04 Amazon Fine Food Reviews Analysis_NaiveBayes

March 26, 2019

1 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.

2 [1]. Reading Data

2.1 [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 [2]:
```
%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
```

In [3]:
```
# 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 point
# 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 5
```

```
        # for tsne assignment you can take 5k data points

        filtered_data = pd.read_sql_query(""" SELECT * FROM Reviews WHERE Score != 3 LIMIT 150

        # Give reviews with Score>3 a positive rating(1), and reviews with a score<3 a negativ
        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 (150000, 10)
```

Out[3]:

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

	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	\
0	1	1	1	1303862400	
1	0	0	0	1346976000	
2	1	1	1	1219017600	

	Summary	Text
0	Good Quality Dog Food	I have bought several of the Vitality canned d...
1	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...
2	"Delight" says it all	This is a confection that has been around a fe...

In [4]:
```
display = pd.read_sql_query("""
SELECT UserId, ProductId, ProfileName, Time, Score, Text, COUNT(*)
FROM Reviews
GROUP BY UserId
HAVING COUNT(*)>1
""", con)
```

In [5]:
```
print(display.shape)
display.head()
```

```
(80668, 7)
```

Out[5]:

	UserId	ProductId	ProfileName	Time	Score	\
0	#oc-R115TNMSPFT9I7	B007Y59HVM	Breyton	1331510400	2	

```
1  #oc-R11D9D7SHXIJB9  B005HG9ET0  Louis E. Emory "hoppy"  1342396800      5
2  #oc-R11DNU2NBKQ23Z  B007Y59HVM        Kim Cieszykowski  1348531200      1
3  #oc-R11O5J5ZVQE25C  B005HG9ET0           Penguin Chick  1346889600      5
4  #oc-R12KPBODL2B5ZD  B007OSBE1U   Christopher P. Presta  1348617600      1

                                                  Text  COUNT(*)
0  Overall its just OK when considering the price...         2
1  My wife has recurring extreme muscle spasms, u...         3
2  This coffee is horrible and unfortunately not ...         2
3  This will be the bottle that you grab from the...         3
4  I didnt like this coffee. Instead of telling y...         2
```

```
In [6]: display[display['UserId']=='AZY10LLTJ71NX']
```

```
Out[6]:              UserId   ProductId                      ProfileName        Time  \
80638  AZY10LLTJ71NX  B006P7E5ZI  undertheshrine "undertheshrine"  1334707200

       Score                                                Text  COUNT(*)
80638      5  I was recommended to try green tea extract to ...         5
```

```
In [7]: display['COUNT(*)'].sum()
```

```
Out[7]: 393063
```

3 [2] Exploratory Data Analysis

3.1 [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 [8]: display= pd.read_sql_query("""
        SELECT *
        FROM Reviews
        WHERE Score != 3 AND UserId="AR5J8UI46CURR"
        ORDER BY ProductID
        """, con)
        display.head()
```

```
Out[8]:        Id   ProductId         UserId      ProfileName  HelpfulnessNumerator  \
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

   HelpfulnessDenominator  Score        Time  \
0                       2      5  1199577600
```

```
1                       2      5  1199577600
2                       2      5  1199577600
3                       2      5  1199577600
4                       2      5  1199577600

                              Summary  \
0  LOACKER QUADRATINI VANILLA WAFERS
1  LOACKER QUADRATINI VANILLA WAFERS
2  LOACKER QUADRATINI VANILLA WAFERS
3  LOACKER QUADRATINI VANILLA WAFERS
4  LOACKER QUADRATINI VANILLA WAFERS

