AMOL

Amazon Fine Food Reviews Analysis

Data Source: https://www.kaggle.com/snap/amazon-fine-food-reviews

EDA: https://nycdatascience.com/blog/student-works/amazon-fine-foods-visualization/

The Amazon Fine Food Reviews dataset consists of reviews of fine foods from Amazon.

Number of reviews: 568,454
Number of users: 256,059
Number of products: 74,258
Timespan: Oct 1999 - Oct 2012
Number of Attributes/Columns in data: 10

Attribute Information:

1. Id
2. ProductId - unique identifier for the product
3. UserId - unqiue identifier for the user
4. ProfileName
5. HelpfulnessNumerator - number of users who found the review helpful
6. HelpfulnessDenominator - number of users who indicated whether they found the review helpful or not
7. Score - rating between 1 and 5
8. Time - timestamp for the review
9. Summary - brief summary of the review
10. Text - text of the review

Objective:

Given a review, determine whether the review is positive (rating of 4 or 5) or negative (rating of 1 or 2).

[Q] How to determine if a review is positive or negative?

[Ans] We could use Score/Rating. A rating of 4 or 5 can be cosnidered as a positive review. A rating of 1 or 2 can be considered as negative one. A review of rating 3 is considered nuetral and such reviews are ignored from our analysis. This is an approximate and proxy way of determining the polarity (positivity/negativity) of a review.

[1]. Reading Data

[1.1] Loading the data

The dataset is available in two forms

1. .csv file
2. SQLite Database

In order to load the data, We have used the SQLITE dataset as it is easier to query the data and visualise the data efficiently.

Here as we only want to get the global sentiment of the recommendations (positive or negative), we will purposefully ignore all Scores equal to 3. If the score is above 3, then the recommendation wil be set to "positive". Otherwise, it will be set to "negative".

In [1]:

```
%matplotlib inline
import warnings
warnings.filterwarnings("ignore")


import sqlite3
import pandas as pd
```

```python
import numpy as np
import nltk
import string
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.feature_extraction.text import TfidfVectorizer

from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics import confusion_matrix
from sklearn import metrics
from sklearn.metrics import roc_curve, auc
from nltk.stem.porter import PorterStemmer

import re
# Tutorial about Python regular expressions: https://pymotw.com/2/re/
import string
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from nltk.stem.wordnet import WordNetLemmatizer

from gensim.models import Word2Vec
from gensim.models import KeyedVectors
import pickle

from tqdm import tqdm
import os
```

In [2]:

```python
# using SQLite Table to read data.
con = sqlite3.connect('database.sqlite')

# filtering only positive and negative reviews i.e.
# not taking into consideration those reviews with Score=3
# SELECT * FROM Reviews WHERE Score != 3 LIMIT 500000, will give top 500000 data points
# you can change the number to any other number based on your computing power

# filtered_data = pd.read_sql_query(""" SELECT * FROM Reviews WHERE Score != 3 LIMIT 500000""", co
n)
# for tsne assignment you can take 5k data points

filtered_data = pd.read_sql_query(""" SELECT * FROM Reviews WHERE Score != 3""", con)

# Give reviews with Score>3 a positive rating(1), and reviews with a score<3 a negative rating(0).
def partition(x):
    if x < 3:
        return 0
    return 1

#changing reviews with score less than 3 to be positive and vice-versa
actualScore = filtered_data['Score']
positiveNegative = actualScore.map(partition)
filtered_data['Score'] = positiveNegative
print("Number of data points in our data", filtered_data.shape)
filtered_data.head(3)
```

```
Number of data points in our data (525814, 10)
```

Out[2]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	1	130386240(
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	0	134697600(

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1	1	121901760

In [3]:

```
'''
display = pd.read_sql_query("""
SELECT UserId, ProductId, ProfileName, Time, Score, Text, COUNT(*)
FROM Reviews
GROUP BY UserId
HAVING COUNT(*)>1
""", con)
'''
```

Out[3]:

```
'\ndisplay = pd.read_sql_query("""\nSELECT UserId, ProductId, ProfileName, Time, Score, Text,
COUNT(*)\nFROM Reviews\nGROUP BY UserId\nHAVING COUNT(*)>1\n""", con)\n'
```

In [4]:

```
'''
print(display.shape)
display.head()
'''
```

Out[4]:

```
'\nprint(display.shape)\ndisplay.head()\n'
```

In [5]:

```
#display[display['UserId']=='AZY10LLTJ71NX']
```

In [6]:

```
#display['COUNT(*)'].sum()
```

[2] Exploratory Data Analysis

[2.1] Data Cleaning: Deduplication

It is observed (as shown in the table below) that the reviews data had many duplicate entries. Hence it was necessary to remove duplicates in order to get unbiased results for the analysis of the data. Following is an example:

In [7]:

```
'''
display= pd.read_sql_query("""
SELECT *
FROM Reviews
WHERE Score != 3 AND UserId="AR5J8UI46CURR"
ORDER BY ProductID
""", con)
display.head()
'''
```

Out[7]:

```
'\ndisplay= pd.read_sql_query("""\nSELECT *\nFROM Reviews\nWHERE Score != 3 AND 
UserId="AR5J8UI46CURR"\nORDER BY ProductID\n""", con)\ndisplay.head()\n'
```

As it can be seen above that same user has multiple reviews with same values for HelpfulnessNumerator, HelpfulnessDenominator, Score, Time, Summary and Text and on doing analysis it was found that

ProductId=B000HDOPZG was Loacker Quadratini Vanilla Wafer Cookies, 8.82-Ounce Packages (Pack of 8)

ProductId=B000HDL1RQ was Loacker Quadratini Lemon Wafer Cookies, 8.82-Ounce Packages (Pack of 8) and so on

It was inferred after analysis that reviews with same parameters other than ProductId belonged to the same product just having different flavour or quantity. Hence in order to reduce redundancy it was decided to eliminate the rows having same parameters.

The method used for the same was that we first sort the data according to ProductId and then just keep the first similar product review and delelte the others. for eg. in the above just the review for ProductId=B000HDL1RQ remains. This method ensures that there is only one representative for each product and deduplication without sorting would lead to possibility of different representatives still existing for the same product.

In [8]:

```
#Sorting data according to ProductId in ascending order
sorted_data=filtered_data.sort_values('ProductId', axis=0, ascending=True, inplace=False, kind='qui
cksort', na_position='last')
```

In [9]:

```
#Deduplication of entries
final=sorted_data.drop_duplicates(subset={"UserId","ProfileName","Time","Text"}, keep='first', inpl
ace=False)
final.shape
```

Out[9]:

```
(364173, 10)
```

In [10]:

```
#Checking to see how much % of data still remains
(final['Id'].size*1.0)/(filtered_data['Id'].size*1.0)*100
```

Out[10]:

```
69.25890143662969
```

Observation:- It was also seen that in two rows given below the value of HelpfulnessNumerator is greater than HelpfulnessDenominator which is not practically possible hence these two rows too are removed from calcualtions

In [11]:

```
'''
display= pd.read_sql_query("""
SELECT *
FROM Reviews
WHERE Score != 3 AND Id=44737 OR Id=64422
ORDER BY ProductID
""", con)

display.head()
'''
```

Out[11]:

```
'\ndisplay= pd.read_sql_query("""\nSELECT *\nFROM Reviews\nWHERE Score != 3 AND Id=44737 OR 
Id=64422\nORDER BY ProductID\n""", con)\n\ndisplay.head()\n'
```

In [12]:

```
final=final[final.HelpfulnessNumerator<=final.HelpfulnessDenominator]
```

In [13]:

```
#Before starting the next phase of preprocessing lets see the number of entries left
print(final.shape)

#How many positive and negative reviews are present in our dataset?
final['Score'].value_counts()
```

```
(364171, 10)
```

Out[13]:

```
1    307061
0     57110
Name: Score, dtype: int64
```

[3] Preprocessing

[3.1]. Preprocessing Review Text

Now that we have finished deduplication our data requires some preprocessing before we go on further with analysis and making the prediction model.

Hence in the Preprocessing phase we do the following in the order below:-

1. Begin by removing the html tags
2. Remove any punctuations or limited set of special characters like , or . or # etc.
3. Check if the word is made up of english letters and is not alpha-numeric
4. Check to see if the length of the word is greater than 2 (as it was researched that there is no adjective in 2-letters)
5. Convert the word to lowercase
6. Remove Stopwords
7. Finally Snowball Stemming the word (it was obsereved to be better than Porter Stemming)

After which we collect the words used to describe positive and negative reviews

In [14]:

```
# printing some random reviews
'''
sent_0 = final['Text'].values[0]
print(sent_0)
print("="*50)

sent_1000 = final['Text'].values[1000]
print(sent_1000)
print("="*50)

sent_1500 = final['Text'].values[1500]
print(sent_1500)
print("="*50)

sent_4900 = final['Text'].values[4900]
print(sent_4900)
print("="*50)
'''
```

Out[14]:

```
'\nsent_0 = final[\'Text\'].values[0]\nprint(sent_0)\nprint("="*50)\n\nsent_1000 =
final[\'Text\'].values[1000]\nprint(sent_1000)\nprint("="*50)\n\nsent_1500 =
final[\'Text\'].values[1500]\nprint(sent_1500)\nprint("="*50)\n\nsent_4900 =
final[\'Text\'].values[4900]\nprint(sent_4900)\nprint("="*50)\n'
```

In [15]:

```
'''
# remove urls from text python: https://stackoverflow.com/a/40823105/4084039
sent_0 = re.sub(r"http\S+", "", sent_0)
```

```
sent_1000 = re.sub(r"http\S+", "", sent_1000)
sent_150 = re.sub(r"http\S+", "", sent_1500)
sent_4900 = re.sub(r"http\S+", "", sent_4900)

print(sent_0)
'''
```

Out[15]:

```
'\n# remove urls from text python: https://stackoverflow.com/a/40823105/4084039\nsent_0 =
re.sub(r"http\\S+", "", sent_0)\nsent_1000 = re.sub(r"http\\S+", "", sent_1000)\nsent_150 =
re.sub(r"http\\S+", "", sent_1500)\nsent_4900 = re.sub(r"http\\S+", "",
sent_4900)\n\nprint(sent_0)\n'
```

In [16]:

```
# https://stackoverflow.com/questions/16206380/python-beautifulsoup-how-to-remove-all-tags-from-an
-element
'''
from bs4 import BeautifulSoup

soup = BeautifulSoup(sent_0, 'lxml')
text = soup.get_text()
print(text)
print("="*50)

soup = BeautifulSoup(sent_1000, 'lxml')
text = soup.get_text()
print(text)
print("="*50)

soup = BeautifulSoup(sent_1500, 'lxml')
text = soup.get_text()
print(text)
print("="*50)

soup = BeautifulSoup(sent_4900, 'lxml')
text = soup.get_text()
print(text)
'''
```

Out[16]:

```
'\nfrom bs4 import BeautifulSoup\n\nsoup = BeautifulSoup(sent_0, \'lxml\')\ntext =
soup.get_text()\nprint(text)\nprint("="*50)\n\nsoup = BeautifulSoup(sent_1000, \'lxml\')\ntext = s
oup.get_text()\nprint(text)\nprint("="*50)\n\nsoup = BeautifulSoup(sent_1500, \'lxml\')\ntext = so
up.get_text()\nprint(text)\nprint("="*50)\n\nsoup = BeautifulSoup(sent_4900, \'lxml\')\ntext = sou
p.get_text()\nprint(text)\n'
```

In [17]:

```python
# https://stackoverflow.com/a/47091490/4084039
import re
from bs4 import BeautifulSoup
def decontracted(phrase):
    # specific
    phrase = re.sub(r"won't", "will not", phrase)
    phrase = re.sub(r"can\'t", "can not", phrase)

    # general
    phrase = re.sub(r"n\'t", " not", phrase)
    phrase = re.sub(r"\'re", " are", phrase)
    phrase = re.sub(r"\'s", " is", phrase)
    phrase = re.sub(r"\'d", " would", phrase)
    phrase = re.sub(r"\'ll", " will", phrase)
    phrase = re.sub(r"\'t", " not", phrase)
    phrase = re.sub(r"\'ve", " have", phrase)
    phrase = re.sub(r"\'m", " am", phrase)
    return phrase
```

In [18]:

```
'''sent_1500 = decontracted(sent_1500)
print(sent_1500)
```

```
print(sent_1500)
print("="*50)
'''
```

Out[18]:

```
'sent_1500 = decontracted(sent_1500)\nprint(sent_1500)\nprint("="*50)\n'
```

In [19]:

```
#remove words with numbers python: https://stackoverflow.com/a/18082370/4084039
#sent_0 = re.sub("\S*\d\S*", "", sent_0).strip()
#print(sent_0)
```

In [20]:

```
#remove spacial character: https://stackoverflow.com/a/5843547/4084039
#sent_1500 = re.sub('[^A-Za-z0-9]+', ' ', sent_1500)
#print(sent_1500)
```

In [21]:

```
# https://gist.github.com/sebleier/554280
# we are removing the words from the stop words list: 'no', 'nor', 'not'
# <br /><br /> ==> after the above steps, we are getting "br br"
# we are including them into stop words list
# instead of <br /> if we have <br/> these tags would have revmoved in the 1st step

stopwords= set(['br', 'the', 'i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "y
ou're", "you've",\
            "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his',
'himself', \
            'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them',
'their',\
            'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll",
'these', 'those', \
            'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having',
'do', 'does', \
            'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', '
while', 'of', \
            'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during',
'before', 'after',\
            'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under'
, 'again', 'further',\
            'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'e
ach', 'few', 'more',\
            'most', 'other', 'some', 'such', 'only', 'own', 'same', 'so', 'than', 'too', 'very', \
            's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll'
, 'm', 'o', 're', \
            've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "do
esn't", 'hadn',\
            "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn',
"mightn't", 'mustn',\
            "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn',
"wasn't", 'weren', "weren't", \
            'won', "won't", 'wouldn', "wouldn't"])
```

In [22]:

```
SORT_DATA = final.sort_values("Time")
```

In [23]:

```
# Combining all the above stundents
from tqdm import tqdm
preprocessed_reviews = []
# tqdm is for printing the status bar
for sentance in tqdm(SORT_DATA['Text'].values):
    sentance = re.sub(r"http\S+", "", sentance)
    sentance = BeautifulSoup(sentance, 'lxml').get_text()
    sentance = decontracted(sentance)
    sentance = re.sub("\S*\d\S*", "", sentance).strip()
    sentance = re.sub('[^A-Za-z]+', ' ', sentance)
```

```
    sentance = re.sub('[^A-Za-z]+', ' ', sentance)
    # https://gist.github.com/sebleier/554280
    sentance = ' '.join(e.lower() for e in sentance.split() if e.lower() not in stopwords)
    preprocessed_reviews.append(sentance.strip())
```

```
100%|██████████| 364171/364171 [04:53<00:00, 1240.26it/s]
```

In [24]:

```
SORT_DATA['Score'].value_counts()
```

Out[24]:

```
1    307061
0     57110
Name: Score, dtype: int64
```

In [25]:

```
DATA = np.array(preprocessed_reviews[0:50000])
LABEL  = np.array(SORT_DATA['Score'][0:50000])
```

In [26]:

```
#DATA_50K = np.array(preprocessed_reviews[0:49000])
#LABEL = final['Score']
#LABEL_50K = np.array(LABEL[0:49000])


from sklearn.model_selection import train_test_split
X_train_temp, X_TEST, Y_train_temp, Y_TEST = train_test_split(DATA, LABEL, test_size=0.33,stratify=
LABEL)
X_TRAIN, X_CV, Y_TRAIN, Y_CV = train_test_split(X_train_temp, Y_train_temp,
test_size=0.33,stratify=Y_train_temp)
```

[5] Assignment 3: KNN

1. **Apply Knn(brute force version) on these feature sets**
   - SET 1:Review text, preprocessed one converted into vectors using (BOW)
   - SET 2:Review text, preprocessed one converted into vectors using (TFIDF)
   - SET 3:Review text, preprocessed one converted into vectors using (AVG W2v)
   - SET 4:Review text, preprocessed one converted into vectors using (TFIDF W2v)
2. **Apply Knn(kd tree version) on these feature sets**
   NOTE: sklearn implementation of kd-tree accepts only dense matrices, you need to convert the sparse matrices of CountVectorizer/TfidfVectorizer into dense matices. You can convert sparse matrices to dense using .toarray() attribute. For more information please visit this link
   - SET 5:Review text, preprocessed one converted into vectors using (BOW) but with restriction on maximum features generated.

     ```
     count_vect = CountVectorizer(min_df=10, max_features=500)
     count_vect.fit(preprocessed_reviews)
     ```

   - SET 6:Review text, preprocessed one converted into vectors using (TFIDF) but with restriction on maximum features generated.

     ```
     tf_idf_vect = TfidfVectorizer(min_df=10, max_features=500)
     tf_idf_vect.fit(preprocessed_reviews)
     ```

   - SET 3:Review text, preprocessed one converted into vectors using (AVG W2v)
   - SET 4:Review text, preprocessed one converted into vectors using (TFIDF W2v)

3. **The hyper paramter tuning(find best K)**

    - Find the best hyper parameter which will give the maximum AUC value
    - Find the best hyper paramter using k-fold cross validation or simple cross validation data
    - Use gridsearch cv or randomsearch cv or you can also write your own for loops to do this task of hyperparameter tuning

4. **Representation of results**

    - You need to plot the performance of model both on train data and cross validation data for each hyper parameter, like shown in the figure
    - Once after you found the best hyper parameter, you need to train your model with it, and find the AUC on test data and plot the ROC curve on both train and test.
    - Along with plotting ROC curve, you need to print the confusion matrix with predicted and original labels of test data points

5. **Conclusion**

    - You need to summarize the results at the end of the notebook, summarize it in the table format. To print out a table please refer to this prettytable library link

Note: Data Leakage

1. There will be an issue of data-leakage if you vectorize the entire data and then split it into train/cv/test.
2. To avoid the issue of data-leakag, make sure to split your data first and then vectorize it.
3. While vectorizing your data, apply the method fit_transform() on you train data, and apply the method transform() on cv/test data.
4. For more details please go through this link.

In [27]:

```
def BRUTE_FORCE(TRAIN_DATA,TRAIN_LABEL,CV_DATA,CV_LABEL,TEST_DATA,TEST_LABEL):
    from sklearn import neighbors
    from sklearn.metrics import roc_auc_score
    import matplotlib.pyplot as plt

    K = [1,5,9,15,31,51]

    from sklearn import  neighbors
    TRAIN_AUC = []
    CV_AUC = []

    for i in K:
        OBJ =neighbors.KNeighborsClassifier(n_neighbors=i,algorithm='brute')
        OBJ.fit(TRAIN_DATA,TRAIN_LABEL)
        Y_TRAIN_PRED =  list(OBJ.predict_proba(TRAIN_DATA)[:,1])
        TRAIN_AUC.append(roc_auc_score(TRAIN_LABEL,Y_TRAIN_PRED))

        Y_CV_PRED =  list(OBJ.predict_proba(CV_DATA)[:,1])
        CV_AUC.append(roc_auc_score(CV_LABEL,Y_CV_PRED))

    plt.plot(K,TRAIN_AUC, label='Train AUC')
    plt.plot(K,CV_AUC, label='CV AUC')
    plt.legend()
    plt.xlabel("K: hyperparameter")
    plt.ylabel("AUC")
    plt.title("AUC PLOTS")
    plt.show()

    BEST_K = K[CV_AUC.index(max(CV_AUC))]
    print("BEST_K is {}".format(BEST_K))
    knn_optimal = neighbors.KNeighborsClassifier(n_neighbors=BEST_K,algorithm='brute')
    knn_optimal.fit(TRAIN_DATA,TRAIN_LABEL)

    PRED_TEST=list(knn_optimal.predict(TEST_DATA))

    PRED_TEST = np.array(PRED_TEST)

    PRED_TRAIN = []
```

```
    PRED_TRAIN=list(knn_optimal.predict(TRAIN_DATA))

    PRED_TRAIN = np.array(PRED_TRAIN)



    neigh = neighbors.KNeighborsClassifier(n_neighbors=BEST_K,algorithm='brute')
    neigh.fit(TRAIN_DATA,TRAIN_LABEL)
    TRAIN_PROBA=  list(neigh.predict_proba(TRAIN_DATA)[:,1])
    TEST_PROBA =  list(neigh.predict_proba(TEST_DATA)[:,1])


    from sklearn import metrics
    fpr_2,tpr_2,tr_2 = metrics.roc_curve(TEST_LABEL,TEST_PROBA)
    fpr_1,tpr_1,tr_1 = metrics.roc_curve(TRAIN_LABEL,TRAIN_PROBA)
    lw=2
    area_train = metrics.auc(fpr_1, tpr_1)
    area_test = metrics.auc(fpr_2, tpr_2)
    plt.plot(fpr_2, tpr_2, color='darkorange',lw=lw, label='ROC curve of Test data (area = %0.2f)'
% area_test)
    plt.plot(fpr_1, tpr_1, color='green',lw=lw, label='ROC curve of Train data(area = %0.2f)' % are
a_train)
    plt.legend()
    plt.title("ROC CURVE")


    from sklearn.metrics import confusion_matrix
    import seaborn as sns

    plt.figure()
    cm = confusion_matrix(TEST_LABEL,PRED_TEST)
    class_label = ["negative", "positive"]
    df_cm_test = pd.DataFrame(cm, index = class_label, columns = class_label)
    sns.heatmap(df_cm_test , annot = True, fmt = "d")
    plt.title("Confusiion Matrix for test data")
    plt.xlabel("Predicted Label")
    plt.ylabel("True Label")
    plt.show()
```

........................BOW ....

In [28]:

```
print(X_TRAIN.shape,Y_TRAIN.shape)
print(X_TEST.shape,Y_TEST.shape)
print(X_CV.shape,Y_CV.shape)
```

```
(22445,) (22445,)
(16500,) (16500,)
(11055,) (11055,)
```

In [29]:

```
from sklearn.feature_extraction.text import CountVectorizer
vectorizer = CountVectorizer()
vectorizer.fit(X_TRAIN)
```

Out[29]:

```
CountVectorizer(analyzer='word', binary=False, decode_error='strict',
                dtype=<class 'numpy.int64'>, encoding='utf-8', input='content',
                lowercase=True, max_df=1.0, max_features=None, min_df=1,
                ngram_range=(1, 1), preprocessor=None, stop_words=None,
                strip_accents=None, token_pattern='(?u)\\b\\w\\w+\\b',
                tokenizer=None, vocabulary=None)
```

In [30]:

In [30]:

```
X_TRAIN_BOW = vectorizer.transform(X_TRAIN)
X_CV_BOW = vectorizer.transform(X_CV)
X_TEST_BOW = vectorizer.transform(X_TEST)
```

In [31]:

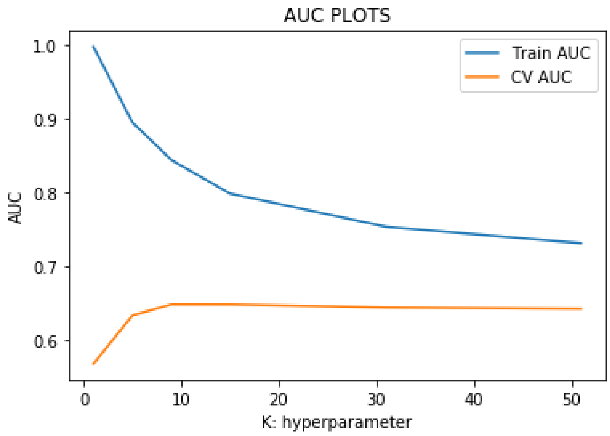
```
print("After vectorizations")
print(X_TRAIN_BOW.shape, Y_TRAIN.shape)
print(X_CV_BOW.shape,Y_CV.shape)
print(X_TEST_BOW.shape, Y_TEST.shape)
print("="*100)
```

```
After vectorizations
(22445, 29101) (22445,)
(11055, 29101) (11055,)
(16500, 29101) (16500,)
====================================================================================================
```

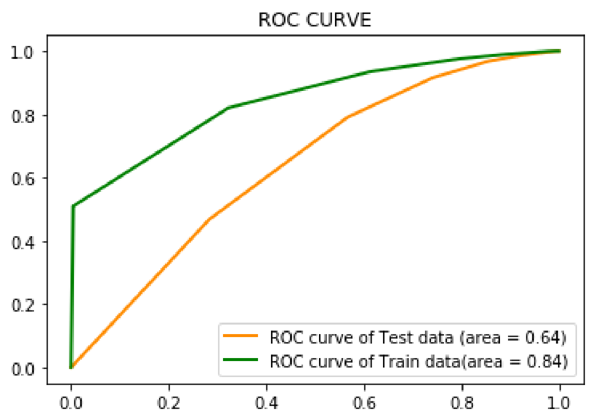
[5.1.1] Applying KNN brute force on BOW, SET 1

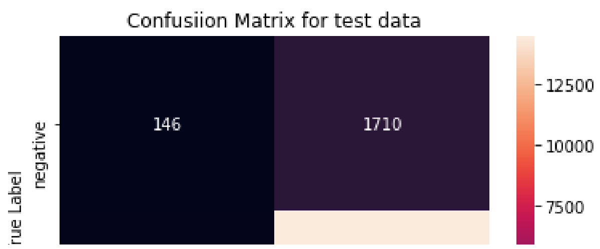
In [32]:

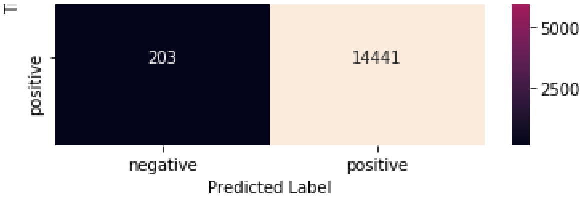
```
BRUTE_FORCE(X_TRAIN_BOW,Y_TRAIN,X_CV_BOW,Y_CV,X_TEST_BOW,Y_TEST)
```




```
BEST_K is 9
```







........................TFIDF.......................

[5.1.4] Applying KNN brute force on TFIDF , SET 2

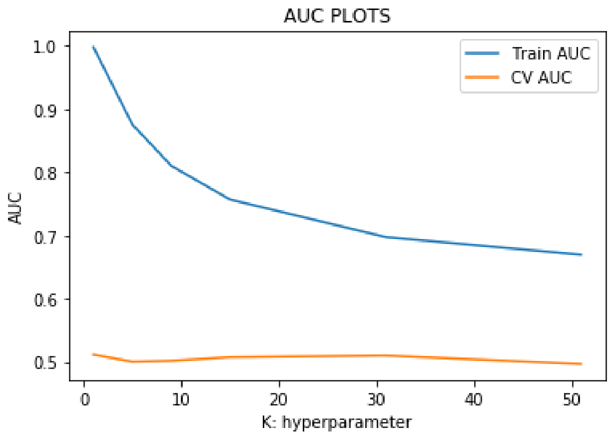
In [33]:

```
from sklearn.feature_extraction.text import TfidfVectorizer
VECTORIZER_TF_IDF = TfidfVectorizer(ngram_range=(1,2), min_df=10)
VECTORIZER_TF_IDF.fit(X_TRAIN)


X_TRAIN_TFIDF = VECTORIZER_TF_IDF.transform(X_TRAIN)
X_CV_TFIDF = VECTORIZER_TF_IDF.transform(X_CV)
X_TEST_TFIDF = VECTORIZER_TF_IDF.transform(X_TEST)



print("After vectorizations")
print(X_TRAIN_TFIDF.shape, Y_TRAIN.shape)
print(X_CV_TFIDF.shape,Y_CV.shape)
print(X_TEST_TFIDF.shape, Y_TEST.shape)
print("="*100)
```

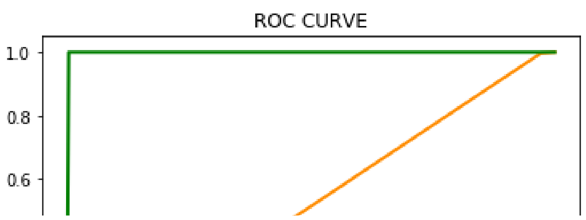
```
After vectorizations
(22445, 12749) (22445,)
(11055, 12749) (11055,)
(16500, 12749) (16500,)
====================================================================================================
```

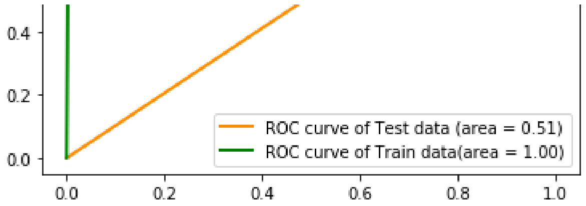
In [34]:

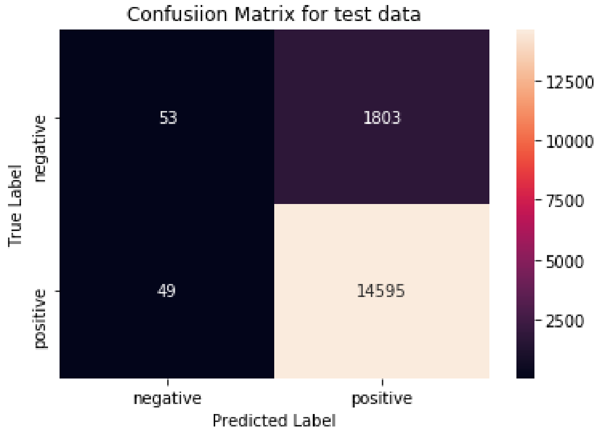
```
BRUTE_FORCE(X_TRAIN_TFIDF,Y_TRAIN,X_CV_TFIDF,Y_CV,X_TEST_TFIDF,Y_TEST)
```




```
BEST_K is 1
```







............AVG W2V........

[5.1.4] Applying KNN brute force on AVG W2V, SET 3

In [35]:

```
# Train your own Word2Vec model using your own text corpus
i=0
list_of_sentance=[]
for sentance in X_TRAIN:
    list_of_sentance.append(sentance.split())

 # min_count = 5 considers only words that occured atleast 5 times
w2v_model=Word2Vec(list_of_sentance,min_count=5,size=50, workers=4)

w2v_words = list(w2v_model.wv.vocab)
print("number of words that occured minimum 5 times ",len(w2v_words))
```

```
C:\Users\KIRTIMAN\Anaconda3\lib\site-packages\gensim\models\base_any2vec.py:743: UserWarning: C ex
tension not loaded, training will be slow. Install a C compiler and reinstall gensim for fast trai
ning.
  "C extension not loaded, training will be slow. "
```

```
number of words that occured minimum 5 times  9176
```

In [36]:

```
def AVGW2V(X_test):

    i=0
    list_of_sentance=[]
    for sentance in X_test:
        list_of_sentance.append(sentance.split())
    test_vectors = []; # the avg-w2v for each sentence/review is stored in this list
    for sent in tqdm(list_of_sentance): # for each review/sentence
        sent_vec = np.zeros(50) # as word vectors are of zero length 50, you might need to change t
his to 300 if you use google's w2v
        cnt_words =0; # num of words with a valid vector in the sentence/review
        for word in sent: # for each word in a review/sentence
            if word in w2v_words:
                vec = w2v_model.wv[word]
                sent_vec += vec
                cnt_words += 1
        if cnt_words != 0:
```

```
            sent_vec /= cnt_words
        test_vectors.append(sent_vec)
    return test_vectors
```

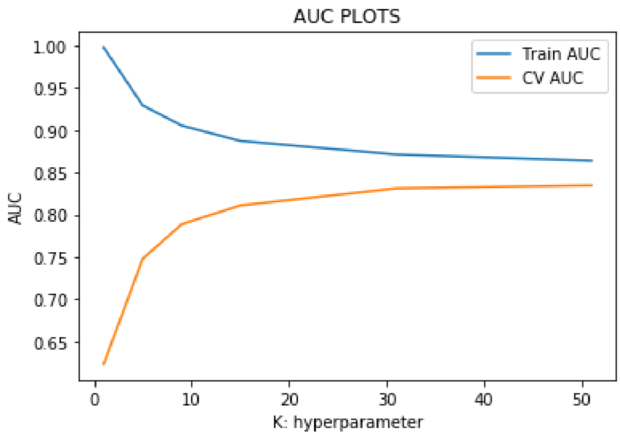
In [37]:

```
AV_TRAIN_BOW = AVGW2V(X_TRAIN)
AV_CV_BOW = AVGW2V(X_CV)
AV_TEST_BOW = AVGW2V(X_TEST)
```

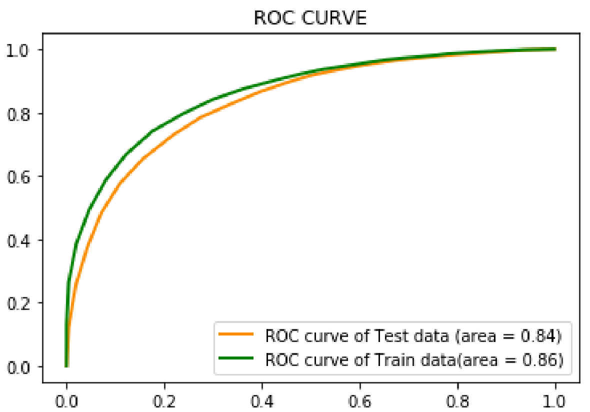
```
100%|██████████| 22445/22445 [00:54<00:00, 408.86it/s]
100%|██████████| 11055/11055 [00:30<00:00, 364.04it/s]
100%|██████████| 16500/16500 [00:45<00:00, 360.83it/s]
```

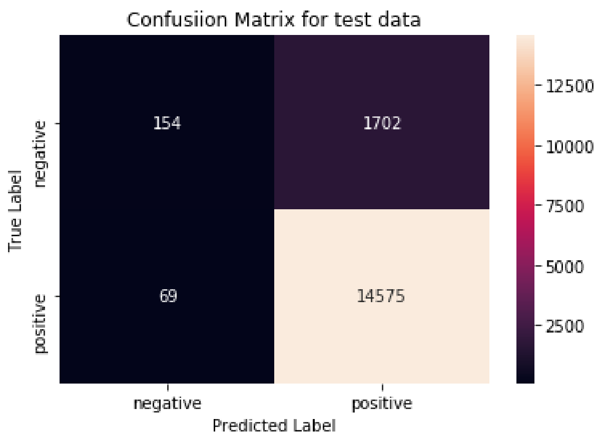
In [38]:

```
BRUTE_FORCE(np.array(AV_TRAIN_BOW),Y_TRAIN,np.array(AV_CV_BOW),Y_CV,np.array(AV_TEST_BOW),Y_TEST)
```




```
BEST_K is 51
```





..............TFIDF W2V..............

[5.1.4] Applying KNN brute force on TFIDF W2V, SET 4

In [39]:

```
model = TfidfVectorizer()
model.fit(X_TRAIN)

dictionary = dict(zip(model.get_feature_names(), list(model.idf_)))

# TF-IDF weighted Word2Vec
tfidf_feat = model.get_feature_names() # tfidf words/col-names
# final_tf_idf is the sparse matrix with row= sentence, col=word and cell_val = tfidf
```

In [40]:

```
def TFIDFW2V(test):
    '''
    Returns tfidf word2vec
    '''
    tfidf_sent_vectors = []; # the tfidf-w2v for each sentence/review is stored in this list
    i=0
    list_of_sentance=[]
    for sentance in test:
        list_of_sentance.append(sentance.split())

    for sent in tqdm(list_of_sentance): # for each review/sentence
        sent_vec = np.zeros(50) # as word vectors are of zero length
        weight_sum =0; # num of words with a valid vector in the sentence/review
        for word in sent: # for each word in a review/sentence
            if word in w2v_words and word in tfidf_feat:
                vec = w2v_model.wv[word]
                tf_idf = dictionary[word]*(sent.count(word)/len(sent))
                sent_vec += (vec * tf_idf)
                weight_sum += tf_idf
        if weight_sum != 0:
            sent_vec /= weight_sum
        tfidf_sent_vectors.append(sent_vec)

    return tfidf_sent_vectors
```

In [41]:

```
AV_TRAIN_TFIDF = TFIDFW2V(X_TRAIN)
AV_CV_TFIDF = TFIDFW2V(X_CV)
AV_TEST_TFIDF = TFIDFW2V(X_TEST)
```

```
100%|██████████| 22445/22445 [08:50<00:00, 42.31it/s]
100%|██████████| 11055/11055 [04:02<00:00, 45.63it/s]
100%|██████████| 16500/16500 [06:22<00:00, 43.15it/s]
```

In [42]:

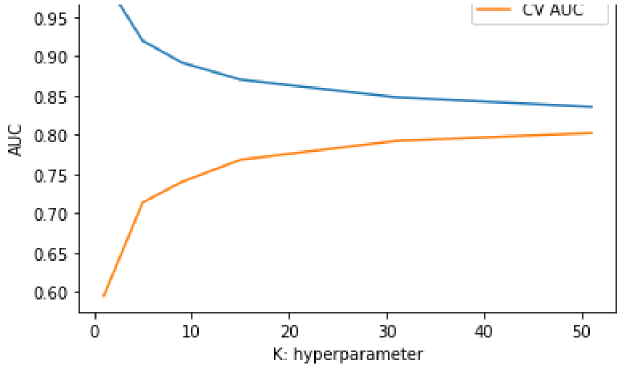
```
len(AV_TRAIN_TFIDF)
```

Out[42]:

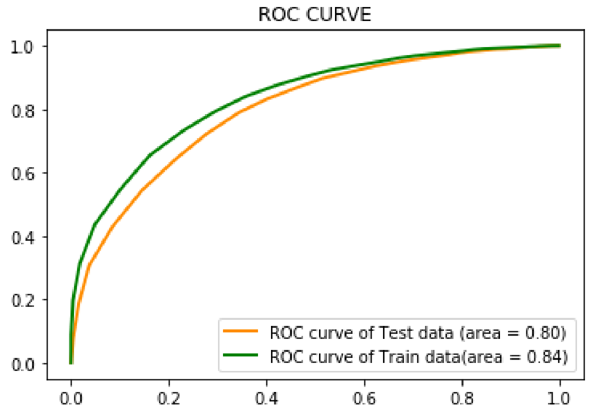
```
22445
```

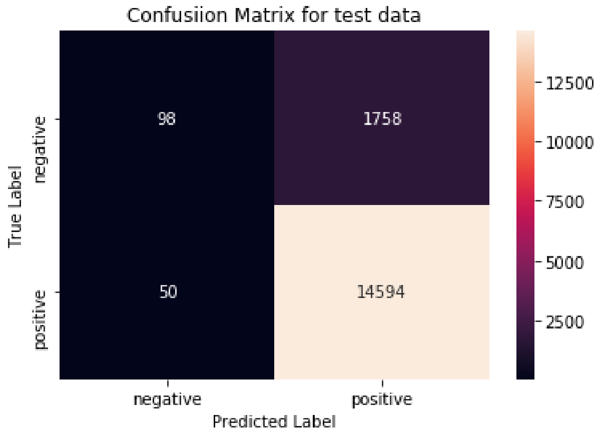
In [43]:

```
BRUTE_FORCE(np.array(AV_TRAIN_TFIDF),Y_TRAIN,np.array(AV_CV_TFIDF),Y_CV,np.array(AV_TEST_TFIDF),Y_TEST)
```




```
BEST_K is 51
```





[5.2] Applying KNN kd-tree,

[5.2.1] Applying KNN kd-tree on BOW, SET 5

In [84]:

```
def KD_TREE(TRAIN_DATA,TRAIN_LABEL,CV_DATA,CV_LABEL,TEST_DATA,TEST_LABEL):
    from sklearn import  neighbors
    from sklearn.metrics import roc_auc_score
    import matplotlib.pyplot as plt
    AUC_TRAIN = []
    AUC_CV = []
    K = [1,5,7,9,11,15,17,19,21,23,31,35,39,41,47,49,51]

    for i in K :
        OBJ_KD= neighbors.KNeighborsClassifier(n_neighbors=i, algorithm='kd_tree')
        OBJ_KD.fit(TRAIN_DATA,TRAIN_LABEL)
        PROB_CV_KD = OBJ_KD.predict_proba(CV_DATA)
        PROB_TRAIN_KD = OBJ_KD.predict_proba(TRAIN_DATA)
        AUC_CV.append(roc_auc_score(CV_LABEL, PROB_CV_KD[:,1]))
        AUC_TRAIN.append(roc_auc_score(TRAIN_LABEL,PROB_TRAIN_KD[:,1]))
```

```
    plt.figure()
    plt.plot(K,AUC_TRAIN, label='Train AUC')
    plt.plot(K,AUC_CV, label='CV AUC')
    plt.legend()
    plt.xlabel("K: hyperparameter")
    plt.ylabel("AUC")
    plt.title("AUC PLOT")
    plt.show()


    BEST_K = K[AUC_CV.index(max(AUC_CV))]
    print("BEST K is {}".format(BEST_K))
    OBJ_KD= neighbors.KNeighborsClassifier(n_neighbors=BEST_K, algorithm='kd_tree')
    OBJ_KD.fit(TRAIN_DATA,TRAIN_LABEL)
    PROB_TEST_KD = OBJ_KD.predict_proba(TEST_DATA)
    PROB_TRAIN_KD = OBJ_KD.predict_proba(TRAIN_DATA)

    from sklearn import metrics
    fpr_2,tpr_2,tr_2 = metrics.roc_curve(TEST_LABEL,PROB_TEST_KD[:,1])
    fpr_1,tpr_1,tr_1 = metrics.roc_curve(TRAIN_LABEL,PROB_TRAIN_KD[:,1])
    lw=2
    area_train = metrics.auc(fpr_1, tpr_1)
    area_test = metrics.auc(fpr_2, tpr_2)
    plt.plot(fpr_2, tpr_2, color='darkorange',lw=lw, label='ROC curve of Test data (area = %0.2f)'
% area_test)
    plt.plot(fpr_1, tpr_1, color='green',lw=lw, label='ROC curve of Train data(area = %0.2f)' % are
a_train)
    plt.legend()
    plt.title("ROC CURVE")

    PRED_LABEL= OBJ_KD.predict(TEST_DATA)
    from sklearn.metrics import confusion_matrix
    import seaborn as sns

    plt.figure()
    cm = confusion_matrix(TEST_LABEL,PRED_LABEL)
    class_label = ["negative", "positive"]
    df_cm_test = pd.DataFrame(cm, index = class_label, columns = class_label)
    sns.heatmap(df_cm_test , annot = True, fmt = "d")
    plt.title("Confusiion Matrix for test data")
    plt.xlabel("Predicted Label")
    plt.ylabel("True Label")
    plt.show()
```

In [49]:

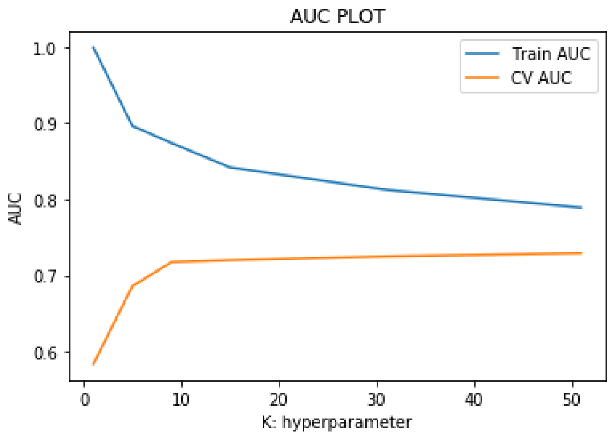
```
DATA_20K = np.array(preprocessed_reviews[0:20000])
LABEL_20K  = np.array(SORT_DATA['Score'][0:20000])





from sklearn.model_selection import train_test_split
X_train_temp_KD, X_TEST_KD, Y_train_temp_KD, Y_TEST_KD = train_test_split(DATA_20K, LABEL_20K,
test_size=0.33)
X_TRAIN_KD, X_CV_KD, Y_TRAIN_KD, Y_CV_KD = train_test_split(X_train_temp_KD, Y_train_temp_KD, test_
size=0.33)
```

In [50]:

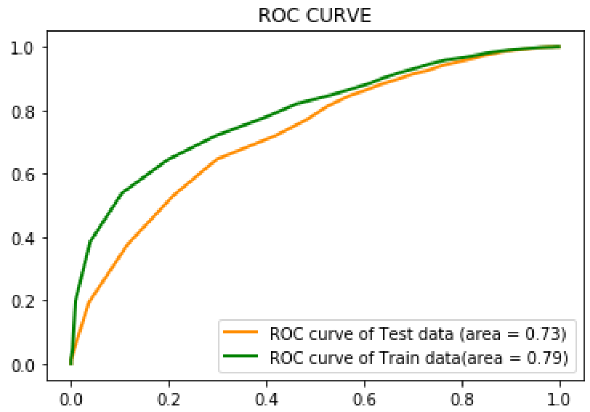
```
# Please write all the code with proper documentation
count_vect = CountVectorizer(min_df=10, max_features=500)
count_vect.fit(X_TRAIN_KD)
XTB = count_vect.transform(X_TRAIN_KD)
XCV = count_vect.transform(X_CV_KD)
XTEST = count_vect.transform(X_TEST_KD)
XTB = XTB.toarray()
XCV = XCV.toarray()
XTEST = XTEST.toarray()
```

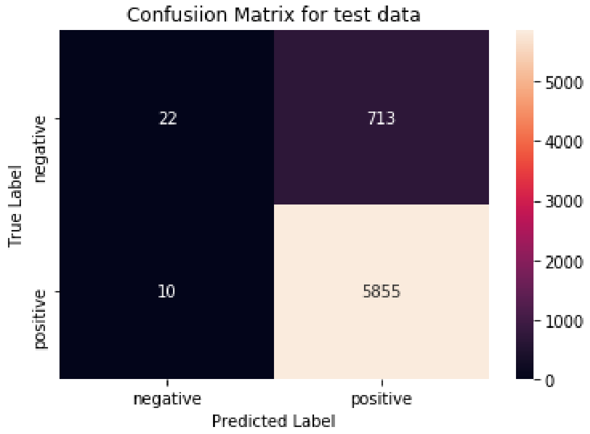
In [53]:

```
KD_TREE(XTB,Y_TRAIN_KD,XCV,Y_CV_KD,XTEST,Y_TEST_KD)
```




```
BEST K is 51
```





