```
import pandas as pd
import numpy as np
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier

import nltk
import re
from nltk.corpus import stopwords
stopword = set(stopwords.words('english'))
stemmer = nltk.SnowballStemmer("english")

data=pd.read_csv("data.csv")
print(data.head())
```

```
   Unnamed: 0  count  hate_speech  offensive_language  neither  class
\
0           0      3            0                   0        3      2

1           1      3            0                   3        0      1

2           2      3            0                   3        0      1

3           3      3            0                   2        1      1

4           4      6            0                   6        0      1


                                               tweet
0  !!! RT @mayasolovely: As a woman you shouldn't...
1  !!!!! RT @mleew17: boy dats cold...tyga dwn ba...
2  !!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby...
3  !!!!!!!!! RT @C_G_Anderson: @viva_based she lo...
4  !!!!!!!!!!!! RT @ShenikaRoberts: The shit you...
```

```
data["labels"]=data["class"].map({0:"Hate Speech", 1:"Offensive
Speech", 2: "No Hate and Offensive Speech"})

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

data.head()
```

```
                                               tweet  \
0  !!! RT @mayasolovely: As a woman you shouldn't...
1  !!!!! RT @mleew17: boy dats cold...tyga dwn ba...
2  !!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby...
3  !!!!!!!!! RT @C_G_Anderson: @viva_based she lo...
4  !!!!!!!!!!!! RT @ShenikaRoberts: The shit you...

                         labels
0  No Hate and Offensive Speech
```

```
1               Offensive Speech
2               Offensive Speech
3               Offensive Speech
4               Offensive Speech
```

```
import re

def clean (text):
    text=str (text). lower()
    text= re. sub('[.?]', '', text)
    text= re. sub('https?://\S+|www.\S+', '', text)
    text= re. sub('<.?>+', '', text)
    text= re. sub(r'[^\w\s]', '', text)
    text= re. sub('\n', '', text)
    text= re. sub('\w\d\w', '', text)
    text=[word for  word in text.split(' ') if word not in stopword]
    text=" ". join(text)
    text= [stemmer.stem(word) for word in text. split(' ')]
    text=" ". join(text)
    return text
data["tweet"] = data["tweet"].apply(clean)

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

cv=CountVectorizer()

X = cv.fit_transform(x)

X_train,X_test,y_train,y_test =
train_test_split(X,y,test_size=0.33,random_state=42)

model = DecisionTreeClassifier()

model.fit(X_train,y_train)
```

```
DecisionTreeClassifier()
```

```
y_pred=model.predict(X_test)

from sklearn.metrics import accuracy_score
print(accuracy_score(y_test,y_pred))
```

```
0.8774911358356767
```

```
i="You are too bad and I dont like your attitude"
i = cv.transform([i]).toarray()
print(model.predict((i)))
```

```
['No Hate and Offensive Speech']
```

```
i="wommen belong in kitchen"
i = cv.transform([i]).toarray()
print(model.predict((i)))
```

```
['No Hate and Offensive Speech']
```

```
i="Fuck you"
i = cv.transform([i]).toarray()
print(model.predict((i)))
```

```
['Offensive Speech']
```

```
i="niggas gonna get corona"
i = cv.transform([i]).toarray()
print(model.predict((i)))
```

```
['No Hate and Offensive Speech']
```