                                                  Text
0  DELICIOUS WAFERS. I FIND THAT EUROPEAN WAFERS ...
1  DELICIOUS WAFERS. I FIND THAT EUROPEAN WAFERS ...
2  DELICIOUS WAFERS. I FIND THAT EUROPEAN WAFERS ...
3  DELICIOUS WAFERS. I FIND THAT EUROPEAN WAFERS ...
4  DELICIOUS WAFERS. I FIND THAT EUROPEAN WAFERS ...
```

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 [9]: #Sorting data according to ProductId in ascending order
        sorted_data=filtered_data.sort_values('ProductId', axis=0, ascending=True, inplace=Fals

In [10]: #Deduplication of entries
         final=sorted_data.drop_duplicates(subset={"UserId","ProfileName","Time","Text"}, keep=
         final.shape

Out[10]: (126359, 10)

In [11]: #Checking to see how much % of data still remains
         (final['Id'].size*1.0)/(filtered_data['Id'].size*1.0)*100

Out[11]: 84.23933333333333
```

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

         display.head()
```

Out[12]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
0	64422	B000MIDROQ	A161DK06JJMCYF	J. E. Stephens "Jeanne"	3	1	5	1224892800	Bought This for My Son at College	My son loves spaghetti so I didn't hesitate or...
1	44737	B001EQ55RW	A2V0I904FH7ABY	Ram	3	2	4	1212883200	Pure cocoa taste with crunchy almonds inside	It was almost a 'love at first bite' - the per...

```
In [13]: final=final[final.HelpfulnessNumerator<=final.HelpfulnessDenominator]

In [14]: #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()
```

```
(126357, 10)
```

```
Out[14]: 1    106326
         0     20031
         Name: Score, dtype: int64
```

4 [3] Preprocessing

4.1 [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 [15]: # 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)
```

```
I grew up reading these Sendak books, and watching the Really Rosie movie that incorporates the
==================================================
Its about time Spanish products started getting their due.. The most famous (rightly so) Spanis
==================================================
I love this stuff.  I nuke a mug a milk until it's very hot, drop in 2 of the triangles, stir u
==================================================
There's nothing like the scent of real lavender! Just a whiff smells so good. Besides enjoying
==================================================
```

```
In [16]: # 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)
```

```
I grew up reading these Sendak books, and watching the Really Rosie movie that incorporates the
```

```
In [17]: # https://stackoverflow.com/questions/16206380/python-beautifulsoup-how-to-remove-all
         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)
```

```
I grew up reading these Sendak books, and watching the Really Rosie movie that incorporates the
==================================================
Its about time Spanish products started getting their due.. The most famous (rightly so) Spanis
==================================================
I love this stuff.  I nuke a mug a milk until it's very hot, drop in 2 of the triangles, stir u
==================================================
There's nothing like the scent of real lavender! Just a whiff smells so good. Besides enjoying
```

In [18]:
```
# 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 [19]:
```
sent_1500 = decontracted(sent_1500)
```

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

```
I love this stuff.  I nuke a mug a milk until it is very hot, drop in 2 of the triangles, stir
==================================================
```

In [20]:
```
#remove words with numbers python: https://stackoverflow.com/a/18082370/4084039
sent_0 = re.sub("\S*\d\S*", "", sent_0).strip()
print(sent_0)
```

```
I grew up reading these Sendak books, and watching the Really Rosie movie that incorporates the
```

In [21]:
```
#remove spacial character: https://stackoverflow.com/a/5843547/4084039
sent_1500 = re.sub('[^A-Za-z0-9]+', ' ', sent_1500)
print(sent_1500)
```

```
I love this stuff I nuke a mug a milk until it is very hot drop in 2 of the triangles stir unti
```

In [22]:
```
# 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', 'ourselve
            "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him'
            'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself',
            'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "
            'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'h
            'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as
            'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'throug
            'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', '
            'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'a
            'most', 'other', 'some', 'such', 'only', 'own', 'same', 'so', 'than', 'to
            's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", '
            've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't
            "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mi
            "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't",
            'won', "won't", 'wouldn', "wouldn't"])
```

In [23]:
```
# 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 in stopw
        preprocessed_reviews.append(sentance.strip())
```

```
100%|| 126357/126357 [00:53<00:00, 2377.58it/s]
```

```
In [24]: preprocessed_reviews[1500]
```

```
Out[24]: 'love stuff nuke mug milk hot drop triangles stir chocolate melts froth aerolatte sim
```

5 [4] Featurization

Splitting the dataset into train, test and cv We are taking 100000 points in total , 33 % of which is kept as dtest

```
In [27]: from sklearn.model_selection import train_test_split
```

```
        # X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.33, shuffle=F
        X=np.asarray(preprocessed_reviews[:100000])
        Y=final['Score'].values
        #print(X.shape," ",Y.shape)
        X_train, X_test, y_train, y_test = train_test_split(X, Y[:100000], test_size=0.33) #
        X_train, X_cv, y_train, y_cv = train_test_split(X_train, y_train, test_size=0.33) # t

        #print(X.shape,Y.shape)
```

5.1 [4.1] BAG OF WORDS

```
In [26]: #BoW
        count_vect = CountVectorizer() #in scikit-learn
        count_vect.fit(X_train)
        print("some feature names ", count_vect.get_feature_names()[:10])
        print('='*50)

        final_countsXtrain = count_vect.transform(X_train)
        final_countsXtest = count_vect.transform(X_test)
        final_countsXcv = count_vect.transform(X_cv)

        print("the shape of out text BOW vectorizer xtrain ",final_countsXtrain.get_shape())
        print("the shape of out text BOW vectorizer xtest ",final_countsXtest.get_shape())
        print("the shape of out text BOW vectorizer xcv ",final_countsXcv.get_shape())
```

```
        #print("the number of unique words ", final_counts.get_shape()[1])
```

```
some feature names  ['aa', 'aaa', 'aaaa', 'aaaaaaahhhhhh', 'aaaaaawwwwwwwww', 'aaaah', 'aaaan
==================================================
the shape of out text BOW vectorizer xtrain  (44890, 40140)
the shape of out text BOW vectorizer xtest  (33000, 40140)
the shape of out text BOW vectorizer xcv  (22110, 40140)
```

bow vectors with added feature of review length

```
In [27]: from scipy.sparse import coo_matrix, hstack
         xtrain_len=[]
         xtest_len=[]
         xcv_len=[]
         for i in X_train:
             xtrain_len.append(len(i))

         for i in X_test:
             xtest_len.append(len(i))

         for i in X_cv:
             xcv_len.append(len(i))
         xtrain_len=coo_matrix(np.reshape(np.array(xtrain_len),(-1,1)))
         xtest_len=coo_matrix(np.reshape(np.array(xtest_len),(-1,1)))
         xcv_len=coo_matrix(np.reshape(np.array(xcv_len),(-1,1)))
         bowtrain=coo_matrix(final_countsXtrain)
         bowtest=coo_matrix(final_countsXtest)
         bowcv=coo_matrix(final_countsXcv)
         bowtrain=hstack([final_countsXtrain,xtrain_len])
         bowtest=hstack([final_countsXtest,xtest_len])
         bowcv=hstack([final_countsXcv,xcv_len])
         print(bowtrain.shape)
         print(bowtest.shape)
         print(bowcv.shape)
```

```
(44890, 40068)
(33000, 40068)
(22110, 40068)
```

5.2 [4.3] TF-IDF

```
In [28]: tf_idf_vect = TfidfVectorizer(ngram_range=(1,2), min_df=10)
         tf_idf_vect.fit(X_train)
         print("some sample features(unique words in the corpus)",tf_idf_vect.get_feature_names
         print('='*50)
```

```
        final_tf_idfXtrain = tf_idf_vect.transform(X_train)
        final_tf_idfXtest = tf_idf_vect.transform(X_test)
        final_tf_idfXcv = tf_idf_vect.transform(X_cv)
        #print("the type of count vectorizer ",type(final_tf_idf))
        print("the shape of out text TFIDF vectorizer xtrain ",final_tf_idfXtrain.get_shape()
        print("the shape of out text TFIDF vectorizer xtest ",final_tf_idfXtest.get_shape())
        print("the shape of out text TFIDF vectorizer xcv ",final_tf_idfXcv.get_shape())
        #print("the number of unique words including both unigrams and bigrams ", final_tf_id
```

```
some sample features(unique words in the corpus) ['aa', 'abdominal', 'ability', 'able', 'able b
==================================================
the shape of out text TFIDF vectorizer xtrain  (44890, 26171)
the shape of out text TFIDF vectorizer xtest  (33000, 26171)
the shape of out text TFIDF vectorizer xcv  (22110, 26171)
```

tfidf vectors with added feature of review length

```
In [29]: tfidftrain=coo_matrix(final_tf_idfXtrain)
         tfidftest=coo_matrix(final_tf_idfXtest)
         tfidfcv=coo_matrix(final_tf_idfXcv)
         tfidftrain=hstack([final_tf_idfXtrain,xtrain_len])
         tfidftest=hstack([final_tf_idfXtest,xtest_len])
         tfidfcv=hstack([final_tf_idfXcv,xcv_len])
         print(tfidftrain.shape)
         print(tfidftest.shape)
         print(tfidfcv.shape)
```

```
(44890, 26172)
(33000, 26172)
(22110, 26172)
```

6 [5] Assignment 4: Apply Naive Bayes

```
<li><strong>Apply Multinomial NaiveBayes on these feature sets</strong>
    <ul>
        <li><font color='red'>SET 1:</font>Review text, preprocessed one converted into vectors
        <li><font color='red'>SET 2:</font>Review text, preprocessed one converted into vectors
    </ul>
</li>
<br>
<li><strong>The hyper paramter tuning(find best Alpha)</strong>
    <ul>
<li>Find the best hyper parameter which will give the maximum <a href='https://www.appliedaicou
<li>Consider a wide range of alpha values for hyperparameter tuning, start as low as 0.00001</l
<li>Find the best hyper paramter using k-fold cross validation or simple cross validation data
<li>Use gridsearch cv or randomsearch cv or you can also write your own for loops to do this ta
```

```
    </ul>
</li>
<br>
<li><strong>Feature importance</strong>
    <ul>
<li>Find the top 10 features of positive class and top 10 features of negative class for both
    </ul>
</li>
<br>
<li><strong>Feature engineering</strong>
    <ul>
<li>To increase the performance of your model, you can also experiment with with feature engine
        <ul>
        <li>Taking length of reviews as another feature.</li>
        <li>Considering some features from review summary as well.</li>
    </ul>
    </ul>
</li>
<br>
<li><strong>Representation of results</strong>
    <ul>
<li>You need to plot the performance of model both on train data and cross validation data for
<img src='train_cv_auc.JPG' width=300px></li>
<li>Once after you found the best hyper parameter, you need to train your model with it, and fi
<img src='train_test_auc.JPG' width=300px></li>
<li>Along with plotting ROC curve, you need to print the <a href='https://www.appliedaicourse.c
<img src='confusion_matrix.png' width=300px></li>
    </ul>
</li>
<br>
<li><strong>Conclusion</strong>
    <ul>
<li>You need to summarize the results at the end of the notebook, summarize it in the table for
    <img src='summary.JPG' width=400px>
</li>
    </ul>
```

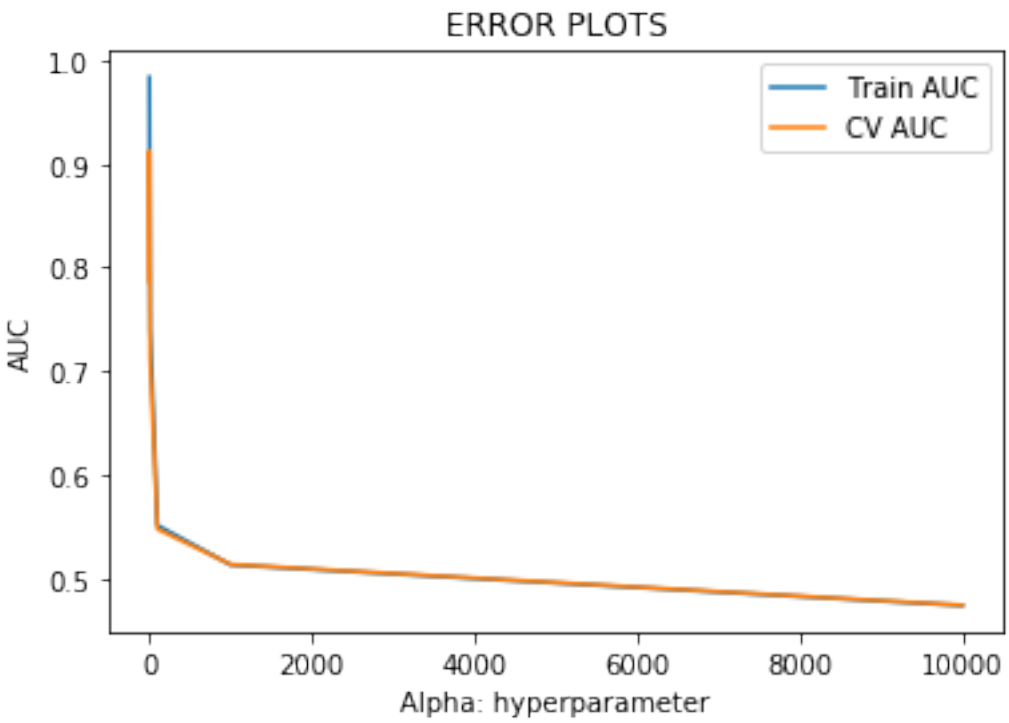
Note: Data Leakage

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.

7 Applying Multinomial Naive Bayes

7.1 [5.1] Applying Naive Bayes on BOW, SET 1

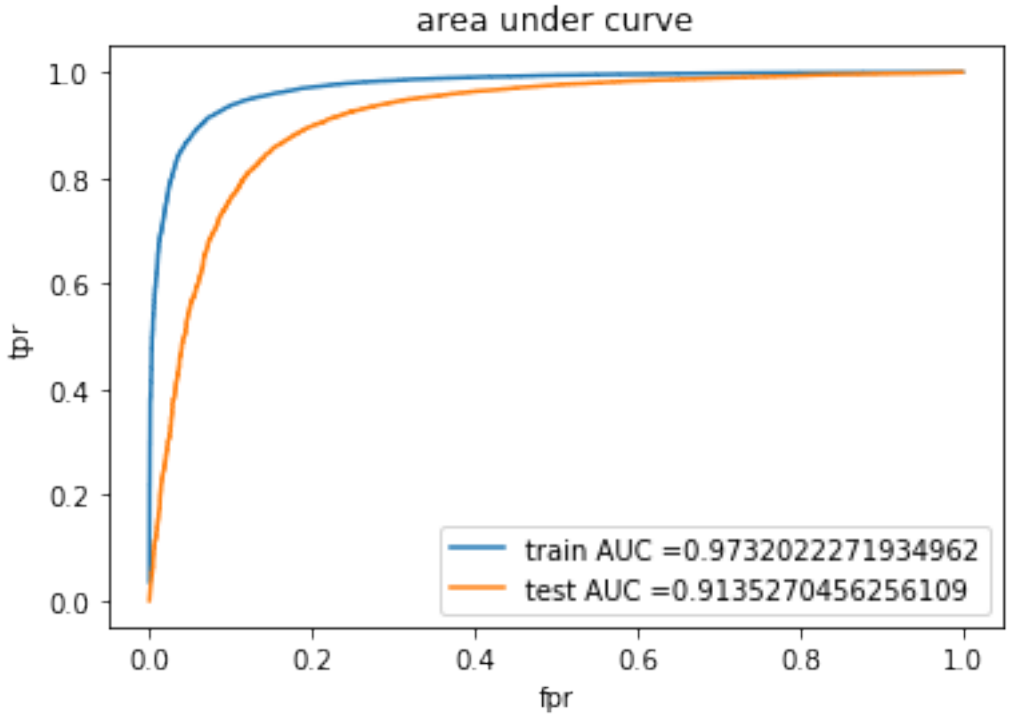
In [30]:
```
# Please write all the code with proper documentation
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import roc_auc_score
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler

x_train=final_countsXtrain
x_test=final_countsXtest
x_cv=final_countsXcv
"""
y_true : array, shape = [n_samples] or [n_samples, n_classes]
True binary labels or binary label indicators.

y_score : array, shape = [n_samples] or [n_samples, n_classes]
Target scores, can either be probability estimates of the positive class, confidence
decisions (as returned by decision_function on some classifiers).
For binary y_true, y_score is supposed to be the score of the class with greater labe

"""
#print(final_countsXtrain.toarray().shape)
train_auc = []
cv_auc = []
A = [0.000001,0.00001,0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000,10000]
for i in A:
    model = MultinomialNB(alpha=i)
    model.fit(x_train, y_train)
    # roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimate
    # not the predicted outputs
    y_train_pred =  model.predict_proba(x_train)[:,1]
    y_cv_pred =  model.predict_proba(x_cv)[:,1]

    train_auc.append(roc_auc_score(y_train,y_train_pred))
    cv_auc.append(roc_auc_score(y_cv, y_cv_pred))

plt.plot(A, train_auc, label='Train AUC')
plt.plot(A, cv_auc, label='CV AUC')
plt.legend()
plt.xlabel("Alpha: hyperparameter")
plt.ylabel("AUC")
plt.title("ERROR PLOTS")
plt.show()
```



In [31]:
```
from sklearn.metrics import roc_curve, auc

best_a = A[cv_auc.index(max(cv_auc))]
print(best_a)
model = MultinomialNB(alpha=best_a)
model.fit(x_train, y_train)
# roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates of
# not the predicted outputs

train_fpr, train_tpr, thresholds = roc_curve(y_train, model.predict_proba(x_train)[:,
test_fpr, test_tpr, thresholds = roc_curve(y_test, model.predict_proba(x_test)[:,1])

plt.plot(train_fpr, train_tpr, label="train AUC ="+str(auc(train_fpr, train_tpr)))
plt.plot(test_fpr, test_tpr, label="test AUC ="+str(auc(test_fpr, test_tpr)))
plt.legend()
plt.xlabel("fpr")
plt.ylabel("tpr")
plt.title("area under curve")
plt.show()

print("="*100)

from sklearn.metrics import confusion_matrix
print("Train confusion matrix")
```

```
print(confusion_matrix(y_train, model.predict(x_train)))
print("Test confusion matrix")
print(confusion_matrix(y_test, model.predict(x_test)))
```

0.1




```
==================================================================================================
Train confusion matrix
[[ 5820  1195]
 [ 1350 36525]]
Test confusion matrix
[[ 3390  1644]
 [ 1403 26563]]
```

7.1.1 [5.1.1] Top 10 important features of positive class from SET 1

```
In [32]: # Please write all the code with proper documentation
         fi=np.array(model.feature_log_prob_)
         print(fi.shape)
         top=np.array(sorted(fi[1,:]))
         toppos=top[len(top)-10:]
         top_pos_index=[]
```

```
        top_pos_feat=[]
        for i in toppos:
            top_pos_feat.append(count_vect.get_feature_names()[fi[1,:].tolist().index(i)])
            top_pos_index.append(fi[1,:].tolist().index(i))
        print(top_pos_index)
        print("Top 10 important features of positive class are")
        print(top_pos_feat)
```