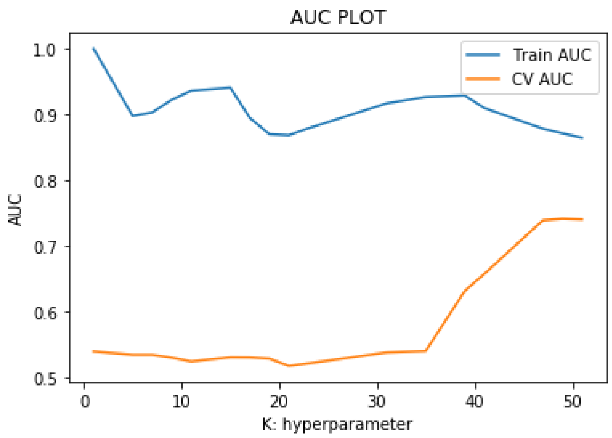
[5.2.3] Applying KNN kd-tree on TFIDF, SET 6

In [54]:

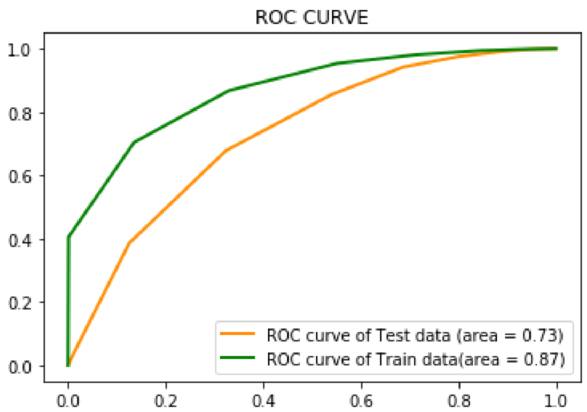
```
TFIDF_VECT = TfidfVectorizer(ngram_range=(1,2),min_df=10, max_features=500)
TFIDF_VECT.fit(X_TRAIN_KD)
XTB_TFIDF = TFIDF_VECT.transform(X_TRAIN_KD)
XCV_TFIDF = TFIDF_VECT.transform(X_CV_KD)
XTEST_TFIDF = TFIDF_VECT.transform(X_TEST_KD)
XTB_TFIDF = XTB_TFIDF.toarray()
XCV_TFIDF = XCV_TFIDF.toarray()
XTEST_TFIDF = XTEST_TFIDF.toarray()
```

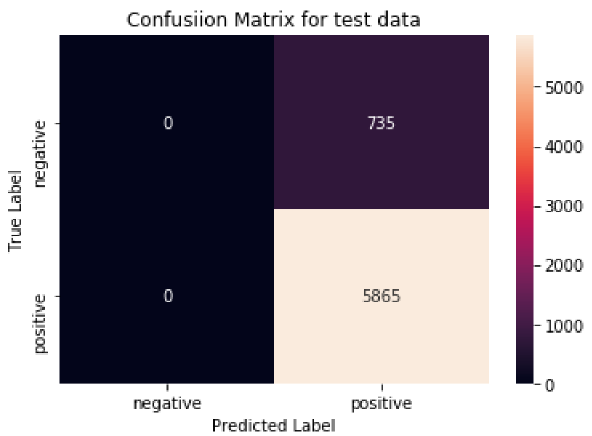
In [75]:

```
KD_TREE(XTB_TFIDF,Y_TRAIN_KD,XCV_TFIDF,Y_CV_KD,XTEST_TFIDF,Y_TEST_KD)
```




```
BEST K is 49
```





[5.2.3] Applying KNN kd-tree on AVG W2V, SET 7

In [56]:

```
# Train your own Word2Vec model using your own text corpus
i=0
list_of_sentance=[]
for sentance in X_TRAIN_KD:
    list_of_sentance.append(sentance.split())

 # min_count = 5 considers only words that occured atleast 5 times
w2v_model=Word2Vec(list_of_sentance,min_count=5,size=50, workers=4)

w2v_words = list(w2v_model.wv.vocab)
print("number of words that occured minimum 5 times ",len(w2v_words))
```

```
C:\Users\KIRTIMAN\Anaconda3\lib\site-packages\gensim\models\base_any2vec.py:743: UserWarning: C ex
tension not loaded, training will be slow. Install a C compiler and reinstall gensim for fast trai
```

```
ning.
  "C extension not loaded, training will be slow. "
```

```
number of words that occured minimum 5 times  5739
```

In [57]:

```python
def AVGW2V(X_test):
    '''
    returns average word2vec
    '''
    i=0
    list_of_sentance=[]
    for sentance in X_test:
        list_of_sentance.append(sentance.split())
    test_vectors = []; # the avg-w2v for each sentence/review is stored in this list
    for sent in tqdm(list_of_sentance): # for each review/sentence
        sent_vec = np.zeros(50) # as word vectors are of zero length 50, you might need to change t
his to 300 if you use google's w2v
        cnt_words =0; # num of words with a valid vector in the sentence/review
        for word in sent: # for each word in a review/sentence
            if word in w2v_words:
                vec = w2v_model.wv[word]
                sent_vec += vec
                cnt_words += 1
        if cnt_words != 0:
            sent_vec /= cnt_words
        test_vectors.append(sent_vec)
    return test_vectors
```

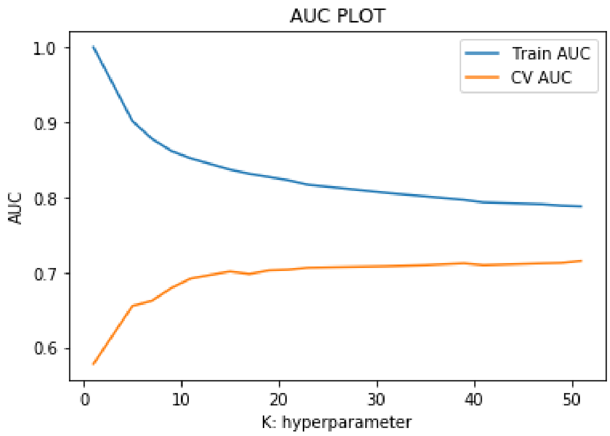
In [58]:

```python
AV_TRAIN = AVGW2V(X_TRAIN_KD)
AV_CV = AVGW2V(X_CV_KD)
AV_TEST = AVGW2V(X_TEST_KD)
```

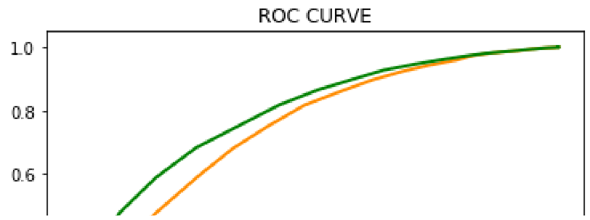
```
100%|██████████| 8978/8978 [00:07<00:00, 1138.77it/s]
100%|██████████| 4422/4422 [00:03<00:00, 1149.56it/s]
100%|██████████| 6600/6600 [00:06<00:00, 1082.55it/s]
```

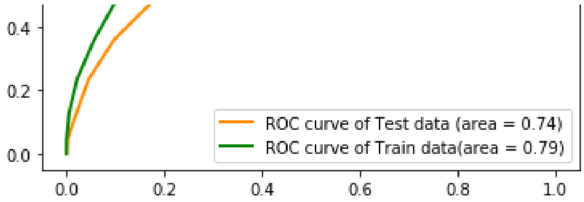
In [74]:

