Tanvi Bhosle.

Internship Task @Youth India Foundation E-School

In [ ]:

```
```

Import Libraries

In [4]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings("ignore")
import neattext.functions as nfx
from textblob import TextBlob
from collections import Counter
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import f1_score
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix,classification_report
```

Import Dataset

In [6]:

```
data=pd.read_csv("C:\\Users\\Admin\\Desktop\\xyz.csv",encoding=("ISO-8859-1"),low_memory=Fa
data.head(5)
```

Out[6]:

	ItemID	Sentiment	SentimentText
0	1	0	is so sad for my APL frie...
1	2	0	I missed the New Moon trail...
2	3	1	omg its already 7:30 :O
3	4	0	.. Omgaga. Im sooo im gunna CRy. I'...
4	5	0	i think mi bf is cheating on me!!! ...

In [7]:

```
data.dtypes
```

Out[7]:

```
ItemID            int64
Sentiment         int64
SentimentText    object
dtype: object
```

In [8]:

```
data["SentimentText"]
```

Out[8]:

```
0                             is so sad for my APL frie...
1                           I missed the New Moon trail...
2                                  omg its already 7:30 :O
3                  .. Omgaga. Im sooo  im gunna CRy. I'...
4                 i think mi bf is cheating on me!!!   ...
                                ...
99984    @Cupcake  seems like a repeating problem   hop...
99985    @cupcake__ arrrr we both replied to each other...
99986                         @CuPcAkE_2120 ya i thought so
99987    @Cupcake_Dollie Yes. Yes. I'm glad you had mor...
99988                        @cupcake_kayla haha yes you do
Name: SentimentText, Length: 99989, dtype: object
```

Text Cleaning

In [9]:

```
data["clean_Text"]=data["SentimentText"].apply(nfx.remove_hashtags)
data[["SentimentText", "clean_Text"]]
```

Out[9]:

	SentimentText	clean_Text
0	is so sad for my APL frie...	is so sad for my APL frie...
1	I missed the New Moon trail...	I missed the New Moon trail...
2	omg its already 7:30 :O	omg its already 7:30 :O
3	.. Omgaga. Im sooo im gunna CRy. I'...	.. Omgaga. Im sooo im gunna CRy. I'...
4	i think mi bf is cheating on me!!! ...	i think mi bf is cheating on me!!! ...
...	...	...
99984	@Cupcake seems like a repeating problem hop...	@Cupcake seems like a repeating problem hop...
99985	@cupcake__ arrrr we both replied to each other...	@cupcake__ arrrr we both replied to each other...
99986	@CuPcAkE_2120 ya i thought so	@CuPcAkE_2120 ya i thought so
99987	@Cupcake_Dollie Yes. Yes. I'm glad you had mor...	@Cupcake_Dollie Yes. Yes. I'm glad you had mor...
99988	@cupcake_kayla haha yes you do	@cupcake_kayla haha yes you do

99989 rows × 2 columns

In [10]:

```
data["clean_Text"]=data["clean_Text"].apply(lambda x:nfx.remove_userhandles(x))
data[["SentimentText", "clean_Text"]]
```

Out[10]:

	SentimentText	clean_Text
0	is so sad for my APL frie...	is so sad for my APL frie...
1	I missed the New Moon trail...	I missed the New Moon trail...
2	omg its already 7:30 :O	omg its already 7:30 :O
3	.. Omgaga. Im sooo im gunna CRy. I'...	.. Omgaga. Im sooo im gunna CRy. I'...
4	i think mi bf is cheating on me!!! ...	i think mi bf is cheating on me!!! ...
...	...	...
99984	@Cupcake seems like a repeating problem hop...	seems like a repeating problem hope you'r...
99985	@cupcake__ arrrr we both replied to each other...	arrrr we both replied to each other over dif...
99986	@CuPcAkE_2120 ya i thought so	ya i thought so
99987	@Cupcake_Dollie Yes. Yes. I'm glad you had mor...	Yes. Yes. I'm glad you had more fun with me.
99988	@cupcake_kayla haha yes you do	haha yes you do

99989 rows × 2 columns

In [11]:

