Amazon Fine Food Reviews Analysis

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

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

```
C:\Users\sujpanda\Anaconda3\lib\site-packages\gensim\utils.py:1212: UserWarnin
g: detected Windows; aliasing chunkize to chunkize_serial
  warnings.warn("detected Windows; aliasing chunkize to chunkize_serial")
```

In [2]:

```
# using SQLite Table to read data.
con = sqlite3.connect('C:\\Users\\sujpanda\\Desktop\\applied\\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 a
# 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 !=
# for tsne assignment you can take 5k data points

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

# Give reviews with Score>3 a positive rating(1), and reviews with a score<3
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 (100000, 10)
```

Out[2]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominat
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	

In [3]:

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

In [4]:
```
print(display.shape)
display.head()
```

(80668, 7)

Out[4]:

	UserId	ProductId	ProfileName	Time	Score	Text	COUNT(*)
0	#oc-R115TNMSPFT9I7	B007Y59HVM	Breyton	1331510400	2	Overall its just OK when considering the price...	2
1	#oc-R11D9D7SHXIJB9	B005HG9ET0	Louis E. Emory "hoppy"	1342396800	5	My wife has recurring extreme muscle spasms, u...	3
2	#oc-R11DNU2NBKQ23Z	B007Y59HVM	Kim Cieszykowski	1348531200	1	This coffee is horrible and unfortunately not ...	2
3	#oc-R11O5J5ZVQE25C	B005HG9ET0	Penguin Chick	1346889600	5	This will be the bottle that you grab from the...	3
4	#oc-R12KPBODL2B5ZD	B007OSBE1U	Christopher P. Presta	1348617600	1	I didnt like this coffee. Instead of telling y...	2

In [5]:
```
display[display['UserId']=='AZY10LLTJ71NX']
```

Out[5]:

	UserId	ProductId	ProfileName	Time	Score	Text	COUNT(*)
80638	AZY10LLTJ71NX	B006P7E5ZI	undertheshrine "undertheshrine"	1334707200	5	I was recommended to try green tea extract to ...	5

In [6]:
```
display['COUNT(*)'].sum()
```

Out[6]: 393063

[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]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenomir
0	78445	B000HDL1RQ	AR5J8UI46CURR	Geetha Krishnan	2	
1	138317	B000HDOPYC	AR5J8UI46CURR	Geetha Krishnan	2	
2	138277	B000HDOPYM	AR5J8UI46CURR	Geetha Krishnan	2	
3	73791	B000HDOPZG	AR5J8UI46CURR	Geetha Krishnan	2	
4	155049	B000PAQ75C	AR5J8UI46CURR	Geetha Krishnan	2	

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, in
```

In [9]:
```
#Deduplication of entries
final=sorted_data.drop_duplicates(subset={"UserId","ProfileName","Time","Text
final.shape
```

Out[9]: (87775, 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]: 87.775

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]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenomir
0	64422	B000MIDROQ	A161DK06JJMCYF	J. E. Stephens "Jeanne"	3	
1	44737	B001EQ55RW	A2V0I904FH7ABY	Ram	3	

In [12]:
```
final=final[final.HelpfulnessNumerator<=final.HelpfulnessDenominator]
```

In [13]:
```
#Before starting the next phase of preprocessing lets see the number of entri
print(final.shape)

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

```
(87773, 10)
```

Out[13]:
```
1    73592
0    14181
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)
```

```
My dogs loves this chicken but its a product from China, so we wont be buying i
t anymore.  Its very hard to find any chicken products made in the USA but they
are out there, but this one isnt.  Its too bad too because its a good product b
ut I wont take any chances till they know what is going on with the china impor
ts.
==================================================
The Candy Blocks were a nice visual for the Lego Birthday party but the candy h
as little taste to it.  Very little of the 2 lbs that I bought were eaten and I
threw the rest away.  I would not buy the candy again.
==================================================
was way to hot for my blood, took a bite and did a jig  lol
==================================================
My dog LOVES these treats. They tend to have a very strong fish oil smell. So i
f you are afraid of the fishy smell, don't get it. But I think my dog likes it
because of the smell. These treats are really small in size. They are great for
training. You can give your dog several of these without worrying about him ove
r eating. Amazon's price was much more reasonable than any other retailer. You
can buy a 1 pound bag on Amazon for almost the same price as a 6 ounce bag at o
ther retailers. It's definitely worth it to buy a big bag if your dog eats them
a lot.
==================================================
```

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)
```

```
My dogs loves this chicken but its a product from China, so we wont be buying i
t anymore.  Its very hard to find any chicken products made in the USA but they
are out there, but this one isnt.  Its too bad too because its a good product b
ut I wont take any chances till they know what is going on with the china impor
ts.
```

In [16]:
```
# https://stackoverflow.com/questions/16206380/python-beautifulsoup-how-to-re
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)
```

```
My dogs loves this chicken but its a product from China, so we wont be buying i
t anymore.  Its very hard to find any chicken products made in the USA but they
are out there, but this one isnt.  Its too bad too because its a good product b
ut I wont take any chances till they know what is going on with the china impor
ts.
==================================================
The Candy Blocks were a nice visual for the Lego Birthday party but the candy h
as little taste to it.  Very little of the 2 lbs that I bought were eaten and I
threw the rest away.  I would not buy the candy again.
==================================================
was way to hot for my blood, took a bite and did a jig  lol
==================================================
My dog LOVES these treats. They tend to have a very strong fish oil smell. So i
f you are afraid of the fishy smell, don't get it. But I think my dog likes it
because of the smell. These treats are really small in size. They are great for
training. You can give your dog several of these without worrying about him ove
r eating. Amazon's price was much more reasonable than any other retailer. You
can buy a 1 pound bag on Amazon for almost the same price as a 6 ounce bag at o
ther retailers. It's definitely worth it to buy a big bag if your dog eats them
a lot.
```

In [17]:

```
# https://stackoverflow.com/a/47091490/4084039
import re

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("="*50)
```

```
was way to hot for my blood, took a bite and did a jig  lol
==================================================
```

In [19]:

```
#remove words with numbers python: https://stackoverflow.com/a/18082370/40840
sent_0 = re.sub("\S*\d\S*", "", sent_0).strip()
print(sent_0)
```

```
My dogs loves this chicken but its a product from China, so we wont be buying i
t anymore.  Its very hard to find any chicken products made in the USA but they
are out there, but this one isnt.  Its too bad too because its a good product b
ut I wont take any chances till they know what is going on with the china impor
ts.
```

In [20]:

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

```
was way to hot for my blood took a bite and did a jig lol
```

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 1s

stopwords= set(['br', 'the', 'i', 'me', 'my', 'myself', 'we', 'our', 'ours',
            "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he
            'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'i
            'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', '
            'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have',
            'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'beca
            'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into',
            'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on',
            'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', '
            'most', 'other', 'some', 'such', 'only', 'own', 'same', 'so', 'th
            's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "shoul
            've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn',
            "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", '
            "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shou
            'won', "won't", 'wouldn', "wouldn't"])
```

In [22]:
```
# Combining all the above stundents
from tqdm import tqdm
preprocessed_reviews = []
# tqdm is for printing the status bar
for sentance in tqdm(final['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)
    # https://gist.github.com/sebleier/554280
    sentance = ' '.join(e.lower() for e in sentance.split() if e.lower() not
    preprocessed_reviews.append(sentance.strip())
```

```
100%|██████████| 87773/87773 [01:02<00:00, 1396.85it/s]
```

In [23]:
```
preprocessed_reviews[1500]
```

Out[23]:
```
'way hot blood took bite jig lol'
```

[3.2] Preprocessing Review Summary

In [24]:
```
from tqdm import tqdm
preprocessed_summary = []
# tqdm is for printing the status bar
for sentance in tqdm(final['Summary'].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)
    # https://gist.github.com/sebleier/554280
    sentance = ' '.join(e.lower() for e in sentance.split() if e.lower() not
    preprocessed_summary.append(sentance.strip())
```

```
 37%|███▋      | 32715/87773 [00:11<00:22, 2405.44it/s]C:\Users\sujpanda\Anacon
da3\lib\site-packages\bs4\__init__.py:219: UserWarning: "b'...'" looks like a f
ilename, not markup. You should probably open this file and pass the filehandle
into Beautiful Soup.
  ' Beautiful Soup.' % markup)
 70%|███████   | 61326/87773 [00:22<00:09, 2647.88it/s]C:\Users\sujpanda\Anacon
da3\lib\site-packages\bs4\__init__.py:219: UserWarning: "b'...'" looks like a f
ilename, not markup. You should probably open this file and pass the filehandle
into Beautiful Soup.
  ' Beautiful Soup.' % markup)
 75%|███████▍  | 65472/87773 [00:23<00:07, 3079.02it/s]C:\Users\sujpanda\Anacon
da3\lib\site-packages\bs4\__init__.py:219: UserWarning: "b'...'" looks like a f
ilename, not markup. You should probably open this file and pass the filehandle
into Beautiful Soup.
  ' Beautiful Soup.' % markup)
 96%|█████████▌| 84147/87773 [00:30<00:01, 2489.15it/s]C:\Users\sujpanda\Anacon
da3\lib\site-packages\bs4\__init__.py:219: UserWarning: "b'...'" looks like a f
ilename, not markup. You should probably open this file and pass the filehandle
into Beautiful Soup.
  ' Beautiful Soup.' % markup)
100%|██████████| 87773/87773 [00:32<00:00, 2730.40it/s]
```

[5] Assignment 8: Decision Trees

1. **Apply Decision Trees 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. **The hyper paramter tuning (best `depth` in range [1, 5, 10, 50, 100, 500, 100], and the best `min_samples_split` in range [5, 10, 100, 500])**

    - Find the best hyper parameter which will give the maximum [AUC (https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/receiver-operating-characteristic-curve-roc-curve-and-auc-1/)](https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/receiver-operating-characteristic-curve-roc-curve-and-auc-1/) 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

3. **Graphviz**

- Visualize your decision tree with Graphviz. It helps you to understand how a decision is being made, given a new vector.
- Since feature names are not obtained from word2vec related models, visualize only BOW & TFIDF decision trees using Graphviz
- Make sure to print the words in each node of the decision tree instead of printing its index.
- Just for visualization purpose, limit max_depth to 2 or 3 and either embed the generated images of graphviz in your notebook, or directly upload them as .png files.

4. **Feature importance**

- Find the top 20 important features from both feature sets Set 1 and Set 2 using `feature_importances_` method of Decision Tree Classifier (https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html) and print their corresponding feature names

5. **Feature engineering**

- To increase the performance of your model, you can also experiment with with feature engineering like :
  - Taking length of reviews as another feature.
  - Considering some features from review summary as well.

6. **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 (https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/confusion-matrix-tpr-fpr-fnr-tnr-1/) with predicted and original labels of test data points. Please visualize your confusion matrices using seaborn heatmaps. (https://seaborn.pydata.org/generated/seaborn.heatmap.html)

7. **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 (http://zetcode.com/python/prettytable/)

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. (https://soundcloud.com/applied-ai-course/leakage-bow-and-tfidf)](https://soundcloud.com/applied-ai-course/leakage-bow-and-tfidf)

Some utility functions

In [25]:
```
## Some utility functions

def check_trade_off(X_train,X_test,y_train,y_test):

    from sklearn.metrics import roc_curve
    from sklearn.metrics import roc_auc_score

    [{'max_depth': [1, 5, 10, 50, 100, 500, 1000]}]

    C_range1 = ['1','5','10','50','100','500','1000']
    C_range = [1, 5, 10, 50, 100, 500, 1000]
    dummy_range = [1,2,3,4,5,6,7]

    auc_scores =[]
    auc_train_scores = []

    i = 0
    for i in C_range:
        clf =DecisionTreeClassifier(max_depth=i)

        # fitting the model on crossvalidation train
        clf.fit(X_train, y_train)


        #evaluate AUC score.
        probs = clf.predict_proba(X_test)
        probs = probs[:, 1]
        # calculate AUC
        auc = roc_auc_score(y_test, probs)
        print('AUC: %.3f' % auc)
        auc_scores.append(auc)

    print('#######################################################')
    print('AUC from train data ###################################')
    i = 0
    for i in C_range:
        clf =DecisionTreeClassifier(max_depth=i)

        # fitting the model on crossvalidation train
        clf.fit(X_train, y_train)

        #evaluate AUC score.
        probs = clf.predict_proba(X_train)
        probs = probs[:, 1]
        # calculate AUC
        auc = roc_auc_score(y_train, probs)
        print('AUC: %.3f' % auc)
        auc_train_scores.append(auc)

    plt.plot(dummy_range, auc_scores,'r')
    plt.plot(dummy_range, auc_train_scores,'b')
    plt.xticks(dummy_range, C_range1, rotation='vertical')
    for xy in zip(dummy_range, auc_scores):
        plt.annotate('(%f, %f)' % xy, xy=xy, textcoords='data')
    for xy in zip(dummy_range, auc_train_scores):
        plt.annotate('(%f, %f)' % xy, xy=xy, textcoords='data')
```

```
    plt.xlabel('max_depth')
    plt.ylabel('auc_scores')
    plt.show()
```

In [26]:

```
def dt_results(maxDepth,minimumSampleSplit,X_train,X_test,y_train,y_test):
    # roc curve and auc
    from sklearn.metrics import roc_curve
    from sklearn.metrics import roc_auc_score
    from matplotlib import pyplot
    # ============================== Decision Tree===========================
    clf = DecisionTreeClassifier(max_depth=maxDepth,min_samples_split=minimum

    # fitting the model
    clf.fit(X_train, y_train)

    # predict the response
    pred = clf.predict(X_test)

    # evaluate accuracy
    acc = accuracy_score(y_test, pred) * 100
    print('\nThe accuracy of the DT classifier for maxDepth = %d and min spli

    probs = clf.predict_proba(X_test)
    probs = probs[:, 1]
    # calculate AUC
    auc = roc_auc_score(y_test, probs)
    print('AUC: %.3f' % auc)
    # calculate roc curve
    fpr, tpr, thresholds = roc_curve(y_test, probs)
    #######################Train data###############################

    clf.fit(X_train, y_train)
    pred_train = clf.predict(X_train)
    probs = clf.predict_proba(X_train)
    probs = probs[:, 1]
    # calculate AUC
    auc = roc_auc_score(y_train, probs)
    print('AUC: %.3f' % auc)
    fpr1, tpr1, thresholds1 = roc_curve(y_train, probs)


    # plot no skill
    pyplot.plot([0, 1], [0, 1], linestyle='--')
    # plot the roc curve for the model
    pyplot.plot(fpr, tpr, marker='.',label='test')
    pyplot.plot(fpr1, tpr1, marker='*',label='train')
    pyplot.legend()
    # show the plot
    pyplot.show()
    from sklearn.metrics import confusion_matrix
    con_mat = confusion_matrix(y_test, pred, [0, 1])
    con_mat_train = confusion_matrix(y_train,pred_train,[0,1])
    return con_mat,con_mat_train,clf
```

In [27]:
```
def showHeatMap(con_mat):
    class_label = ["negative", "positive"]
    df_cm = pd.DataFrame(con_mat, index = class_label, columns = class_label)
    sns.heatmap(df_cm, annot = True, fmt = "d")
    plt.title("Confusion Matrix")
    plt.xlabel("Predicted Label")
    plt.ylabel("True Label")
    plt.show()
```

Applying Decision Trees

[5.1] Applying Decision Trees on BOW, SET 1

In [28]:
```
from sklearn.cross_validation import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score
from sklearn.cross_validation import cross_val_score
from collections import Counter
from sklearn.metrics import accuracy_score
from sklearn import cross_validation
from sklearn.grid_search import GridSearchCV
import warnings
warnings.filterwarnings("ignore")
```

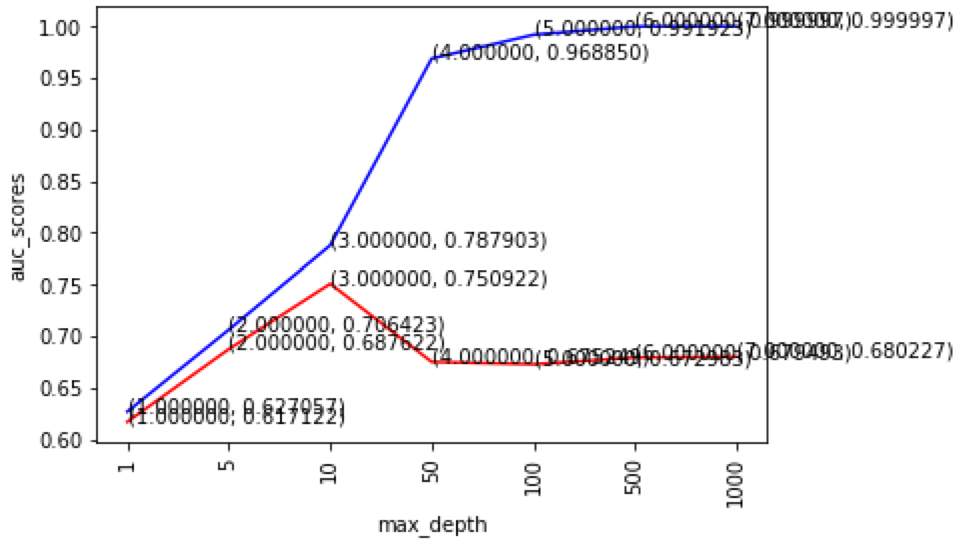
```
C:\Users\sujpanda\Anaconda3\lib\site-packages\sklearn\cross_validation.py:41: D
eprecationWarning: This module was deprecated in version 0.18 in favor of the m
odel_selection module into which all the refactored classes and functions are m
oved. Also note that the interface of the new CV iterators are different from t
hat of this module. This module will be removed in 0.20.
  "This module will be removed in 0.20.", DeprecationWarning)
C:\Users\sujpanda\Anaconda3\lib\site-packages\sklearn\grid_search.py:42: Deprec
ationWarning: This module was deprecated in version 0.18 in favor of the model_
selection module into which all the refactored classes and functions are moved.
This module will be removed in 0.20.
  DeprecationWarning)
```

In [143]:
```
X_1, X_test, y_1, y_test = cross_validation.train_test_split(preprocessed_rev
```

In [144]:

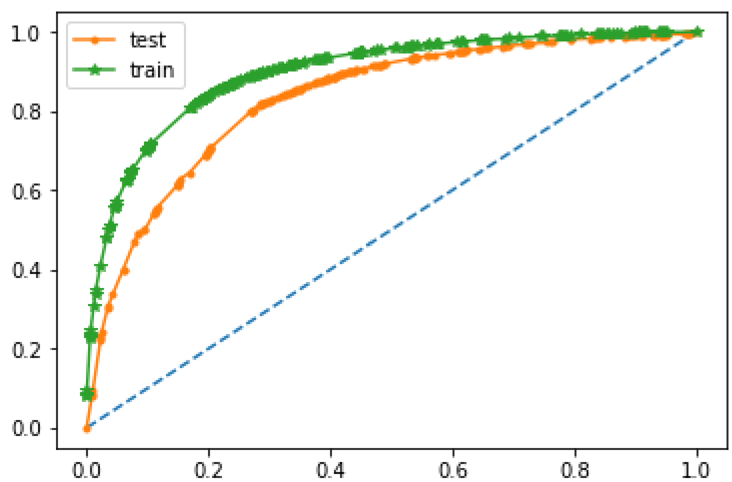
```
count_vect = CountVectorizer()
final_counts = count_vect.fit_transform(X_1)
final_test_count = count_vect.transform(X_test)

# split the train data set into cross validation train and cross validation t
X_tr, X_cv, y_tr, y_cv = cross_validation.train_test_split(X_1, y_1, test_siz

final_counts_tr_cv = count_vect.transform(X_tr)
final_test_count_cv = count_vect.transform(X_cv)

tuned_parameters = [{'max_depth': [1, 5, 10, 50, 100, 500, 1000],'min_samples

#Using GridSearchCV
model = GridSearchCV(DecisionTreeClassifier(), tuned_parameters, scoring = 'r
model.fit(final_counts_tr_cv, y_tr)

print(model.best_estimator_)
print(model.score(final_test_count_cv, y_cv))

check_trade_off(final_counts_tr_cv,final_test_count_cv,y_tr,y_cv)
```

```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=50,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=500,
            min_weight_fraction_leaf=0.0, presort=False, random_state=None,
            splitter='best')
0.8217225902108642
AUC: 0.617
AUC: 0.688
AUC: 0.751
AUC: 0.675
AUC: 0.673
AUC: 0.679
AUC: 0.680
#######################################################
AUC from train data ###################################
AUC: 0.627
AUC: 0.706
AUC: 0.788
AUC: 0.969
AUC: 0.992
AUC: 1.000
AUC: 1.000
```



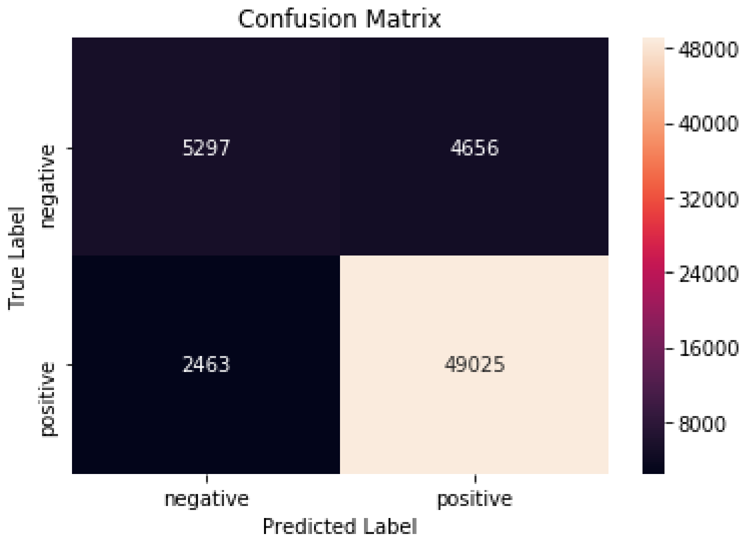
In [145]:
```
1  con_mat,con_mat_train,clf = dt_results(50,500,final_counts,final_test_count,y
```

```
The accuracy of the DT classifier for maxDepth = 50 and min split = 500 is 85.8
72702%
AUC: 0.832
AUC: 0.898
```



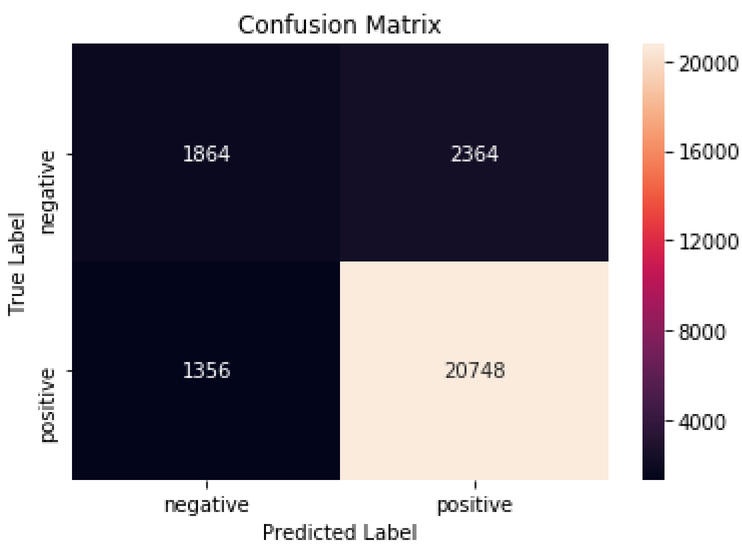
Observation: My model predicted with accuracy 85% with AUC: 0.833. There is no much difference between train and test ROC curve

In [146]:
```
showHeatMap(con_mat_train)
```



Observation: My model would have predicted 2463 + 4656 points wrongly even with train data

In [147]:
```
showHeatMap(con_mat)
```



Observation: My model misclassified 1356 + 2364 points

[5.1.1] Top 20 important features from SET 1

In [148]:
```
feature_names = np.array(count_vect.get_feature_names())
featureDict = dict(zip(feature_names, clf.feature_importances_))
sortedFeatures = sorted(featureDict.items(), key=lambda x: x[1],reverse=True)
print(type(sortedFeatures))
for i in range(0,20):
    print(sortedFeatures[i])
```

```
<class 'list'>
('not', 0.10556541192374382)
('great', 0.06328011594676017)
('worst', 0.04198736350781279)
('disappointed', 0.041460848322431014)
('money', 0.03469326192544119)
('horrible', 0.027550887530162 25)
('return', 0.02734130622754287)
('best', 0.02631809150225828)
('delicious', 0.02371161236410449)
('love', 0.02230452293131868)
('terrible', 0.021929296168877045)
('good', 0.019721769822422328)
('awful', 0.01744950993510839)
('waste', 0.01686007552218982)
('perfect', 0.015630476695974057)
('loves', 0.015271687168840937)
('disappointing', 0.015263760669904461)
('threw', 0.014663220813299996)
('nice', 0.01335400790815546)
('bad', 0.012334416821108925)
```

[5.1.2] Graphviz visualization of Decision Tree on BOW, SET 1

In [149]:
```
import graphviz
import pydotplus
import collections
from sklearn import tree
clf = DecisionTreeClassifier(max_depth=2)
clf.fit(final_counts, y_1)
dot_data = tree.export_graphviz(clf, feature_names = count_vect.get_feature_n
graph = pydotplus.graph_from_dot_data(dot_data)
colors = ('turquoise', 'orange')
edges = collections.defaultdict(list)
for edge in graph.get_edge_list():
    edges[edge.get_source()].append(int(edge.get_destination()))

for edge in edges:
    edges[edge].sort()
    for i in range(2):
        dest = graph.get_node(str(edges[edge][i]))[0]
        dest.set_fillcolor(colors[i])

graph.write_png('tree.png')
```

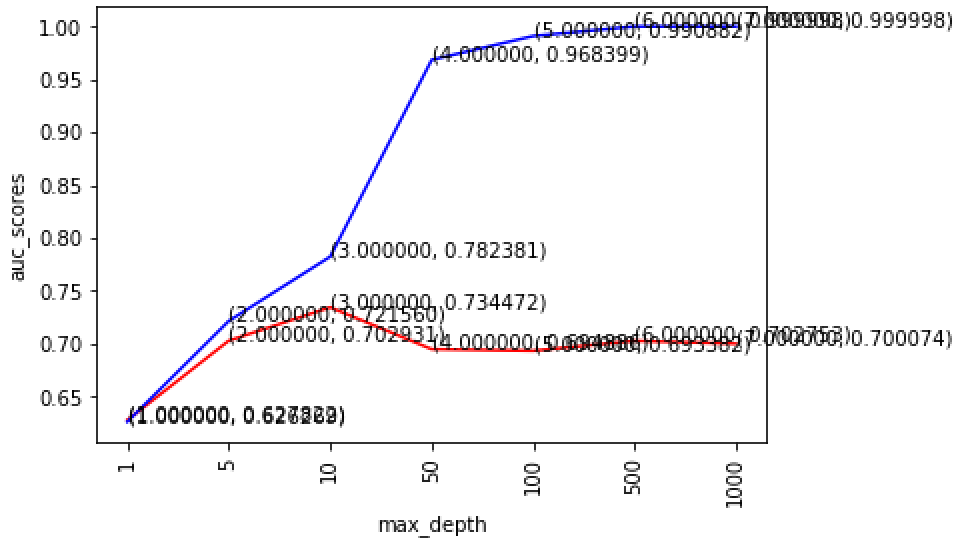
Out[149]: True

[5.2] Applying Decision Trees on TFIDF, SET 2

In [30]:
```
X_1, X_test, y_1, y_test = cross_validation.train_test_split(preprocessed_rev
```

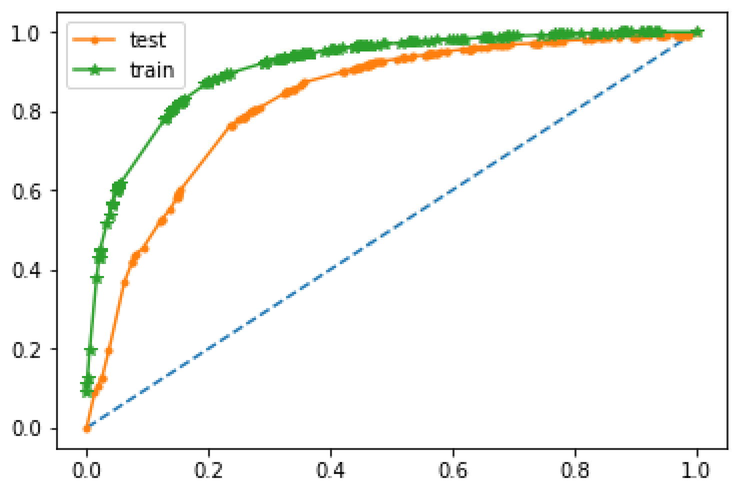
In [31]:
```
tf_idf_vect = TfidfVectorizer(ngram_range=(1,2), min_df=10)
tf_idf_vect.fit(X_1)
final_tf_idf = tf_idf_vect.transform(X_1)
final_test_count = tf_idf_vect.transform(X_test)

# split the train data set into cross validation train and cross validation t
X_tr, X_cv, y_tr, y_cv = cross_validation.train_test_split(X_1, y_1, test_siz

final_counts_tr_cv = tf_idf_vect.transform(X_tr)
final_test_count_cv = tf_idf_vect.transform(X_cv)

tuned_parameters = [{'max_depth': [1, 5, 10, 50, 100, 500, 1000],'min_samples

#Using GridSearchCV
model = GridSearchCV(DecisionTreeClassifier(), tuned_parameters, scoring = 'r
model.fit(final_counts_tr_cv, y_tr)

print(model.best_estimator_)
print(model.score(final_test_count_cv, y_cv))

check_trade_off(final_counts_tr_cv,final_test_count_cv,y_tr,y_cv)
```

```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=50,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=500,
            min_weight_fraction_leaf=0.0, presort=False, random_state=None,
            splitter='best')
0.8175195450851837
AUC: 0.628
AUC: 0.703
AUC: 0.734
AUC: 0.695
AUC: 0.693
AUC: 0.703
AUC: 0.700
#######################################################
AUC from train data ###################################
AUC: 0.626
AUC: 0.722
AUC: 0.782
AUC: 0.968
AUC: 0.991
AUC: 1.000
AUC: 1.000
```



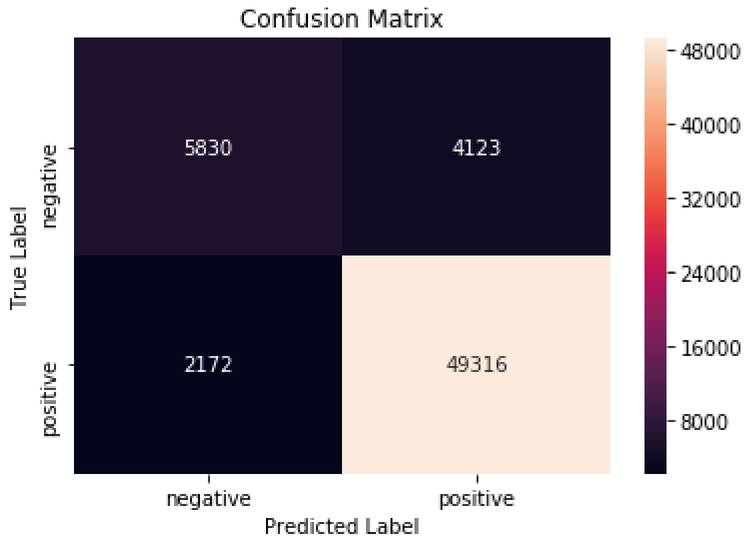
In [32]:
```
1 con_mat,con_mat_train,clf = dt_results(50,500,final_tf_idf,final_test_count,y
```

```
The accuracy of the DT classifier for maxDepth = 50 and min split = 500 is 86.1
80313%
AUC: 0.825
AUC: 0.911
```



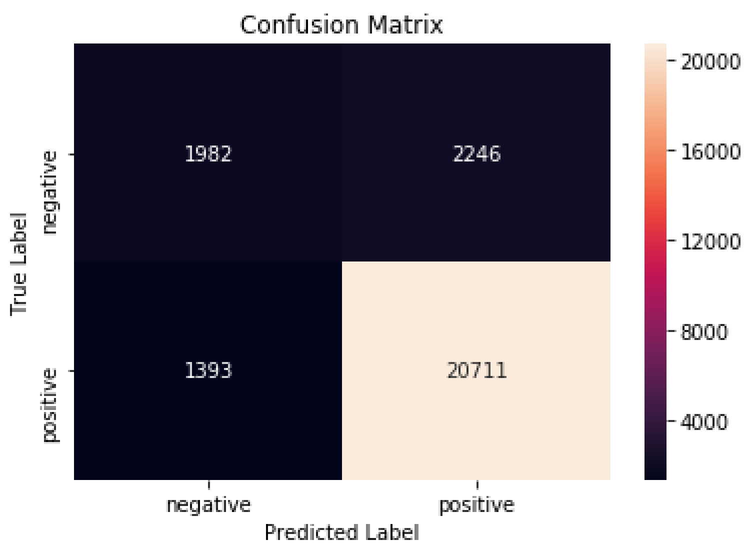
Observation: My model has predicted 86% accuracy with AUC: 0.825

In [33]:
```
showHeatMap(con_mat_train)
```



Observation: My model predicted 2172 + 4123 points wrongly in train data

In [35]:
```
showHeatMap(con_mat)
```



Observation: My model predicted 1393 + 2246 points wrongly

[5.2.1] Top 20 important features from SET 2

In [36]:
```
feature_names = np.array(tf_idf_vect.get_feature_names())
featureDict = dict(zip(feature_names, clf.feature_importances_))
sortedFeatures = sorted(featureDict.items(), key=lambda x: x[1],reverse=True)
print(type(sortedFeatures))
for i in range(0,20):
    print(sortedFeatures[i])
```

```
<class 'list'>
('not', 0.10516319649370136)
('great', 0.05574250687918252)
('disappointed', 0.039630778078183465)
('worst', 0.03771103774223818)
('horrible', 0.026077555904695973)
('not buy', 0.02573570932893479)
('not worth', 0.025698879081758955)
('return', 0.022609828389066555)
('terrible', 0.022198814641220536)
('bad', 0.021413273008521163)
('awful', 0.018711373537716715)
('waste', 0.018038406237556696)
('love', 0.015707544506488345)
('delicious', 0.015009217862943312)
('refund', 0.014709302315080703)
('not recommend', 0.014437736293673174)
('disappointing', 0.014042244798650627)
('waste money', 0.013672608221397712)
```

[5.2.2] Graphviz visualization of Decision Tree on TFIDF, SET 2

In [39]:
```
import graphviz
import pydotplus
import collections
from sklearn import tree
clf = DecisionTreeClassifier(max_depth=2)
clf.fit(final_tf_idf, y_1)
dot_data = tree.export_graphviz(clf, feature_names = tf_idf_vect.get_feature_
graph = pydotplus.graph_from_dot_data(dot_data)
colors = ('turquoise', 'orange')
edges = collections.defaultdict(list)
for edge in graph.get_edge_list():
    edges[edge.get_source()].append(int(edge.get_destination()))

for edge in edges:
    edges[edge].sort()
    for i in range(2):
        dest = graph.get_node(str(edges[edge][i]))[0]
        dest.set_fillcolor(colors[i])

graph.write_png('treetfidf.png')
```

Out[39]: True

[5.3] Applying Decision Trees on AVG W2V, SET 3

In [40]:

```
X_train, X_test, y_1, y_test = cross_validation.train_test_split(preprocessed
```

In [41]:

```
i=0
list_of_sentance=[]
for sentance in X_train:
    list_of_sentance.append(sentance.split())

w2v_model=Word2Vec(list_of_sentance,min_count=5,size=50, workers=4)
w2v_words = list(w2v_model.wv.vocab)

# average Word2Vec
# compute average word2vec for each review.
sent_vectors = []; # the avg-w2v for each sentence/review is stored in this l
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 migh
    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
    sent_vectors.append(sent_vec)
print(sent_vectors[0])



i=0
list_of_test_sentance=[]
for sentance in X_test:
    list_of_test_sentance.append(sentance.split())

test_sent_vectors = [];

for sent in tqdm(list_of_test_sentance): # for each review/sentence
    sent_vec = np.zeros(50) # as word vectors are of zero length 50, you migh
    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_sent_vectors.append(sent_vec)
print(test_sent_vectors[0])

```

```
100%|██████████| 61441/61441 [02:47<00:00, 366.80it/s]

[-1.10754768e+00  2.00193054e-01 -4.16910136e-01  3.16457859e-01
  1.67205180e-01  2.26236710e-01  1.63417535e-01  7.77784129e-04
 -2.26487562e-01 -6.12884563e-01 -3.77290405e-01  6.69056482e-02
  8.02721659e-01  3.09768991e-02  1.54656387e+00  1.16882547e-01
  5.74342138e-01 -3.87559963e-01  9.58900229e-01 -6.66402416e-01
  7.56670846e-02  3.80231652e-01 -1.59714155e-01  1.76354673e-01
  1.56241445e-01 -2.10104528e-01 -7.18050218e-01 -7.24558968e-01
```

```
 -3.62788136e-01 -1.63233715e-01  1.31064441e-01 -1.40604076e-01
  1.01911919e-01  2.49025648e-01  8.28877521e-01  3.88748020e-01
  2.57363992e-01  1.51706172e-01  1.89670206e-01 -3.30035001e-02
  3.07788125e-01  1.34390374e-01 -1.03560609e+00 -4.67261180e-01
 -9.94279975e-02  3.30757251e-01  5.00301523e-01 -4.65017455e-01
  1.55648759e-01 -5.28523285e-01]
```

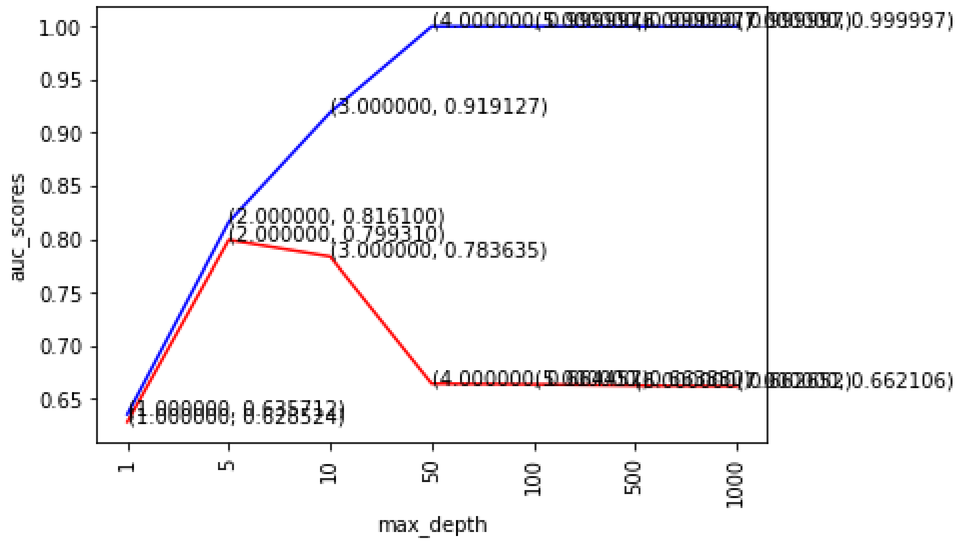
```
100%|██████████| 26332/26332 [01:13<00:00, 360.05it/s]
```

```
[-0.74806017 -0.48290271 -0.47126988  0.4013922   0.29665439 -0.45408076
  1.00896888 -0.37358878 -0.78329729 -0.19524113  0.06300695 -0.8424012
  0.0484644   1.09750664  0.04071722 -0.30163987  0.01318479  0.61584562
 -0.12420691 -0.39806303  0.19882826 -0.36472114  0.78072151 -0.86098378
  0.0171463   1.47464345 -0.01129889 -0.1263671  -0.42119131 -0.42079127
  0.62090134  0.06183675 -0.6788543   0.67373388  0.35000022  0.01159972
 -0.26829883  0.48789187  0.06580197  0.46019822 -0.29895352  0.44290327
  0.57227689  0.28952171 -0.47447127  0.20441469  0.1476062  -1.02437518
  0.49800761 -0.89132388]
```

In [42]:
```
# split the train data set into cross validation train and cross validation t
X_tr, X_cv, y_tr, y_cv = cross_validation.train_test_split(X_train, y_1, test

i=0
list_of_cv_sentance=[]
for sentance in X_tr:
    list_of_cv_sentance.append(sentance.split())

cv_train_sent_vectors = [];

for sent in tqdm(list_of_cv_sentance): # for each review/sentence
    sent_vec = np.zeros(50) # as word vectors are of zero length 50, you migh
    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
    cv_train_sent_vectors.append(sent_vec)
print(cv_train_sent_vectors[0])

i=0
list_of_cv_test_sentance=[]
for sentance in X_cv:
    list_of_cv_test_sentance.append(sentance.split())

cv_test_sent_vectors = [];

for sent in tqdm(list_of_cv_test_sentance): # for each review/sentence
    sent_vec = np.zeros(50) # as word vectors are of zero length 50, you migh
    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
    cv_test_sent_vectors.append(sent_vec)
print(cv_test_sent_vectors[0])

tuned_parameters = [{'max_depth': [1, 5, 10, 50, 100, 500, 1000],'min_samples

#Using GridSearchCV
model = GridSearchCV(DecisionTreeClassifier(), tuned_parameters, scoring = 'r
model.fit(cv_train_sent_vectors, y_tr)

print(model.best_estimator_)
print(model.score(cv_test_sent_vectors, y_cv))

check_trade_off(cv_train_sent_vectors,cv_test_sent_vectors,y_tr,y_cv)
```

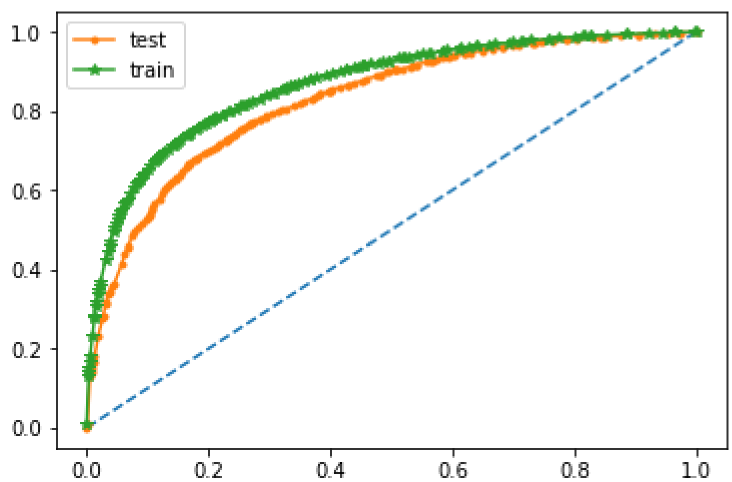
```
100%|██████████| 43008/43008 [02:04<00:00, 346.43it/s]
```

```
[-1.17079823e+00  6.30106942e-01 -3.98037020e-01  1.26815554e-01
  4.77530123e-01 -4.38024840e-01  3.82910454e-01 -7.14504091e-02
 -1.97967817e-01 -2.57848058e-01  2.28154914e-01 -7.28499447e-01
  5.58125364e-01  4.65385314e-01  1.19031763e+00 -2.57445815e-01
 -2.23603821e-02 -6.10887574e-02  6.34071672e-01 -8.80125334e-01
 -5.31466887e-01  2.93856806e-01  2.20533359e-01 -1.97766191e-01
 -2.19944531e-01  2.29905581e-02 -2.82737630e-01 -6.82443065e-01
  1.40747673e-01 -1.31457557e-01  1.03656121e+00  8.39704917e-02
 -1.07011952e-03  1.03453724e+00  8.46687205e-01  4.14343814e-01
  6.55215979e-03 -6.73246777e-01 -8.67709714e-02 -1.68713270e-01
 -3.82535082e-01 -4.62584463e-01 -3.31906426e-01 -1.80009013e-01
 -9.42691050e-02  7.90953606e-02 -6.51220036e-02 -2.22707379e-01
  4.61115037e-01 -1.56566691e-01]

100%|██████████| 18433/18433 [00:51<00:00, 359.95it/s]

[-0.65424046 -0.15814827 -0.54246559  0.41331746 -0.15781516  0.28864999
  0.18132091  0.29827008 -0.07347252 -0.01320163  0.11105867  0.3110009
  0.3764931  -0.14698873  0.9797962  -0.1694225  -0.13996538 -0.04780426
  0.18079759 -0.49656016 -0.24204404  0.03810901 -0.09690808 -0.25461577
 -0.1482041   0.70717171 -0.13982589  0.10306405 -0.55370422  0.03788182
  0.01893827 -0.03094728 -0.26235353  0.26219302  0.43014521  0.06087685
 -0.19330437 -0.36747148  0.34109956 -0.18614172  0.23745384 -0.01328977
 -0.53631501  0.02791026 -0.35983008  0.3108568   0.96507663 -0.31063059
 -0.25054761 -0.21689876]
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=10,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=500,
            min_weight_fraction_leaf=0.0, presort=False, random_state=None,
            splitter='best')
0.8177623575622466
AUC: 0.629
AUC: 0.799
AUC: 0.784
AUC: 0.664
AUC: 0.664
AUC: 0.663
AUC: 0.662
#######################################################
AUC from train data ###################################
AUC: 0.636
AUC: 0.816
AUC: 0.919
AUC: 1.000
AUC: 1.000
AUC: 1.000
AUC: 1.000
```



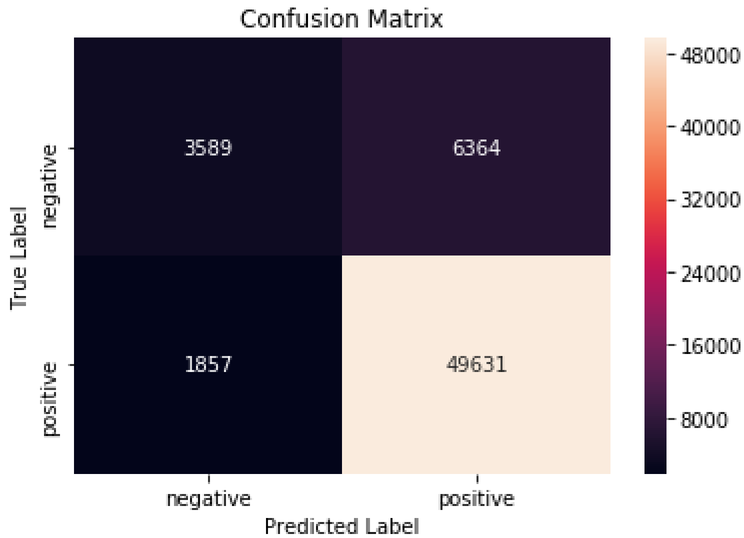
In [43]:
```
1 con_mat,con_mat_train,clf = dt_results(10,500,sent_vectors,test_sent_vectors,
```

```
The accuracy of the DT classifier for maxDepth = 10 and min split = 500 is 85.6
56236%
AUC: 0.826
AUC: 0.867
```



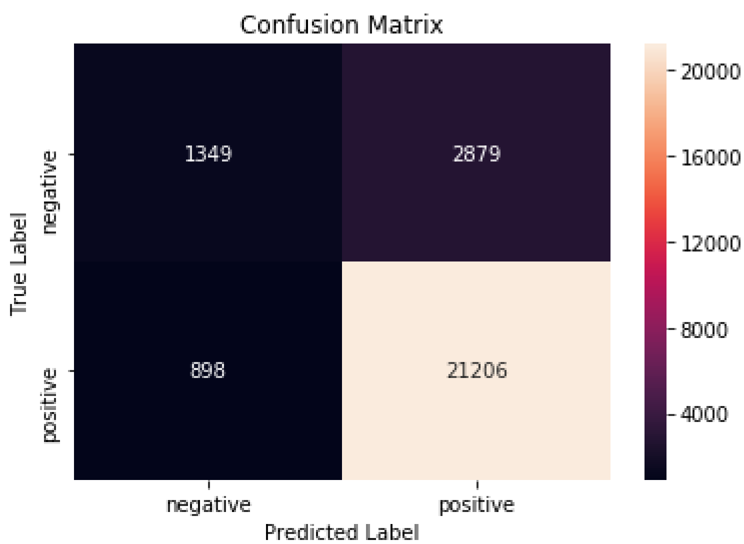
Observation: My model predicted with 85% accuracy with AUC: 0.832. Not much diffenece between Train ROC and Test ROC.

In [44]:
```
showHeatMap(con_mat_train)
```



Observation: My model predicted 1857 + 6364 points wrongly for train data set .

In [45]:
```
showHeatMap(con_mat)
```



Observation: My model predicted 898 + 2879 points wrongly

[5.4] Applying Decision Trees on TFIDF W2V, SET 4

In [29]:
```
X_train, X_test, y_1, y_test = cross_validation.train_test_split(preprocessed
```

In [30]:
```
model = TfidfVectorizer()
X_train_transformed = model.fit_transform(X_train)

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

In [31]:
```
# 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())
```

In [32]:
```
w2v_model=Word2Vec(list_of_sentance,min_count=5,size=50, workers=4)
w2v_words = list(w2v_model.wv.vocab)
```

In [33]:
```
# 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_sent_vectors = []; # the tfidf-w2v for each sentence/review is stored i
row=0;
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 = tf_idf_matrix[row, tfidf_feat.index(word)]
            # to reduce the computation we are
            # dictionary[word] = idf value of word in whole courpus
            # sent.count(word) = tf valeus of word in this review
            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)
    row += 1
```

```
100%|██████████| 61441/61441 [46:16<00:00, 32.22it/s]
```

In [34]:
```
i=0
list_of_test_sentance=[]
for sentance in X_test:
    list_of_test_sentance.append(sentance.split())
```

In [35]:

```
# 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_test_sent_vectors = []; # the tfidf-w2v for each sentence/review is sto
row=0;
for sent in tqdm(list_of_test_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 = tf_idf_matrix[row, tfidf_feat.index(word)]
            # to reduce the computation we are
            # dictionary[word] = idf value of word in whole courpus
            # sent.count(word) = tf valeus of word in this review
            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_test_sent_vectors.append(sent_vec)
    row += 1
```

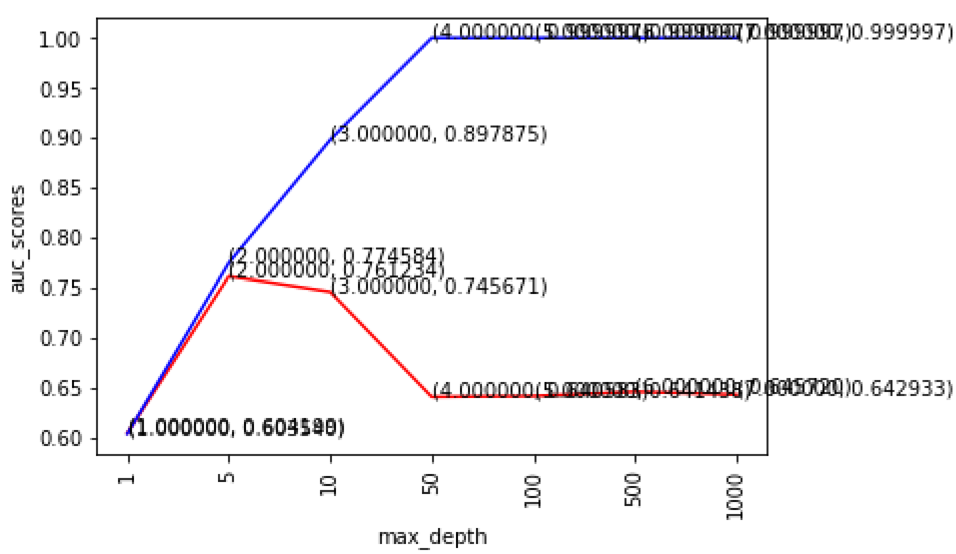
```
100%|██████████| 26332/26332 [1:03:08<00:00,  6.95it/s]
```

In [36]:

```
X_tr, X_cv, y_tr, y_cv = cross_validation.train_test_split(X_train, y_1, test

i=0
list_of_cv_sentance=[]
for sentance in X_tr:
    list_of_cv_sentance.append(sentance.split())


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_cv_sent_vectors = []; # the tfidf-w2v for each sentence/review is store
row=0;
for sent in tqdm(list_of_cv_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 = tf_idf_matrix[row, tfidf_feat.index(word)]
            # to reduce the computation we are
            # dictionary[word] = idf value of word in whole courpus
            # sent.count(word) = tf valeus of word in this review
            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_cv_sent_vectors.append(sent_vec)
    row += 1

i=0
list_of_cv_test_sentance=[]
for sentance in X_cv:
    list_of_cv_test_sentance.append(sentance.split())


tfidf_cv_test_sent_vectors = []; # the tfidf-w2v for each sentence/review is
row=0;
for sent in tqdm(list_of_cv_test_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 = tf_idf_matrix[row, tfidf_feat.index(word)]
            # to reduce the computation we are
            # dictionary[word] = idf value of word in whole courpus
            # sent.count(word) = tf valeus of word in this review
            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_cv_test_sent_vectors.append(sent_vec)
    row += 1
```

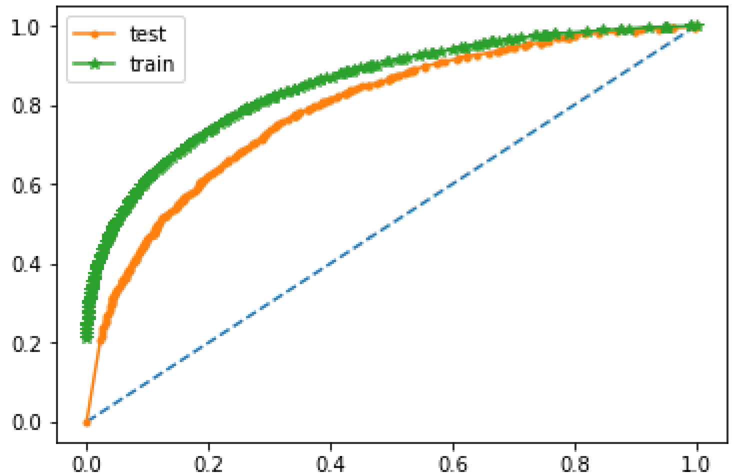
```


tuned_parameters = [{'max_depth': [1, 5, 10, 50, 100, 500, 1000],'min_samples

#Using GridSearchCV
model = GridSearchCV(DecisionTreeClassifier(), tuned_parameters, scoring = 'r
model.fit(tfidf_cv_sent_vectors, y_tr)

print(model.best_estimator_)
print(model.score(tfidf_cv_test_sent_vectors, y_cv))

check_trade_off(tfidf_cv_sent_vectors,tfidf_cv_test_sent_vectors,y_tr,y_cv)

```

```
100%|██████████| 43008/43008 [34:39<00:00, 20.68it/s]
100%|██████████| 18433/18433 [21:01<00:00, 14.62it/s]

DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=1000,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=500,
            min_weight_fraction_leaf=0.0, presort=False, random_state=None,
            splitter='best')
0.7829388087147336
AUC: 0.604
AUC: 0.761
AUC: 0.746
AUC: 0.641
AUC: 0.641
AUC: 0.646
AUC: 0.643
#######################################################
AUC from train data ###################################
AUC: 0.604
AUC: 0.775
AUC: 0.898
AUC: 1.000
AUC: 1.000
AUC: 1.000
AUC: 1.000
```



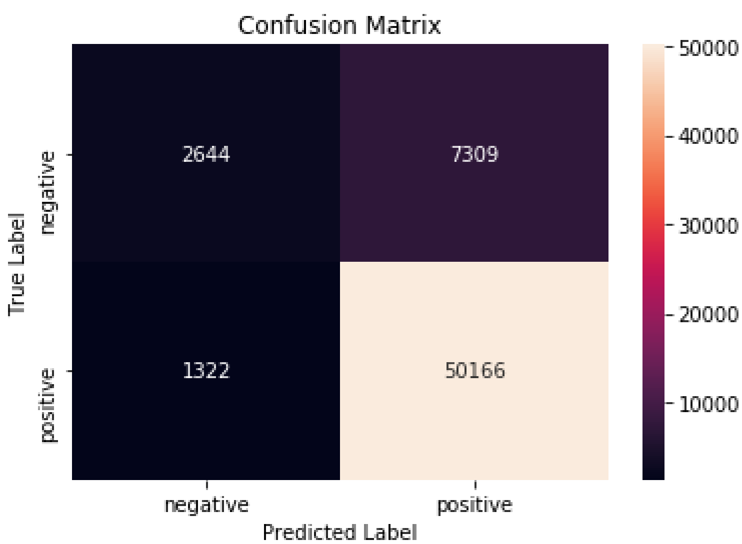
In [37]:
```
con_mat,con_mat_trai,clf = dt_results(1000,500,tfidf_sent_vectors,tfidf_test_
```

```
The accuracy of the DT classifier for maxDepth = 1000 and min split = 500 is 8
4.968859%
AUC: 0.791
AUC: 0.853
```



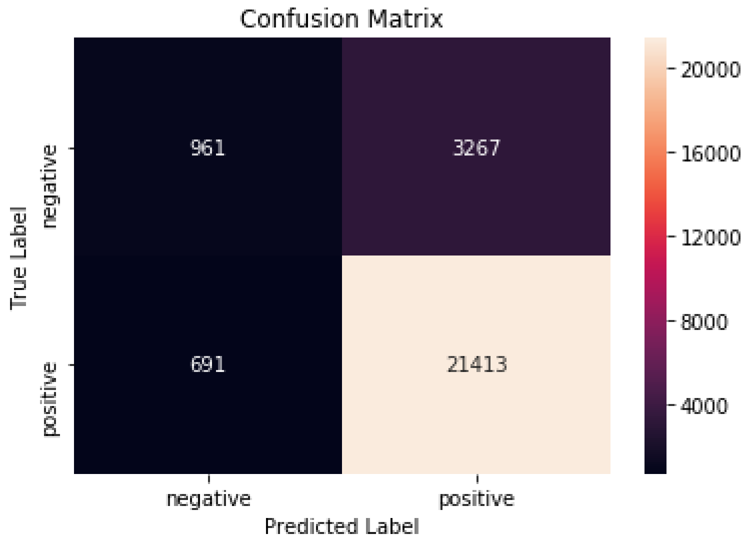
Observation: My model predicted with 84 % accuracy with AUC: 0.791

In [39]:
```
showHeatMap(con_mat_trai)
```



Observation : My model predicted 1322 + 7309 points wrongly for train data.

In [40]:
```
showHeatMap(con_mat)
```



Observation : My model predicted 691 + 3267 points wrongly

Repeat with extra features

In [41]:
```
mylen = np.vectorize(len)
newarr = mylen(preprocessed_summary)
```

In [42]:
```
newproce_reviews = np.asarray(preprocessed_reviews)
```

In [43]:
```
newproce_summary = np.asanyarray(preprocessed_summary)
```

In [44]:
```
df = pd.DataFrame({'desc':newproce_reviews, 'summary':newproce_summary,'len':
```

In [45]:
```
df.head()
```

Out[45]:

	desc	summary	len
0	dogs loves chicken product china wont buying a...	made china	10
1	dogs love saw pet store tag attached regarding...	dog lover delites	17
2	infestation fruitflies literally everywhere fl...	one fruitfly stuck	18
3	worst product gotten long time would rate no s...	not work not waste money	24
4	wish would read reviews making purchase basica...	big rip	7

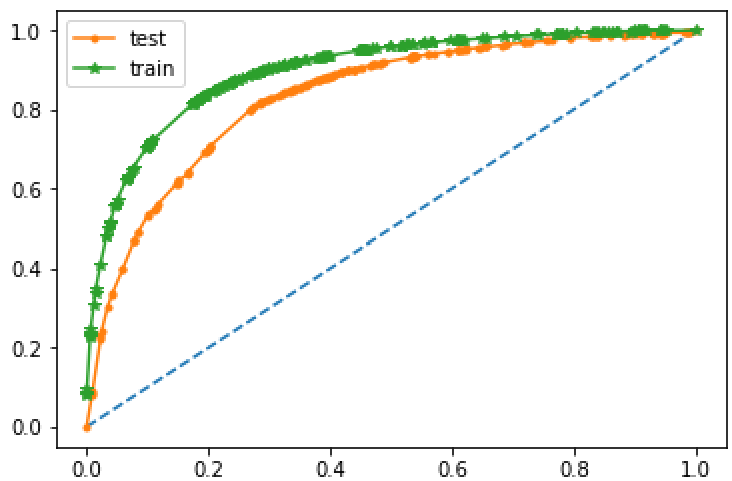
In [46]:
```
X_1, X_test, y_1, y_test = cross_validation.train_test_split(df, final['Score
```

In [47]:
```
import scipy
count_vect = CountVectorizer()
final_counts = count_vect.fit_transform(X_1['desc'])
final_test_count = count_vect.transform(X_test['desc'])

# split the train data set into cross validation train and cross validation t
X_tr, X_cv, y_tr, y_cv = cross_validation.train_test_split(X_1, y_1, test_siz

final_counts_tr_cv = count_vect.transform(X_tr['desc'])
final_test_count_cv = count_vect.transform(X_cv['desc'])

from scipy.sparse import csr_matrix, issparse

##################Adding len as feature##########################
#if issparse(final_counts_tr_cv):
    #print('sparse matrix')
len_sparse = scipy.sparse.coo_matrix(X_tr['len'])
len_sparse = len_sparse.transpose()

final_counts_tr_cv = scipy.sparse.hstack([final_counts_tr_cv, len_sparse])
print(final_counts_tr_cv.shape)

len_test_sparse = scipy.sparse.coo_matrix(X_cv['len'])
len_test_sparse = len_test_sparse.transpose()
final_test_count_cv = scipy.sparse.hstack([final_test_count_cv,len_test_spars
print("final_counts_tr_cv.shape after length = ",final_counts_tr_cv.shape)

##################Adding summary as feature##########################
final_summary_count = count_vect.transform(X_tr['summary'])
final_test_summary_count_cv = count_vect.transform(X_cv['summary'])
columns=count_vect.get_feature_names()

print("sujet",final_summary_count[:,12].shape)
final_counts_tr_cv = scipy.sparse.hstack([final_counts_tr_cv, final_summary_c
print("final_counts_tr_cv.shape after f1= ",final_counts_tr_cv.shape)

final_test_count_cv = scipy.sparse.hstack([final_test_count_cv,final_test_sum


final_counts_tr_cv = scipy.sparse.hstack([final_counts_tr_cv, final_summary_c
print("final_counts_tr_cv.shape after f2= ",final_counts_tr_cv.shape)


final_test_count_cv = scipy.sparse.hstack([final_test_count_cv,final_test_sum

##############finding the new C #########################################

tuned_parameters = [{'max_depth': [1, 5, 10, 50, 100, 500, 1000],'min_samples

#Using GridSearchCV
model = GridSearchCV(DecisionTreeClassifier(), tuned_parameters, scoring = 'r
model.fit(final_counts_tr_cv, y_tr)

print(model.best_estimator_)
print(model.score(final_test_count_cv, y_cv))

```

```
(43008, 46447)
final_counts_tr_cv.shape after length =  (43008, 46447)
sujet (43008, 1)
final_counts_tr_cv.shape after f1=  (43008, 46448)
final_counts_tr_cv.shape after f2=  (43008, 46449)
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=50,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=500,
            min_weight_fraction_leaf=0.0, presort=False, random_state=None,
            splitter='best')
0.8124306837028364
```

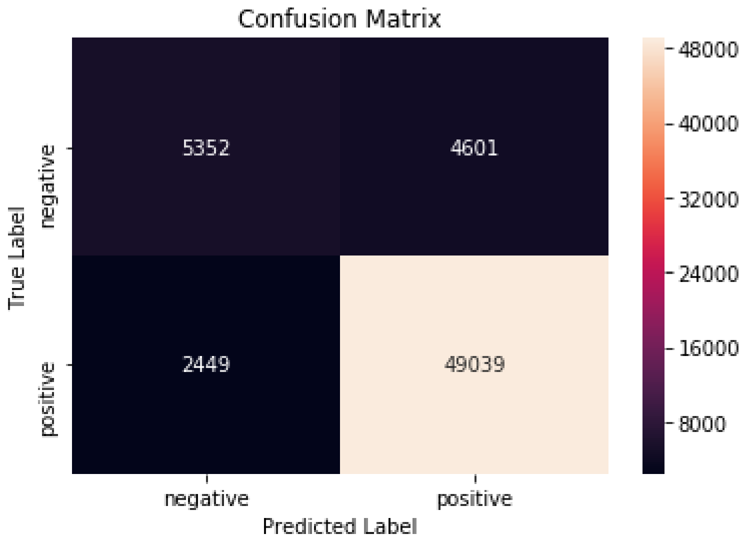
In [48]:
```
1 con_mat,con_mat_train,clf = dt_results(50,500,final_counts,final_test_count,y
```

```
The accuracy of the DT classifier for maxDepth = 50 and min split = 500 is 85.8
61309%
AUC: 0.833
AUC: 0.898
```



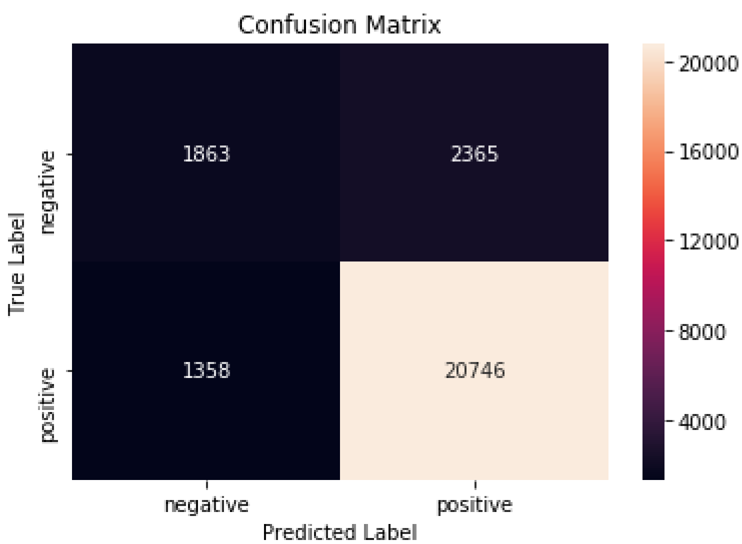
Observation: My model predicted with 85% accuracy with AUC: 0.833

In [49]:
```
showHeatMap(con_mat_train)
```



Observation: My model predicted 2449 + 4601 points wrongly even with train data

In [50]:
```
showHeatMap(con_mat)
```



Observation: My model predicted 1358 + 2365 points wrongly

[6] Conclusions

Method	No of samples	depth	split	accuracy	AUC Score
BOW	100000	50	500	85	0.833
TFIDF	100000	50	500	86	826
AVG W2VE	100000	10	500	85	0.832
TFIDF W2VE	100000	1000	500	85	0.817

Method	No of samples	depth	split	accuracy	AUC Score
BOW1	100000	50	500	85	0.833

In [ ]:
```
1
```