
```

In [27]:

```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import train_test_split
```

In [28]:

```
cv= CountVectorizer()
x= cv.fit_transform(x)
```

In [29]:

```
x
```

Out[29]:

```
<24783x26127 sparse matrix of type '<class 'numpy.int64'>'
	with 198269 stored elements in Compressed Sparse Row format>
```

In [30]:

```
x_train, x_test, y_train, y_test= train_test_split(x, y, test_size=0.33, random_state=42
```

In [31]:

```
x_train
```

Out[31]:

```
<16604x26127 sparse matrix of type '<class 'numpy.int64'>'
	with 132883 stored elements in Compressed Sparse Row format>
```

In [32]:

```
from sklearn.tree import DecisionTreeClassifier
```

In [33]:

```
dt= DecisionTreeClassifier()
dt.fit(x_train, y_train)
```

Out[33]:

```
DecisionTreeClassifier()
```

Loading [MathJax]/extensions/Safe.js

```
In [34]: y_pred= dt.predict(x_test)
```

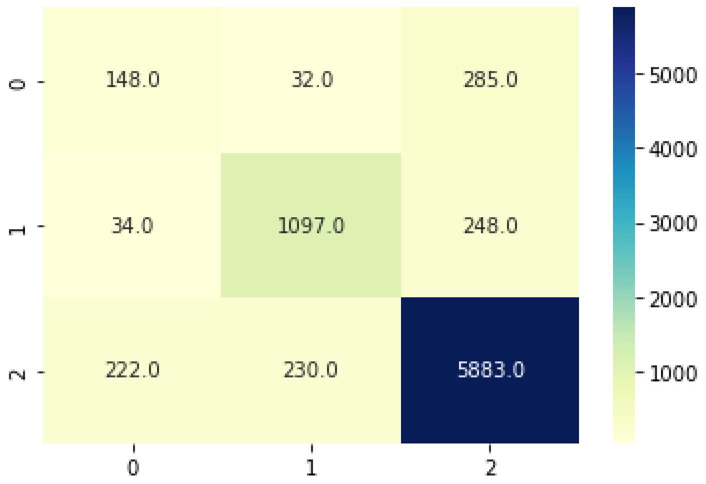
```
In [35]: from sklearn.metrics import confusion_matrix
         cm= confusion_matrix(y_test, y_pred)
         cm
```

```
Out[35]: array([[ 148,   32,  285],
                [  34, 1097,  248],
                [ 222,  230, 5883]], dtype=int64)
```

```
In [36]: import seaborn as sns
         import matplotlib.pyplot as ply
         %matplotlib inline
```

```
In [37]: sns.heatmap(cm, annot= True, fmt=".1f", cmap="YlGnBu")
```

```
Out[37]: <AxesSubplot:>
```



```
In [38]: from sklearn.metrics import accuracy_score
         accuracy_score(y_test, y_pred)
```

```
Out[38]: 0.8715001833965033
```

```
In [39]: sample= "Let's unite and kill all the people who are protesting against the government"
         sample= clean_data(sample)
         sample
```

```
Out[39]: 'let unit kill peopl protest govern'
```

```
In [40]: data1= cv.transform([sample]).toarray()
```

```
In [41]: data1
```

```
Out[41]: array([[0, 0, 0, ..., 0, 0, 0]], dtype=int64)
```

```
In [42]: dt.predict(data1)
```

```
Out[42]: array(['Hate speech'], dtype=object)
```

```
In [43]: sample1= "Yummy, I wanna eat you up"
         sample1= clean_data(sample1)
         sample1
```

```
Out[43]: 'yummi wanna eat'
```

Loading [MathJax]/extensions/Safe.js

```
In [44]: data2= cv.transform([sample1]).toarray()
```

```
In [45]: data2
```

```
Out[45]: array([[0, 0, 0, ..., 0, 0, 0]], dtype=int64)
```

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

```
Out[46]: array(['No hate nor offensive'], dtype=object)
```

```
In [ ]:
```

Loading [MathJax]/extensions/Safe.js