```
data["clean_Text"]=data["clean_Text"].apply(nfx.remove_multiple_spaces)
data[["SentimentText", "clean_Text"]]
```

Out[11]:

	SentimentText	clean_Text
0	is so sad for my APL frie...	is so sad for my APL friend.....
1	I missed the New Moon trail...	I missed the New Moon trailer...
2	omg its already 7:30 :O	omg its already 7:30 :O
3	.. Omgaga. Im sooo im gunna CRy. I'...	.. Omgaga. Im sooo im gunna CRy. I've been at...
4	i think mi bf is cheating on me!!! ...	i think mi bf is cheating on me!!! T_T
...	...	...
99984	@Cupcake seems like a repeating problem hop...	seems like a repeating problem hope you're ab...
99985	@cupcake__ arrrr we both replied to each other...	arrrr we both replied to each other over diff...
99986	@CuPcAkE_2120 ya i thought so	ya i thought so
99987	@Cupcake_Dollie Yes. Yes. I'm glad you had mor...	Yes. Yes. I'm glad you had more fun with me.
99988	@cupcake_kayla haha yes you do	haha yes you do

99989 rows × 2 columns

In [13]:

```
data["clean_Text"]=data["clean_Text"].apply(nfx.remove_urls)
data[["SentimentText", "clean_Text"]]
```

Out[13]:

	SentimentText	clean_Text
0	is so sad for my APL frie...	is so sad for my APL friend.....
1	I missed the New Moon trail...	I missed the New Moon trailer...
2	omg its already 7:30 :O	omg its already 7:30 :O
3	.. Omgaga. Im sooo im gunna CRy. I'...	.. Omgaga. Im sooo im gunna CRy. I've been at...
4	i think mi bf is cheating on me!!! ...	i think mi bf is cheating on me!!! T_T
...	...	...
99984	@Cupcake seems like a repeating problem hop...	seems like a repeating problem hope you're ab...
99985	@cupcake__ arrrr we both replied to each other...	arrrr we both replied to each other over diff...
99986	@CuPcAkE_2120 ya i thought so	ya i thought so
99987	@Cupcake_Dollie Yes. Yes. I'm glad you had mor...	Yes. Yes. I'm glad you had more fun with me.
99988	@cupcake_kayla haha yes you do	haha yes you do

99989 rows × 2 columns

In [14]:

```
data["clean_Text"]=data["clean_Text"].apply(nfx.remove_puncts)
data[["SentimentText", "clean_Text"]]
```

Out[14]:

	SentimentText	clean_Text
0	is so sad for my APL frie...	is so sad for my APL friend
1	I missed the New Moon trail...	I missed the New Moon trailer
2	omg its already 7:30 :O	omg its already 7:30 :O
3	.. Omgaga. Im sooo im gunna CRy. I'...	Omgaga Im sooo im gunna CRy Ive been at this...
4	i think mi bf is cheating on me!!! ...	i think mi bf is cheating on me TT
...	...	...
99984	@Cupcake seems like a repeating problem hop...	seems like a repeating problem hope youre abl...
99985	@cupcake__ arrrr we both replied to each other...	arrrr we both replied to each other over diff...
99986	@CuPcAkE_2120 ya i thought so	ya i thought so
99987	@Cupcake_Dollie Yes. Yes. I'm glad you had mor...	Yes Yes Im glad you had more fun with me
99988	@cupcake_kayla haha yes you do	haha yes you do

99989 rows × 2 columns

In [15]:

```
data[["SentimentText", "clean_Text"]]
```

Out[15]:

	SentimentText	clean_Text
0	is so sad for my APL frie...	is so sad for my APL friend
1	I missed the New Moon trail...	I missed the New Moon trailer
2	omg its already 7:30 :O	omg its already 7:30 :O
3	.. Omgaga. Im sooo im gunna CRy. I'...	Omgaga Im sooo im gunna CRy Ive been at this...
4	i think mi bf is cheating on me!!! ...	i think mi bf is cheating on me TT
...	...	...
99984	@Cupcake seems like a repeating problem hop...	seems like a repeating problem hope youre abl...
99985	@cupcake__ arrrr we both replied to each other...	arrrr we both replied to each other over diff...
99986	@CuPcAkE_2120 ya i thought so	ya i thought so
99987	@Cupcake_Dollie Yes. Yes. I'm glad you had mor...	Yes Yes Im glad you had more fun with me
99988	@cupcake_kayla haha yes you do	haha yes you do

99989 rows × 2 columns

Predicted Sentiment

In [16]:

```
def get_sentiment(text):
    blob = TextBlob(text)
    sentiment_polarity = blob.sentiment.polarity
    sentiment_subjectivity = blob.sentiment.subjectivity
    if sentiment_polarity > 0:
        prdicted_sentiment = "Happy"
    elif sentiment_polarity < 0:
        prdicted_sentiment = "Sad"
    else:
        prdicted_sentiment = "Neutral"
    result = {"polarity": sentiment_polarity,
              "subjectivity": sentiment_subjectivity,
              "pred sentiment": prdicted_sentiment}
    return result
```

In [17]:

```
get_sentiment(data["clean_Text"].iloc[10])
```

Out[17]:

```
{'polarity': 0.22727272727272727,
 'subjectivity': 0.5454545454545454,
 'pred sentiment': 'Happy'}
```

In [18]:

```
data["pred_sentiment"]=data["clean_Text"].apply(get_sentiment)
data["pred_sentiment"]
```

Out[18]:

```
0        {'polarity': -0.5, 'subjectivity': 1.0, 'pred ...
1        {'polarity': 0.13636363636363635, 'subjectivit...
2        {'polarity': 0.05, 'subjectivity': 1.0, 'pred ...
3        {'polarity': 0.0, 'subjectivity': 0.0, 'pred s...
4        {'polarity': 0.0, 'subjectivity': 0.0, 'pred s...
                              ...
99984    {'polarity': 0.5, 'subjectivity': 0.625, 'pred...
99985    {'polarity': -0.15625, 'subjectivity': 0.525, ...
99986    {'polarity': 0.0, 'subjectivity': 0.0, 'pred s...
99987    {'polarity': 0.43333333333333335, 'subjectivit...
99988    {'polarity': 0.2, 'subjectivity': 0.3, 'pred s...
Name: pred_sentiment, Length: 99989, dtype: object
```

In [19]:

```
data["pred_sentiment"].iloc[10]
```

Out[19]:

```
{'polarity': 0.22727272727272727,
 'subjectivity': 0.5454545454545454,
 'pred sentiment': 'Happy'}
```

In [20]:

```
j=pd.json_normalize(data["pred_sentiment"])
j
```

Out[20]:

	polarity	subjectivity	pred sentiment
0	-0.500000	1.000000	Sad
1	0.136364	0.454545	Happy
2	0.050000	1.000000	Happy
3	0.000000	0.000000	Neutral
4	0.000000	0.000000	Neutral
...	...	...	...
99984	0.500000	0.625000	Happy
99985	-0.156250	0.525000	Sad
99986	0.000000	0.000000	Neutral
99987	0.433333	0.566667	Happy
99988	0.200000	0.300000	Happy

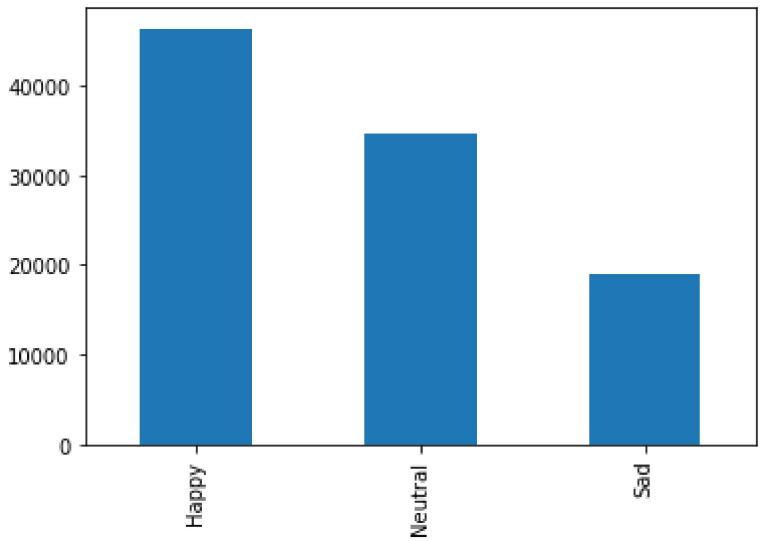
99989 rows × 3 columns

In [21]:

```
j["pred sentiment"].value_counts().plot(kind="bar")
j["pred sentiment"].value_counts()
```

Out[21]:

```
Happy      46325
Neutral    34697
Sad        18967
Name: pred sentiment, dtype: int64
```



In [22]:

```
pos_response=data[j["pred sentiment"]=="Happy"]["clean_Text"]
pos_response
pos_response.apply(nfx.remove_stopwords).tolist()
```

Out[22]:

```
['missed New Moon trailer',
 'omg 7:30 :O',
 'worry',
 'think positive',
 'thanks haters face day 112102',
 'weekend sucked far',
 'jb isnt showing australia',
 'ok thats win',
 'lt way feel right',
 'Feeling strangely fine Im gonna listen Semisonic celebrate',
 'cut beard growing year Im gonna start happy meantime',
 'Youre cause following youre pretty awesome',
 'Feeeling like shit right want sleep nooo 3 hours dancing art assignment
finish',
 'miss guys think im wearing skinny jeans cute sweater heels sure today',
 'SOX Floyd great relievers need scolding',
 '(: wrote week got new york office',
 'vear Lakers Thats magic fun'.
```

In [23]:

```
neg_response=data[j["pred sentiment"]=="Sad"]["clean_Text"]
neg_response
neg_response.apply(nfx.remove_stopwords).tolist()
```

Out[23]:

```
['sad APL friend',
 'awhhe man Im completely useless rt Funny twitter',
 'HUGE roll thunder nowSO scary',
 'sad Iran',
 'ltSad level 3 writing massive blog tweet Myspace comp shut lost *lays fe
tal position*',
 'BoRinG ): whats wrong tell :/',
 'cant bothered wish spend rest life sat going gigs seriously',
 'hate athlete appears tear ACL live television',
 'Dont Like Room Boring Sick Wardrobe Cant Waiit Till Walk Yay',
 'type spaz downloads virus brother thats :\\ MSN fucked forever :(',
 'makes sad look Muslims reality',
 'entertainment complained properly experimental experiment melody',
 'friends leaving cause stupid love',
 'hate u leysh t9ar5 =(((((((',
 'hate allergies hair cut tomorrow Im taking public poll',
 'slow $1 tix',
 'idk wat 2 trust im sorry 4 da pain caused nebody ima dis time 2 straight
```

In [24]:

```
neutral_response=data[j["pred sentiment"]=="Neutral"]["clean_Text"]
neutral_response
neutral_response.apply(nfx.remove_stopwords).tolist()
```

Out[24]:

```
['Omgaga Im sooo im gunna CRy Ive dentist 11 suposed 2 crown (30mins)',
 'think mi bf cheating TT',
 'Juuuuuuuuuuuuuuuuussssst Chillin',
 'Sunny Work Tomorrow :| TV Tonight',
 'handed uniform today miss',
 'hmmmm wonder number',
 'wompppp wompp',
 'Headed Hospitol : pull Golf Tourny 3rd place Think ReRipped Yeah',
 'goodbye exams HELLO ALCOHOL TONIGHT',
 'didnt realize deep Geez girl warning atleast',
 'Meet Meat',
 'horsie moving Saturday morning',
 'Sat offNeed work 6 days week',
 'times like million',
 'uploading pictures friendster',
 'ampampFightiin Wiit Babes',
 '*enough said*',
 'need anyways: CHRIS CORNELL CHICAGO TONIGHT'.
```

In [25]:

```
for line in pos_response:
    print(line)
    for token in line.split():
        print(token)
```

```
 I missed the New Moon trailer
I
missed
the
New
Moon
trailer
 omg its already 7:30 :O
omg
its
already
7:30
:O
 or i just worry too much
or
i
just
worry
too
much
```

In [29]:

```
pos_token=[token for line in pos_response for token in line.split()]
pos_token
```

Out[29]:

```
['I',
 'missed',
 'the',
 'New',
 'Moon',
 'trailer',
 'omg',
 'its',
 'already',
 '7:30',
 ':O',
 'or',
 'i',
 'just',
 'worry',
 'too',
 'much',
 'I'.
```

In [30]:

```
neg_token=[token for line in neg_response for token in line.split()]
neg_token
```

Out[30]:

```
['is',
 'so',
 'sad',
 'for',
 'my',
 'APL',
 'friend',
 'awhhe',
 'man',
 'Im',
 'completely',
 'useless',
 'rt',
 'now',
 'Funny',
 'all',
 'I',
 'can'.
```

In [31]:

```
neutral_token=[token for line in neutral_response for token in line.split()]
neutral_token
```

Out[31]:

```
['Omgaga',
 'Im',
 'sooo',
 'im',
 'gunna',
 'CRy',
 'Ive',
 'been',
 'at',
 'this',
 'dentist',
 'since',
 '11',
 'I',
 'was',
 'suposed',
 '2',
 'just',
```

Most Common Words Used

In [32]:

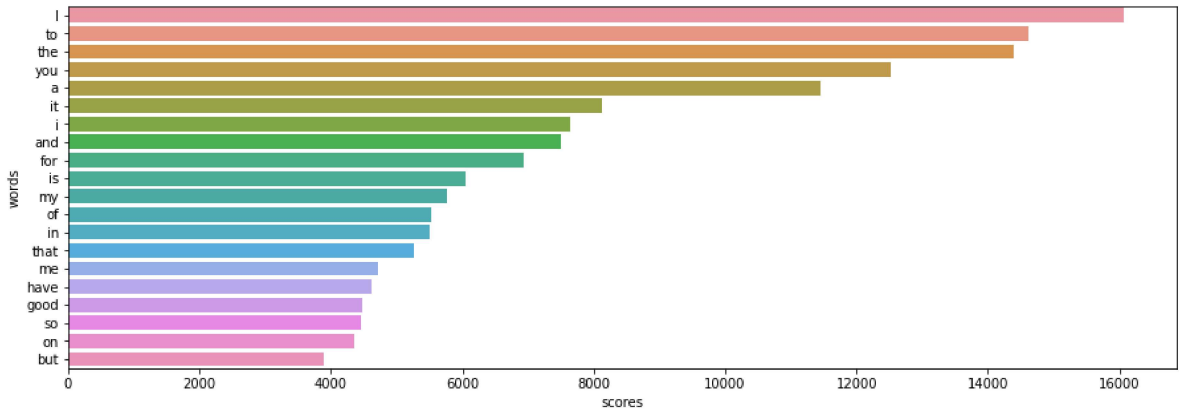
```
def get_tokens(docx,num=20):
    word_tokens = Counter(docx)
    most_common = word_tokens.most_common(num)
    result = dict(most_common)
    return result
```

In [33]:

```
most_comm_pos_words=get_tokens(pos_token)
most_comm_pos_words
pos=pd.DataFrame(most_comm_pos_words.items(),columns=["words","scores"])
pos
plt.figure(figsize=(15,5))
sns.barplot(y="words",x="scores",data=pos)
```

Out[33]:

```
<AxesSubplot:xlabel='scores', ylabel='words'>
```

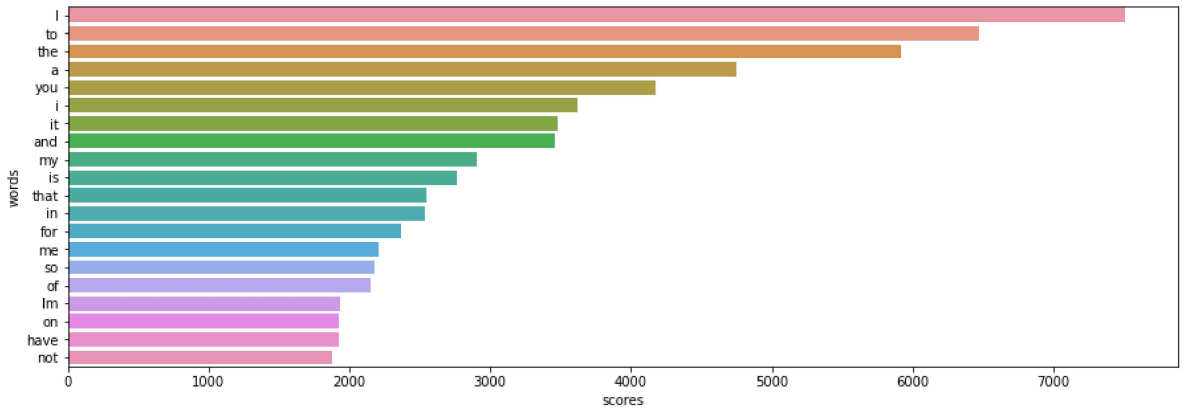


In [34]:

```
most_comm_neg_words=get_tokens(neg_token)
most_comm_neg_words
neg=pd.DataFrame(most_comm_neg_words.items(),columns=["words","scores"])
neg
plt.figure(figsize=(15,5))
sns.barplot(y="words",x="scores",data=neg)
```

Out[34]:

```
<AxesSubplot:xlabel='scores', ylabel='words'>
```

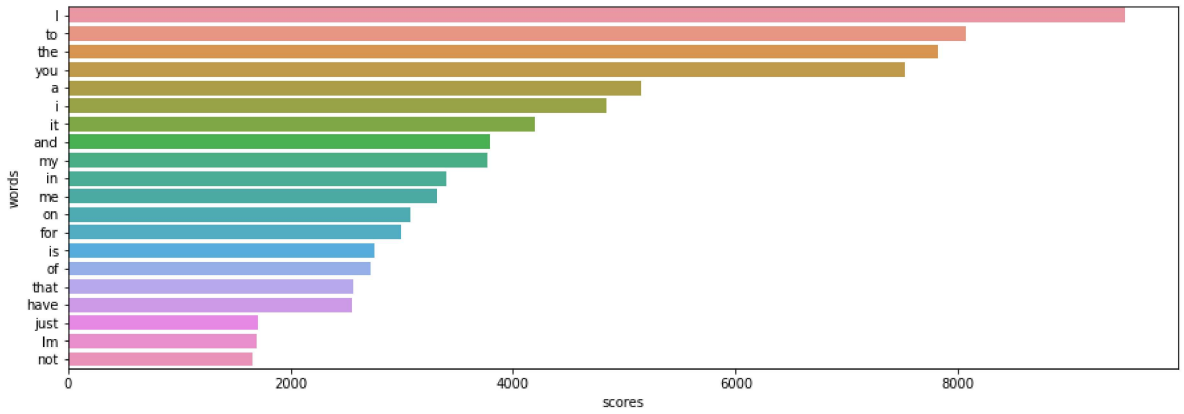


In [35]:

```
most_comm_neutral_words=get_tokens(neutral_token)
most_comm_neutral_words
neutral=pd.DataFrame(most_comm_neutral_words.items(),columns=["words","scores"])
neutral
plt.figure(figsize=(15,5))
sns.barplot(y="words",x="scores",data=neutral)
```

Out[35]:

```
<AxesSubplot:xlabel='scores', ylabel='words'>
```



Logistic Regression Model

In [37]:

```
y=data["Sentiment"]
x=data["clean_Text"]
```

In [40]:

```
y.shape
x.shape
```

Out[40]:

```
(99989,)
```

In [65]:

```
index = data.index
data['random_number'] = np.random.randn(len(index))
train = data[data['random_number'] <= 0.8]
test = data[data['random_number'] > 0.8]
```

In [66]:

```
vectorizer = CountVectorizer(token_pattern=r'\b\w+\b')
train_matrix = vectorizer.fit_transform(train['SentimentText'])
test_matrix = vectorizer.transform(test['SentimentText'])
```

In [67]:

```
X_train = train_matrix
X_test = test_matrix
y_train = train['Sentiment']
y_test = test['Sentiment']
```

In [68]:

```
lr = LogisticRegression()
lr.fit(X_train,y_train)
```

Out[68]:

```
LogisticRegression()
```

In [69]:

```
predictions = lr.predict(X_test)
predictions
```

Out[69]:

```
array([0, 0, 0, ..., 1, 1, 1], dtype=int64)
```

In [70]:

```
new = np.asarray(y_test)
confusion_matrix(predictions,y_test)
```

Out[70]:

```
array([[6464, 2090],
       [2919, 9940]], dtype=int64)
```

In [64]:

```
print(classification_report(predictions,y_test))
```

```
              precision    recall  f1-score   support

           0       0.70      0.76      0.73      9810
           1       0.82      0.78      0.80     14411

    accuracy                           0.77     24221
   macro avg       0.76      0.77      0.76     24221
weighted avg       0.77      0.77      0.77     24221
```

Accuracy of the model is 77%.

In [ ]:

```
```