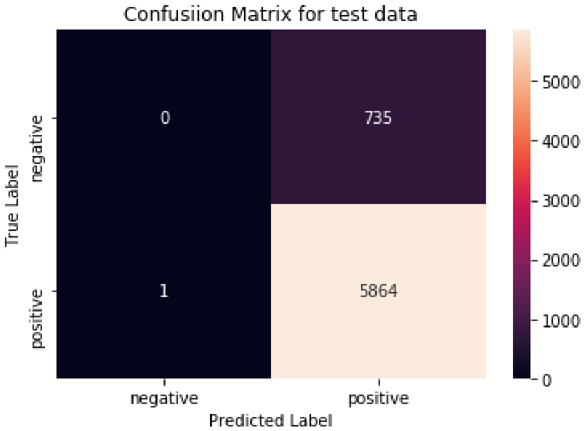
```python
KD_TREE(AV_TRAIN,Y_TRAIN_KD,AV_CV,Y_CV_KD,AV_TEST,Y_TEST_KD)
```




```
BEST K is 51
```







[5.2.4] Applying KNN kd-tree on TFIDF W2V, SET 8

In [60]:

```
model = TfidfVectorizer()
model.fit(X_TRAIN_KD)

dictionary = dict(zip(model.get_feature_names(), list(model.idf_)))

# TF-IDF weighted Word2Vec
tfidf_feat = model.get_feature_names() # tfidf words/col-names
# final_tf_idf is the sparse matrix with row= sentence, col=word and cell_val = tfidf
```

In [61]:

```
def TFIDFW2V(test):
    '''
    Returns tfidf word2vec
    '''
    tfidf_sent_vectors = []; # the tfidf-w2v for each sentence/review is stored in this list
    i=0
    list_of_sentance=[]
    for sentance in test:
        list_of_sentance.append(sentance.split())

    for sent in tqdm(list_of_sentance): # for each review/sentence
        sent_vec = np.zeros(50) # as word vectors are of zero length
        weight_sum =0; # num of words with a valid vector in the sentence/review
        for word in sent: # for each word in a review/sentence
            if word in w2v_words and word in tfidf_feat:
                vec = w2v_model.wv[word]
                tf_idf = dictionary[word]*(sent.count(word)/len(sent))
                sent_vec += (vec * tf_idf)
                weight_sum += tf_idf
        if weight_sum != 0:
            sent_vec /= weight_sum
        tfidf_sent_vectors.append(sent_vec)

    return tfidf_sent_vectors
```

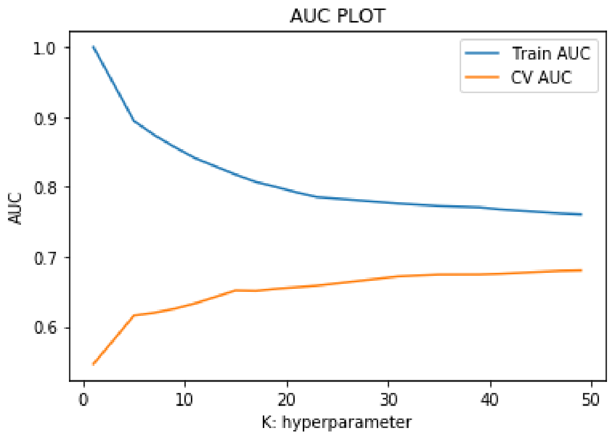
In [62]:

```
TRAIN_TFIDF_W2V = TFIDFW2V(X_TRAIN_KD)
CV_TFIDF_W2V = TFIDFW2V(X_CV_KD)
TEST_TFIDF_W2V = TFIDFW2V(X_TEST_KD)
```

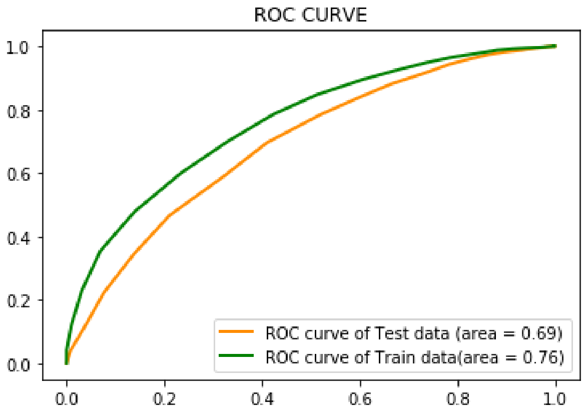
```
100%|██████████| 8978/8978 [01:05<00:00, 136.07it/s]
100%|██████████| 4422/4422 [00:30<00:00, 142.71it/s]
100%|██████████| 6600/6600 [00:48<00:00, 135.57it/s]
```

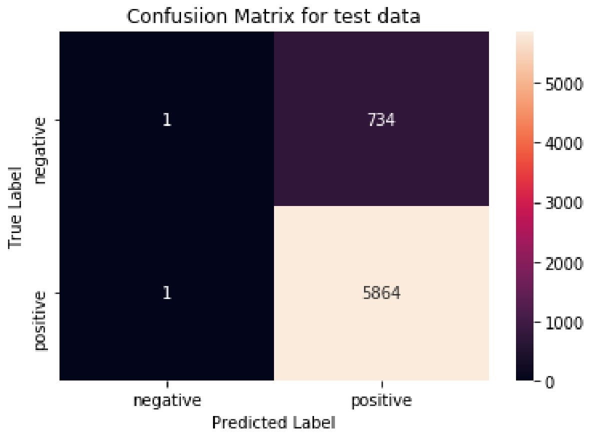
In [83]:

```
KD_TREE(TRAIN_TFIDF_W2V,Y_TRAIN_KD,CV_TFIDF_W2V,Y_CV_KD,TEST_TFIDF_W2V,Y_TEST_KD)
```




```
BEST K is 49
```





[6] Conclusions

In [ ]:

```
# Please compare all your models using Prettytable library
```

In [78]:

```
from prettytable import PrettyTable
```

In [85]:

```
from prettytable import PrettyTable
X = PrettyTable()
print(" "*40+"CONCLUSION")
print("="*100)
X.field_names = ["ALGORITHM","METHOD", "BEST_K", "TRAIN AUC ","TEST_AUC"]
X.add_row(["BRUTE FORCE","BOW", 9,0.84,0.64])
X.add_row(["KD TREE","BOW", 51,0.79,0.73])

X.add_row(["BRUTE FORCE","TFIDF", 1,1,0.51])
X.add_row(["KD TREE","TFIDF", 49,0.87,0.73])

X.add_row(["BRUTE FORCE","AVG W2V", 51,0.86,0.84])
X.add_row(["KD TREE","AVG W2V", 51,0.79,0.74])

X.add_row(["BRUTE FORCE"," TFIDF W2V", 51,0.84,0.80])
X.add_row(["KD TREE","TFIDF W2V", 49,0.76,0.69])

print(X)
```

```
                                        CONCLUSION
====================================================================================================

+-------------+------------+--------+------------+----------+
|  ALGORITHM  |   METHOD   | BEST_K | TRAIN AUC  | TEST_AUC |
+-------------+------------+--------+------------+----------+
| BRUTE FORCE |    BOW     |   9    |    0.84    |   0.64   |
|   KD TREE   |    BOW     |   51   |    0.79    |   0.73   |
| BRUTE FORCE |   TFIDF    |   1    |     1      |   0.51   |
|   KD TREE   |   TFIDF    |   49   |    0.87    |   0.73   |
| BRUTE FORCE |  AVG W2V   |   51   |    0.86    |   0.84   |
|   KD TREE   |  AVG W2V   |   51   |    0.79    |   0.74   |
| BRUTE FORCE |  TFIDF W2V |   51   |    0.84    |   0.8    |
|   KD TREE   | TFIDF W2V  |   49   |    0.76    |   0.69   |
+-------------+------------+--------+------------+----------+
```

1.Maximum AUC is coming Brute Force AvgW2v ,as we seen from table TFIDFw2V also has very similar AUC.

1. Brute Force AVGW2V ,TFIDFW2V and KD Tree AVGw2V and TFIDFW2V give very similar result.
2. As we seen from confusion matrix, model is not able to classify negative points very well,one of the reason behind that is For Brute force or for KDTree we took 50K and 20K points respectively and Data is highly imabalanced.

** REFRENCE: I Have refered to an sample solution .