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

In [2]:
```
dataset=pd.read_csv('twitter_data.csv')
dataset
```

Out[2]:

	count	hate_speech_count	offensive_language_count	neither_count	class	tweet
0	3	0	0	3	2	!!! RT @mayasolovely: As a woman you shouldn't...
1	3	0	3	0	1	!!!!! RT @mleew17: boy dats cold...tyga dwn ba...
2	3	0	3	0	1	!!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby...
3	3	0	2	1	1	!!!!!!!!! RT @C_G_Anderson: @viva_based she lo...
4	6	0	6	0	1	!!!!!!!!!!!! RT @ShenikaRoberts: The shit you...
...	...	...	...	...	...	...
24778	3	0	2	1	1	you's a muthaf***in lie “@LifeAsKing: @2...
24779	3	0	1	2	2	you've gone and broke the wrong heart baby, an...
24780	3	0	3	0	1	young buck wanna eat!!.. dat nigguh like I ain...
24781	6	0	6	0	1	youu got wild bitches tellin you lies
24782	3	0	0	3	2	~~Ruffled | Ntac Eileen Dahlia - Beautiful col...

24783 rows × 6 columns

In [3]:
```
dataset.isnull().sum()
```

Out[3]:
```
count                       0
hate_speech_count           0
offensive_language_count    0
neither_count               0
class                       0
tweet                       0
dtype: int64
```

In [4]:
```
dataset["labels"]=dataset["class"].map({0:"Hate Speech",
                                        1:"Offensive language",
                                        2:"No hate or offensive Language"})
```

In [5]:
```
data=dataset[["tweet","labels"]]
data
```

Out[5]:

	tweet	labels
0	!!! RT @mayasolovely: As a woman you shouldn't...	No hate or offensive Language
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 “@LifeAsKing: @2...	Offensive language
24779	you've gone and broke the wrong heart baby, an...	No hate or offensive Language
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 or offensive Language

24783 rows × 2 columns

In [6]:
```
import re
import nltk
import string
```

In [7]:
```
stopwords = set(nltk.corpus.stopwords.words("english"))
```

In [8]:
```
stopwords.add("rt")
```

```
stopwords
```

Out[8]:
```
{'a',
 'about',
 'above',
 'after',
 'again',
 'against',
 'ain',
 'all',
 'am',
 'an',
 'and',
 'any',
 'are',
 'aren',
 "aren't",
 'as',
 'at',
 'be',
 'because',
 'been',
 'before',
 'being',
 'below',
 'between',
 'both',
 'but',
 'by',
 'can',
 'couldn',
 "couldn't",
 'd',
 'did',
 'didn',
 "didn't",
 'do',
 'does',
 'doesn',
 "doesn't",
 'doing',
 'don',
 "don't",
 'down',
 'during',
 'each',
 'few',
 'for',
 'from',
 'further',
 'had',
 'hadn',
 "hadn't",
 'has',
 'hasn',
 "hasn't",
 'have',
 'haven',
 "haven't",
 'having',
 'he',
 'her',
 'here',
 'hers',
 'herself',
 'him',
 'himself',
 'his',
 'how',
 'i',
 'if',
 'in',
 'into',
 'is',
 'isn',
 "isn't",
 'it',
 "it's",
 'its',
 'itself',
 'just',
 'll',
 'm',
```

```
'ma',
'me',
'mightn',
"mightn't",
'more',
'most',
'mustn',
"mustn't",
'my',
'myself',
'needn',
"needn't",
'no',
'nor',
'not',
'now',
'o',
'of',
'off',
'on',
'once',
'only',
'or',
'other',
'our',
'ours',
'ourselves',
'out',
'over',
'own',
're',
'rt',
's',
'same',
'shan',
"shan't",
'she',
"she's",
'should',
"should've",
'shouldn',
"shouldn't",
'so',
'some',
'such',
't',
'than',
'that',
"that'll",
'the',
'their',
'theirs',
'them',
'themselves',
'then',
'there',
'these',
'they',
'this',
'those',
'through',
'to',
'too',
'under',
'until',
'up',
've',
'very',
'was',
'wasn',
"wasn't",
'we',
'were',
'weren',
"weren't",
'what',
'when',
'where',
'which',
'while',
'who',
'whom',
'why',
```

```
'will',
'with',
'won',
"won't",
'wouldn',
"wouldn't",
'y',
'you',
"you'd",
"you'll",
"you're",
"you've",
'your',
'yours',
'yourself',
'yourselves'}
```

In [9]:
```
stemmer = nltk.SnowballStemmer("english")
```

In [10]:
```
def data_clean(text):
    text = str(text).lower()
    text = re.sub("https?://S+wwe\.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)
    words = [stemmer.stem(word) for word in text.split(' ') if word not in stopwords]
    text = " ".join(words)
    return text
```

In [11]:
```
data["tweet"] = data["tweet"].apply(data_clean)
```

```
C:\Users\999\AppData\Local\Temp\ipykernel_8956\2274407121.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#retu
rning-a-view-versus-a-copy
  data["tweet"] = data["tweet"].apply(data_clean)
```

In [12]:
```
data
```

Out[12]:

	tweet	labels
0	mayasolov woman shouldnt complain clean hous ...	No hate or offensive Language
1	boy dat coldtyga dwn bad cuffin dat hoe place	Offensive language
2	urkindofbrand dawg ever fuck bitch start cri...	Offensive language
3	cganderson vivabas look like tranni	Offensive language
4	shenikarobert shit hear might true might fake...	Offensive language
...	...	...
24778	yous muthafin lie coreyemanuel right tl tras...	Offensive language
24779	youv gone broke wrong heart babi drove redneck...	No hate or offensive Language
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 or offensive Language

24783 rows × 2 columns

In [13]:
```
x = np.array(data["tweet"])
y = np.array(data["labels"])
```

In [15]:
```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import train_test_split
```

In [16]:
```
cv = CountVectorizer()
x = cv.fit_transform(x)
```

In [17]:
```
x
```

Out[17]:
```
<24783x26151 sparse matrix of type '<class 'numpy.int64'>'
        with 191166 stored elements in Compressed Sparse Row format>
```

In [18]:
```
x_train,x_test,y_train,y_test = train_test_split(x, y, test_size=0.3)
```

In [19]:
```
from sklearn.tree import DecisionTreeClassifier
```

In [29]:
```
dt=DecisionTreeClassifier(max_depth=100)
dt.fit(x_train,y_train)
```

Out[29]:
```
DecisionTreeClassifier(max_depth=100)
```

In [22]:
```
y_pred=dt.predict(x_test)
```

In [23]:
```
from sklearn.metrics import classification_report,confusion_matrix
```

In [24]:
```
cm=confusion_matrix(y_test,y_pred)
cm
```

Out[24]:
```
array([[ 157,   36,  239],
       [  41, 1018,  152],
       [ 249,  226, 5317]], dtype=int64)
```

In [30]:
```
print(classification_report(y_test,y_pred))
```

```
                              precision    recall  f1-score   support

                  Hate Speech       0.35      0.36      0.36       432
No hate or offensive Language       0.80      0.84      0.82      1211
           Offensive language       0.93      0.92      0.92      5792

                    accuracy                           0.87      7435
                   macro avg       0.69      0.71      0.70      7435
                weighted avg       0.88      0.87      0.87      7435
```

In [31]:
```
dt.max_depth
```

Out[31]:
```
100
```

In [ ]: