Tweeter sentiment analysis

Steps to be followed

- read the data
- Text preprocessing
- Convert text to vectors
- apply ml model
- metrics
- prediction

In [54]:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

In [56]:

```
data=pd.read_csv(r"C:\Users\ADMIN\Downloads\train_E6oV3lV.csv")
data
```

Out[56]:

	id	label	tweet
0	1	0	@user when a father is dysfunctional and is s...
1	2	0	@user @user thanks for #lyft credit i can't us...
2	3	0	bihday your majesty
3	4	0	#model i love u take with u all the time in ...
4	5	0	factsguide: society now #motivation
...	...	...	...
31957	31958	0	ate @user isz that youuu?ð□□□ð□□□ð□□□ð□□□ð□□□ð...
31958	31959	0	to see nina turner on the airwaves trying to...
31959	31960	0	listening to sad songs on a monday morning otw...
31960	31961	1	@user #sikh #temple vandalised in in #calgary,...
31961	31962	0	thank you @user for you follow

31962 rows × 3 columns

In [58]:

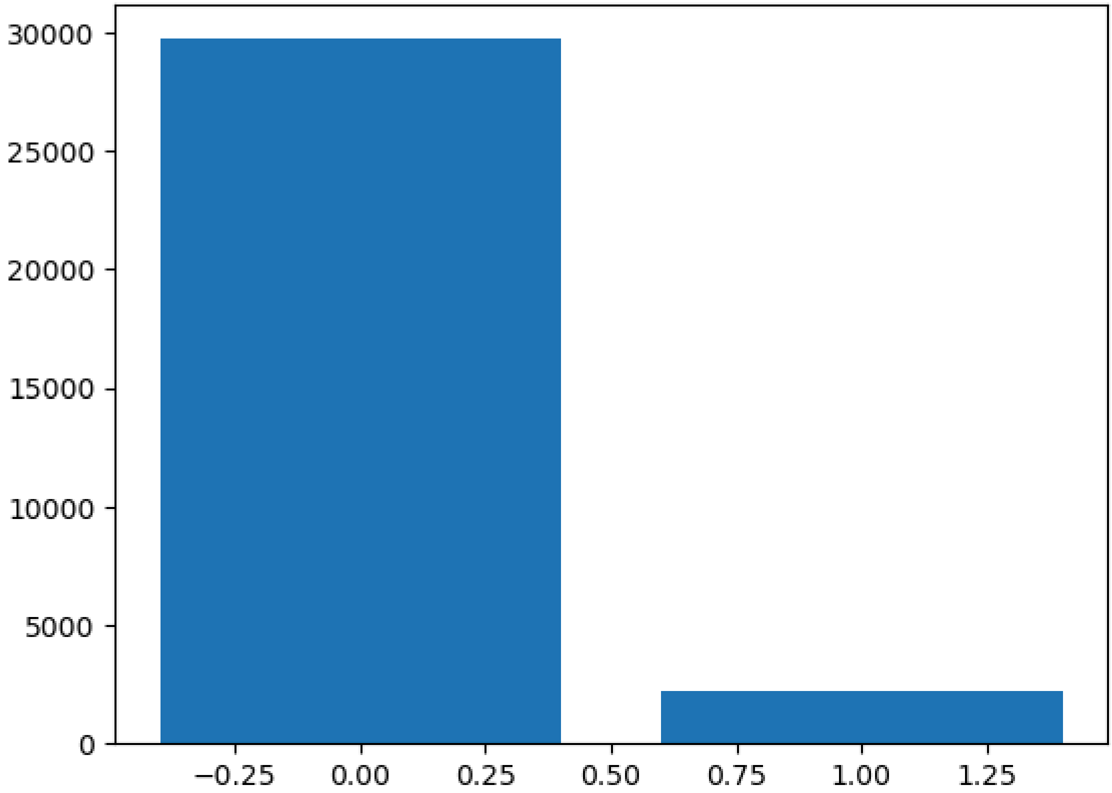
```
data['label'].value_counts()
```

Out[58]:

```
label
0    29720
1     2242
Name: count, dtype: int64
```

In [60]:

```
keys=data['label'].value_counts().keys()
values=data['label'].value_counts().values
plt.bar(keys,values)
plt.show()
```



In [62]:

```
data.isnull().sum()
```

Out[62]:

```
id       0
label    0
tweet    0
dtype: int64
```

In [70]:

```
import pandas as pd
import nltk
import re
import string
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from nltk.stem import PorterStemmer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn import metrics

eng_stop_words=stopwords.words('english')
```

In [72]:

```
ps=PorterStemmer()
corpus=set()
def preprocess(text):
    ## removing unwanted space
    text=text.strip()
    ## removing html tags
    text=re.sub('<[^>]*>','',text)
    ## removing any numerical values
    text=re.sub('[^a-zA-Z]',' ',text)
    ## lower case the words
    text=text.lower()
    ## remove stopwords
    words=text.split()
    words=[w for w in words if w not in eng_stop_words]
    ## stemming the word for sentiment analysis do not remove the stop word
    words=[ps.stem(w) for w in words]
    words=' '.join(words)
    return words
```

In [91]:
```
data['preprocess_review']=data.tweet.apply(preprocess)
data
```

Out[91]:

	id	label	tweet	preprocess_review
0	1	0	@user when a father is dysfunctional and is s...	user father dysfunct selfish drag kid dysfunct...
1	2	0	@user @user thanks for #lyft credit i can't us...	user user thank lyft credit use caus offer whe...
2	3	0	bihday your majesty	bihday majesti
3	4	0	#model i love u take with u all the time in ...	model love u take u time ur
4	5	0	factsguide: society now #motivation	factsguid societi motiv
...	...	...	...	...
31957	31958	0	ate @user isz that youuu? ð□□□ð□□□ð□□□ð□□□ð□□□ð...	ate user isz youuu
31958	31959	0	to see nina turner on the airwaves trying to...	see nina turner airwav tri wrap mantl genuin h...
31959	31960	0	listening to sad songs on a monday morning otw...	listen sad song monday morn otw work sad
31960	31961	1	@user #sikh #temple vandalised in in #calgary,...	user sikh templ vandalis calgari wso condemn act
31961	31962	0	thank you @user for you follow	thank user follow

31962 rows × 4 columns

In [97]:
```
data.drop('id',axis=1,inplace=True)
data
```

Out[97]:

	label	tweet	preprocess_review
0	0	@user when a father is dysfunctional and is s...	user father dysfunct selfish drag kid dysfunct...
1	0	@user @user thanks for #lyft credit i can't us...	user user thank lyft credit use caus offer whe...
2	0	bihday your majesty	bihday majesti
3	0	#model i love u take with u all the time in ...	model love u take u time ur
4	0	factsguide: society now #motivation	factsguid societi motiv
...	...	...	...
31957	0	ate @user isz that youuu? ð□□□ð□□□ð□□□ð□□□ð□□□ð...	ate user isz youuu
31958	0	to see nina turner on the airwaves trying to...	see nina turner airwav tri wrap mantl genuin h...
31959	0	listening to sad songs on a monday morning otw...	listen sad song monday morn otw work sad
31960	1	@user #sikh #temple vandalised in in #calgary,...	user sikh templ vandalis calgari wso condemn act
31961	0	thank you @user for you follow	thank user follow

31962 rows × 3 columns

In [100…

```
data
```

Out[100…

	label	tweet	preprocess_review
0	0	@user when a father is dysfunctional and is s...	user father dysfunct selfish drag kid dysfunct...
1	0	@user @user thanks for #lyft credit i can't us...	user user thank lyft credit use caus offer whe...
2	0	bihday your majesty	bihday majesti
3	0	#model i love u take with u all the time in ...	model love u take u time ur
4	0	factsguide: society now #motivation	factsguid societi motiv
...	...	...	...
31957	0	ate @user isz that youuu? ð□□□ð□□□□ð□□□□ð□□□□ð...	ate user isz youuu
31958	0	to see nina turner on the airwaves trying to...	see nina turner airwav tri wrap mantl genuin h...
31959	0	listening to sad songs on a monday morning otw...	listen sad song monday morn otw work sad
31960	1	@user #sikh #temple vandalised in in #calgary,...	user sikh templ vandalis calgari wso condemn act
31961	0	thank you @user for you follow	thank user follow

31962 rows × 3 columns

- apply train test split

In [107…

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test=train_test_split(data.preprocess_review,
                                                 data.label,
                                                 test_size=0.2,
                                                 random_state=42,
                                                 stratify=data.label)
```

In [113…

```
X_train.shape,y_train.shape
```

Out[113…

```
((25569,), (25569,))
```

In [116…

```
X_test.shape,y_test.shape
```

Out[116…

```
((6393,), (6393,))
```

- create word embedding
- we are using here tf-idf

In [174…

```
tf_idf=TfidfVectorizer()
```

In [176…

```
tf_idf
```

Out[176…

```
▾ TfidfVectorizer
TfidfVectorizer()
```

In [178…

```
X_train_tf=tf_idf.fit_transform(X_train)
X_train_tf
```

Out[178…

```
<25569x27138 sparse matrix of type '<class 'numpy.float64'>'
	with 195606 stored elements in Compressed Sparse Row format>
```

In [180…

```
len(tf_idf.vocabulary_)
```

Out[180…

```
27138
```

In [182…

```
X_train_tf.shape
```

Out[182…

```
(25569, 27138)
```

In [184…

```
X_train_tf[0].toarray()
```

Out[184…

```
array([[0., 0., 0., ..., 0., 0., 0.]])
```

- apply same on test data

In [187…

```
X_test_tf=tf_idf.transform(X_test)
X_test_tf
```

Out[187…

```
<6393x27138 sparse matrix of type '<class 'numpy.float64'>'
	with 44401 stored elements in Compressed Sparse Row format>
```

Model creation

In [190…

```
from sklearn.naive_bayes import MultinomialNB
naive_bays_classifier=MultinomialNB()
naive_bays_classifier.fit(X_train_tf,y_train)
```

Out[190…

```
▾ MultinomialNB
MultinomialNB()
```

In [192…

```
y_pred=naive_bays_classifier.predict(X_test_tf)
y_pred
```

Out[192…

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

In [194…

```
## testing all together
review=['moview is good not a comedy movie']
test_preprocesed=preprocess(review[0])
test_input=tf_idf.transform([test_preprocesed])
test_input

res=naive_bays_classifier.predict(test_input)[0]
res
```

```
if res==1:
    print('Good review')
else:
    print('Bad review')
```

```
Bad review
```

In [196…

```
## testing all together
review=['Movie was mindblowing']
test_preprocesed=preprocess(review[0])
test_input=tf_idf.transform([test_preprocesed])
test_input

res=naive_bays_classifier.predict(test_input)[0]
res

if res==1:
    print('Good review')
else:
    print('Bad review')
```

```
Bad review
```

In [ ]:

In [199…

```
from sklearn.metrics import accuracy_score,f1_score
accuracy_score=accuracy_score(y_test,y_pred)
print(f'the accuracy score is {accuracy_score}')
```

```
the accuracy score is 0.9405599874863132
```

In [ ]:

In [ ]: