In [51]:
```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import accuracy_score, classification_report
```

In [52]:
```
data = pd.read_csv('Tweets.csv')
data
```

Out[52]:

	tweet_id	airline_sentiment	airline_sentiment_confidence	neg
0	570306133677760513	neutral	1.0000	
1	570301130888122368	positive	0.3486	
2	570301083672813571	neutral	0.6837	
3	570301031407624196	negative	1.0000	
4	570300817074462722	negative	1.0000	
...	...	...	...	
14635	569587686496825344	positive	0.3487	
14636	569587371693355008	negative	1.0000	S
14637	569587242672398336	neutral	1.0000	
14638	569587188687634433	negative	1.0000	S
14639	569587140490866689	neutral	0.6771	

14640 rows × 15 columns

In [53]:

```
import nltk
from nltk.corpus import stopwords
```

```
nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\PC-18\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

Out[53]: True

In [85]:
```
def remove_stopwords(text):
    stop_words = set(stopwords.words('english'))
    words = text.split()
    filtered_words = [word for word in words if word.lower() not in stop_
    return ' '.join(filtered_words)
```

In [86]:
```
data['text'] = data['text'].apply(remove_stopwords)
data['text']
```

Out[86]:
```
0                          @VirginAmerica @dhepburn said.
1        @VirginAmerica plus added commercials experien...
2        @VirginAmerica today... Must mean need take an...
3        @VirginAmerica really aggressive blast obnoxio...
4                      @VirginAmerica really big bad thing
                                   ...
14635     @AmericanAir thank got different flight Chicago.
14636    @AmericanAir leaving 20 minutes Late Flight. w...
14637    @AmericanAir Please bring American Airlines #B...
14638    @AmericanAir money, change flight, answer phon...
14639    @AmericanAir 8 ppl need 2 know many seats next...
Name: text, Length: 14640, dtype: object
```

In [56]:
```
data['cleaned_data']=data['text'].str.replace(r'[^\w\s]','',regex=True)
data['cleaned_data']
```

Out[56]:
```
0                             VirginAmerica dhepburn said
1        VirginAmerica plus added commercials experienc...
2        VirginAmerica today Must mean need take anothe...
3        VirginAmerica really aggressive blast obnoxiou...
4                      VirginAmerica really big bad thing
                                   ...
14635       AmericanAir thank got different flight Chicago
14636    AmericanAir leaving 20 minutes Late Flight war...
14637    AmericanAir Please bring American Airlines Bla...
14638    AmericanAir money change flight answer phones ...
14639    AmericanAir 8 ppl need 2 know many seats next ...
Name: cleaned_data, Length: 14640, dtype: object
```

In [57]:
```
print(data.head())
```

```
             tweet_id airline_sentiment  airline_sentiment_confidence  \
0  570306133677760513           neutral                        1.0000
1  570301130888122368          positive                        0.3486
2  570301083672813571           neutral                        0.6837
3  570301031407624196          negative                        1.0000
4  570300817074462722          negative                        1.0000

  negativereason  negativereason_confidence         airline  \
0            NaN                        NaN  Virgin America
1            NaN                     0.0000  Virgin America
2            NaN                        NaN  Virgin America
3     Bad Flight                     0.7033  Virgin America
4     Can't Tell                     1.0000  Virgin America

  airline_sentiment_gold        name negativereason_gold  retweet_count  \
0                    NaN     cairdin                 NaN              0
1                    NaN    jnardino                 NaN              0
2                    NaN  yvonnalynn                 NaN              0
3                    NaN    jnardino                 NaN              0
4                    NaN    jnardino                 NaN              0

                                             text tweet_coord  \
0                    @VirginAmerica @dhepburn said.         NaN
1  @VirginAmerica plus added commercials experien...         NaN
2  @VirginAmerica today... Must mean need take an...         NaN
3  @VirginAmerica really aggressive blast obnoxio...         NaN
4               @VirginAmerica really big bad thing         NaN

              tweet_created tweet_location               user_timezone  \
0  2015-02-24 11:35:52 -0800           NaN  Eastern Time (US & Canada)
1  2015-02-24 11:15:59 -0800           NaN  Pacific Time (US & Canada)
2  2015-02-24 11:15:48 -0800      Lets Play  Central Time (US & Canada)
3  2015-02-24 11:15:36 -0800           NaN  Pacific Time (US & Canada)
4  2015-02-24 11:14:45 -0800           NaN  Pacific Time (US & Canada)

                                       cleaned_data
0                       VirginAmerica dhepburn said
1  VirginAmerica plus added commercials experienc...
2  VirginAmerica today Must mean need take anothe...
3  VirginAmerica really aggressive blast obnoxiou...
4                 VirginAmerica really big bad thing
```

In [58]:
```
X =data['cleaned_data']
X
```

Out[58]:
```
0                              VirginAmerica dhepburn said
1        VirginAmerica plus added commercials experienc...
2        VirginAmerica today Must mean need take anothe...
3        VirginAmerica really aggressive blast obnoxiou...
4                      VirginAmerica really big bad thing
                                   ...
14635       AmericanAir thank got different flight Chicago
14636    AmericanAir leaving 20 minutes Late Flight war...
14637    AmericanAir Please bring American Airlines Bla...
14638    AmericanAir money change flight answer phones ...
14639    AmericanAir 8 ppl need 2 know many seats next ...
Name: cleaned_data, Length: 14640, dtype: object
```

In [59]:
```
y = data['airline_sentiment']
y
```

Out[59]:
```
0         neutral
1        positive
2         neutral
3        negative
4        negative
           ...
14635    positive
14636    negative
14637     neutral
14638    negative
14639     neutral
Name: airline_sentiment, Length: 14640, dtype: object
```

In [60]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
X_train, X_test, y_train, y_test
```

Out[60]:
```
(750              united offering us 8 rooms 32 people FAIL
 6875     JetBlue JFK NYC staff amazing lax JetBlue Send...
 7598     JetBlue well last update right direction least...
 14124    AmericanAir flight 3056 still sitting DFW wait...
 6187     southwestair companion pass broken today purch...
                                ...
 5191     SouthwestAir replacing vitaminwater beer Bravo...
 13418    AmericanAir LAX service reps hand 800 number c...
 5390      SouthwestAir hold hour  chance someone help here
 860      united wouldhow contact discuss poor experienc...
 7270     JetBlue thats ok sure seemed like JetBlue twee...
 Name: cleaned_data, Length: 11712, dtype: object,
 4794     SouthwestAir early frontrunner best airline os...
 10480    USAirways flt EWR Cancelled Flightled yet flts...
 8067     JetBlue going BDL DCA flights yesterday today ...
 8880                          JetBlue depart Washington DC
 8292             JetBlue probably find them ticket s there
                                ...
 11765               USAirways hold 2 hours know keep money
 14156                   AmericanAir hard catering ready go
 10963    USAirways AmericanAir Im finalstretch chairman...
 4877              SouthwestAir Well need something aim for
 5206      SouthwestAir please please please answer phone
 Name: cleaned_data, Length: 2928, dtype: object,
 750      negative
 6875     negative
 7598     positive
 14124    negative
 6187      neutral
           ...
 5191     positive
 13418    negative
 5390     negative
 860      negative
 7270      neutral
 Name: airline_sentiment, Length: 11712, dtype: object,
 4794     positive
 10480    negative
 8067     negative
 8880      neutral
 8292     negative
           ...
 11765    negative
 14156    negative
 10963     neutral
 4877      neutral
 5206     negative
 Name: airline_sentiment, Length: 2928, dtype: object)
```

In [61]:
```
vectorizer = TfidfVectorizer()
vectorizer
```

Out[61]:
```
▾ TfidfVectorizer
TfidfVectorizer()
```

In [62]:
```
X_train_tfidf = vectorizer.fit_transform(X_train)
X_train_tfidf
```

Out[62]:
```
<Compressed Sparse Row sparse matrix of dtype 'float64'
	with 119138 stored elements and shape (11712, 14536)>
```

In [63]:
```
X_test_tfidf = vectorizer.transform(X_test)
X_test_tfidf
```

Out[63]:
```
<Compressed Sparse Row sparse matrix of dtype 'float64'
	with 28012 stored elements and shape (2928, 14536)>
```

In [64]:
```
model = MultinomialNB()
model
```

Out[64]:
```
▾ MultinomialNB
MultinomialNB()
```

In [65]:
```
model.fit(X_train_tfidf, y_train)
```

Out[65]:
```
▾ MultinomialNB
MultinomialNB()
```

In [66]:
```
y_pred = model.predict(X_test_tfidf)
y_pred
```

Out[66]:
```
array(['positive', 'negative', 'negative', ..., 'negative', 'negative',
       'negative'], dtype='<U8')
```

In [67]:
```
accuracy = accuracy_score(y_test, y_pred)
accuracy
```

Out[67]:
```
0.6926229508196722
```

In [68]:
```
report = classification_report(y_test, y_pred)
report
```

Out[68]:
```
'              precision    recall  f1-score   support\n\n    negative
0.68      0.99      0.81      1889\n     neutral       0.73      0.12
0.21       580\n    positive       0.92      0.18      0.30       459\n\
n    accuracy                           0.69      2928\n   macro avg
0.78      0.43      0.44      2928\nweighted avg       0.73      0.69
0.61      2928\n'
```

In [69]:
```
from sklearn.feature_extraction.text import CountVectorizer
```

In [70]:
```
from sklearn.naive_bayes import BernoulliNB,MultinomialNB
```

In [71]:
```
vectorizer = CountVectorizer()
vectorizer
```

Out[71]:
```
▾ CountVectorizer
CountVectorizer()
```

In [72]:
```
vectorizer1 = CountVectorizer(binary = True)
vectorizer2 = CountVectorizer(binary = False)
```

In [73]:
```
X_train_counts = vectorizer.fit_transform(X_train)
X_train_counts
```

Out[73]:
```
<Compressed Sparse Row sparse matrix of dtype 'int64'
        with 119138 stored elements and shape (11712, 14536)>
```

In [74]:
```
X_test_counts = vectorizer.transform(X_test)
X_test_counts
```

Out[74]:
```
<Compressed Sparse Row sparse matrix of dtype 'int64'
        with 28012 stored elements and shape (2928, 14536)>
```

In [75]:
```
model = MultinomialNB()
model
```

Out[75]:
```
▾ MultinomialNB
MultinomialNB()
```

In [76]:
```
model.fit(X_train_counts, y_train)
```

Out[76]:
```
▾ MultinomialNB
MultinomialNB()
```

In [77]:
```
model1= BernoulliNB()
model1
```

Out[77]:
```
▾ BernoulliNB
BernoulliNB()
```

In [78]:
```
model1.fit(X_train_counts, y_train)
```

Out[78]:
```
▾ BernoulliNB
BernoulliNB()
```

In [79]:
```
y_pred = model.predict(X_test_counts)
y_pred
```

Out[79]:
```
array(['positive', 'negative', 'negative', ..., 'negative', 'negative',
       'negative'], dtype='<U8')
```

In [80]:
```
y_pred1 = model1.predict(X_test_counts)
y_pred1
```

Out[80]:
```
array(['positive', 'negative', 'negative', ..., 'negative', 'negative',
       'negative'], dtype='<U8')
```

In [81]:
```
accuracy = accuracy_score(y_test, y_pred)
accuracy
```

Out[81]:
```
0.7762978142076503
```

In [82]:
```
accuracy1 = accuracy_score(y_test, y_pred1)
accuracy1
```

Out[82]:
```
0.7421448087431693
```

In [83]:
```
report = classification_report(y_test, y_pred)
report
```

Out[83]:
```
'              precision    recall  f1-score   support\n\n    negative
0.78      0.97      0.86      1889\n     neutral       0.73      0.34
0.46       580\n    positive       0.80      0.54      0.65       459\n\
n    accuracy                           0.78      2928\n   macro avg
0.77      0.62      0.66      2928\nweighted avg       0.77      0.78
0.75      2928\n'
```

In [84]:
```
report1 = classification_report(y_test, y_pred1)
report1
```

Out[84]:
```
'              precision    recall  f1-score   support\n\n    negative
0.74      0.97      0.84      1889\n     neutral       0.69      0.30
0.42       580\n    positive       0.86      0.35      0.49       459\n\
n    accuracy                           0.74      2928\n   macro avg
0.76      0.54      0.58      2928\nweighted avg       0.75      0.74
0.70      2928\n'
```

In [ ]:
```
#CountVectorizer with Multinomial Naive Bayes gives the best result.
```