# 1stopproject1

October 6, 2023

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

[2]:
```
dataset = pd.read_csv("labeled_data.csv")
```

[3]:
```
dataset
```

[3]:
```
       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
…             …    …          …                   …       …    …
24778       25291      3            0                   2        1      1
24779       25292      3            0                   1        2      2
24780       25294      3            0                   3        0      1
24781       25295      6            0                   6        0      1
24782       25296      3            0                   0        3      2

                                                   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…
…                                                   …
24778  you's a muthaf***in lie &#8220;@LifeAsKing: @2…
24779  you've gone and broke the wrong heart baby, an…
24780  young buck wanna eat!!.. dat nigguh like I ain…
24781                youu got wild bitches tellin you lies
24782  ~~Ruffled | Ntac Eileen Dahlia - Beautiful col…

[24783 rows x 7 columns]
```

[4]:
```
dataset.isnull()
```

[4]:
```
       Unnamed: 0  count  hate_speech  offensive_language  neither  class  \
0           False  False        False               False    False  False
1           False  False        False               False    False  False
2           False  False        False               False    False  False
3           False  False        False               False    False  False
4           False  False        False               False    False  False
…             …    …            …                   …        …      …
24778       False  False        False               False    False  False
24779       False  False        False               False    False  False
24780       False  False        False               False    False  False
24781       False  False        False               False    False  False
24782       False  False        False               False    False  False

       tweet
0      False
1      False
2      False
3      False
4      False
…      …
24778  False
24779  False
24780  False
24781  False
24782  False

[24783 rows x 7 columns]
```

[5]:
```
dataset.isnull().sum()
```

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

[6]:
```
dataset.describe()
```

[6]:
```
         Unnamed: 0         count   hate_speech  offensive_language  \
count  24783.000000  24783.000000  24783.000000        24783.000000
mean   12681.192027      3.243473      0.280515            2.413711
std     7299.553863      0.883060      0.631851            1.399459
min        0.000000      3.000000      0.000000            0.000000
25%     6372.500000      3.000000      0.000000            2.000000
```

```
50%    12703.000000      3.000000      0.000000            3.000000
75%    18995.500000      3.000000      0.000000            3.000000
max    25296.000000      9.000000      7.000000            9.000000

            neither         class
count  24783.000000  24783.000000
mean       0.549247      1.110277
std        1.113299      0.462089
min        0.000000      0.000000
25%        0.000000      1.000000
50%        0.000000      1.000000
75%        0.000000      1.000000
max        9.000000      2.000000
```

[7]:
```
dataset["labels"] = dataset["class"].map({0: "Hate Speech",
                                          1: "Offensive Language",
                                          2: "No hate or offensive language"})
```

[8]:
```
dataset
```

[8]:
```
       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
…             …    …          …                   …      …    …
24778       25291      3            0                   2        1      1
24779       25292      3            0                   1        2      2
24780       25294      3            0                   3        0      1
24781       25295      6            0                   6        0      1
24782       25296      3            0                   0        3      2

                                                    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…
…                                                    …
24778  you's a muthaf***in lie &#8220;@LifeAsKing: @2…
24779  you've gone and broke the wrong heart baby, an…
24780  young buck wanna eat!!.. dat nigguh like I ain…
24781                youu got wild bitches tellin you lies
24782  ~~Ruffled | Ntac Eileen Dahlia - Beautiful col…

                           labels
```

```
0      No hate or offensive language
1                  Offensive Language
2                  Offensive Language
3                  Offensive Language
4                  Offensive Language
…                                 …
24778              Offensive Language
24779  No hate or offensive language
24780              Offensive Language
24781              Offensive Language
24782  No hate or offensive language

[24783 rows x 8 columns]
```

[9]:
```
data = dataset[["tweet", "labels"]]
```

[10]:
```
data
```

[10]:
```
                                                   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…
…                                                   …
24778  you's a muthaf***in lie &#8220;@LifeAsKing: @2…
24779  you've gone and broke the wrong heart baby, an…
24780  young buck wanna eat!!.. dat nigguh like I ain…
24781              youu got wild bitches tellin you lies
24782  ~~Ruffled | Ntac Eileen Dahlia - Beautiful col…

                             labels
0      No hate or offensive language
1                  Offensive Language
2                  Offensive Language
3                  Offensive Language
4                  Offensive Language
…                                 …
24778              Offensive Language
24779  No hate or offensive language
24780              Offensive Language
24781              Offensive Language
24782  No hate or offensive language

[24783 rows x 2 columns]
```

[11]:
```
import re
import nltk
nltk.download("stopwords")
import string
```

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

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

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

[14]:
```
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
```

[15]:
```
data["tweet"] = data['tweet'].apply(clean_data)
```

```
C:\Users\prave\AppData\Local\Temp\ipykernel_2520\962770371.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
  data["tweet"] = data['tweet'].apply(clean_data)
```

[16]:
```
data
```

[16]:

| | tweet |
|---|---|
| 0 | rt mayasolov as a woman you shouldnt complain… |
| 1 | rt boy dat coldtyga dwn bad for cuffin dat ho… |
| 2 | rt urkindofbrand dawg rt you ever fuck a bitc… |
| 3 | rt cganderson vivabas she look like a tranni |
| 4 | rt shenikarobert the shit you hear about me m… |

```
…                                                     …
24778  yous a muthafin lie  coreyemanuel right his tl…
24779  youv gone and broke the wrong heart babi and d…
24780  young buck wanna eat dat nigguh like i aint fu…
24781                 youu got wild bitch tellin you lie
24782  ruffl ntac eileen dahlia beauti color combin o…

                             labels
0      No hate or offensive language
1                  Offensive Language
2                  Offensive Language
3                  Offensive Language
4                  Offensive Language
…                                …
24778              Offensive Language
24779  No hate or offensive language
24780              Offensive Language
24781              Offensive Language
24782  No hate or offensive language

[24783 rows x 2 columns]
```

[17]:
```
X = np.array(data["tweet"])
Y = np.array(data["labels"])
```

[18]:
```
X
```

[18]:
```
array([' rt mayasolov as a woman you shouldnt complain about clean up your hous
amp as a man you should alway take the trash out',
       ' rt boy dat coldtyga dwn bad for cuffin dat hoe in the place',
       ' rt urkindofbrand dawg rt you ever fuck a bitch and she start to cri you
be confus as shit',
       …,
       'young buck wanna eat dat nigguh like i aint fuckin dis up again',
       'youu got wild bitch tellin you lie',
       'ruffl ntac eileen dahlia beauti color combin of pink orang yellow amp
white a coll '],
      dtype=object)
```

[19]:
```
Y
```

[19]:
```
array(['No hate or offensive language', 'Offensive Language',
       'Offensive Language', …, 'Offensive Language',
       'Offensive Language', 'No hate or offensive language'],
      dtype=object)
```

[20]:
```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import train_test_split
```

[21]:
```
cv = CountVectorizer()
X = cv.fit_transform(X)
```

[22]:
```
X
```

[22]:
```
<24783x25784 sparse matrix of type '<class 'numpy.int64'>'
        with 291489 stored elements in Compressed Sparse Row format>
```

[23]:
```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.33,␣
 ↪random_state=42)
```

[24]:
```
X_train
```

[24]:
```
<16604x25784 sparse matrix of type '<class 'numpy.int64'>'
        with 195519 stored elements in Compressed Sparse Row format>
```

[25]:
```
from sklearn.tree import DecisionTreeClassifier
```

[26]:
```
dt = DecisionTreeClassifier()
dt.fit(X_train, Y_train)
```

[26]:
```
DecisionTreeClassifier()
```

[27]:
```
Y_pred = dt.predict(X_test)
```

[28]:
```
Y_pred
```

[28]:
```
array(['Offensive Language', 'Offensive Language', 'Offensive Language',
       …, 'No hate or offensive language',
       'No hate or offensive language', 'Offensive Language'],
      dtype=object)
```

[29]:
```
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(Y_test, Y_pred)
```

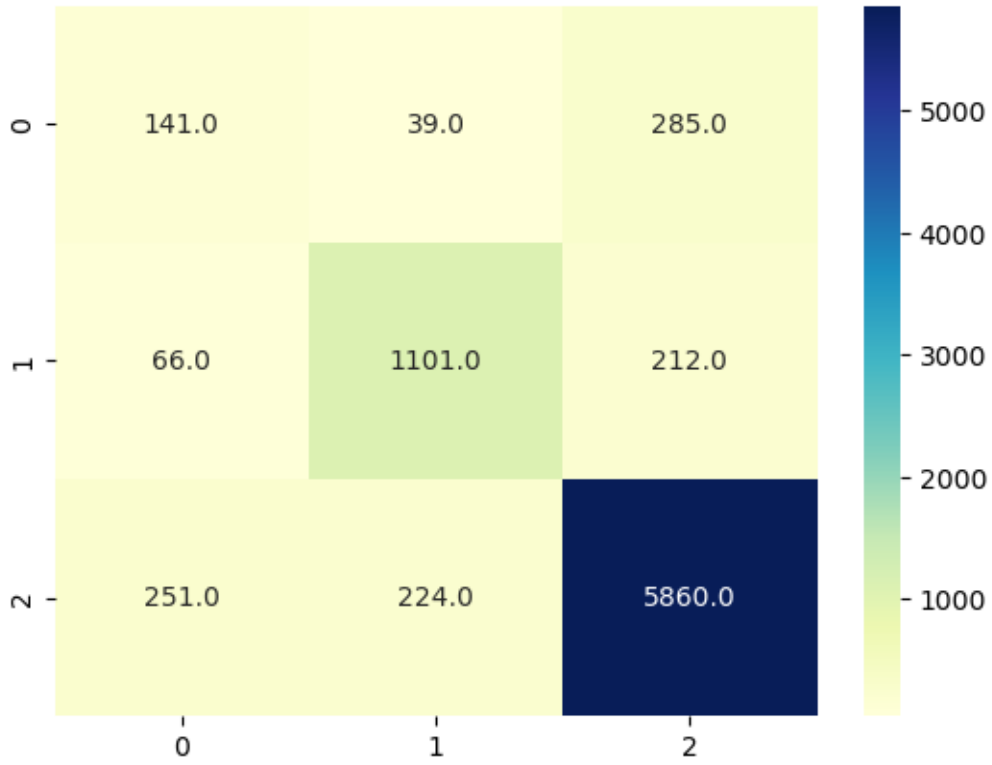
[30]:
```
cm
```

[30]:
```
array([[ 141,   39,  285],
       [  66, 1101,  212],
       [ 251,  224, 5860]], dtype=int64)
```

[31]:
```
import seaborn as sns
import matplotlib.pyplot as ply
%matplotlib inline
```

[32]:
```
sns.heatmap(cm, annot = True, fmt =".1f", cmap="YlGnBu")
```

[32]: <Axes: >




[33]:
```
from sklearn.metrics import accuracy_score
accuracy_score(Y_test,Y_pred)
```

[33]: 0.8683213106736765

[34]:
```
sample = "Let's unite and make this world a better place"
sample = clean_data(sample)
```

[35]:
```
sample
```

[35]: 'let unit and make this world a better place'

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

[37]:
```
data1
```

[37]: array([[0, 0, 0, …, 0, 0, 0]], dtype=int64)

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

[38]:
```
array(['No hate or offensive language'], dtype=object)
```

[ ]:
```

```