```
(2, 40067)
[13244, 27448, 20587, 35042, 34936, 24292, 15333, 15002, 20115, 23640]
Top 10 important features of positive class are
['flavor', 'product', 'love', 'tea', 'taste', 'one', 'great', 'good', 'like', 'not']
```

[5.1.2] Top 10 important features of negative class from SET 1

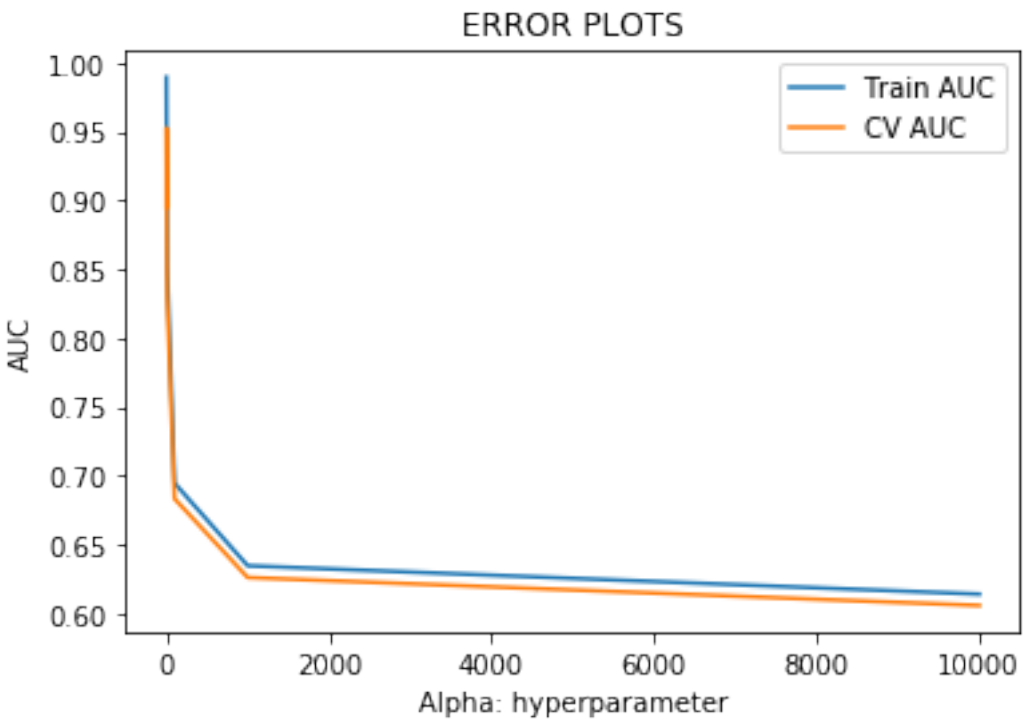
```
In [33]: # Please write all the code with proper documentation
        fi=np.array(model.feature_log_prob_)
        print(fi.shape)
        top=np.array(sorted(fi[0,:]))
        topneg=top[len(top)-10:]
        top_neg_index=[]
        top_neg_feat=[]
        for i in topneg:
            top_neg_feat.append(count_vect.get_feature_names()[fi[0,:].tolist().index(i)])
            top_neg_index.append(fi[0,:].tolist().index(i))
        print(top_neg_index)
        print("Top 10 important features of negative class are")

        print(top_neg_feat)
```

```
(2, 40067)
[13244, 13547, 23481, 15002, 24292, 34936, 27448, 39462, 20115, 23640]
Top 10 important features of negative class are
['flavor', 'food', 'no', 'good', 'one', 'taste', 'product', 'would', 'like', 'not']
```

7.2 [5.2] Applying Naive Bayes on TFIDF, SET 2

```
In [34]: # Please write all the code with proper documentation
        from sklearn.neighbors import KNeighborsClassifier
        from sklearn.metrics import roc_auc_score
        import matplotlib.pyplot as plt
        from sklearn.preprocessing import StandardScaler

        x_train=final_tf_idfXtrain
        x_test=final_tf_idfXtest
        x_cv=final_tf_idfXcv
        """
```

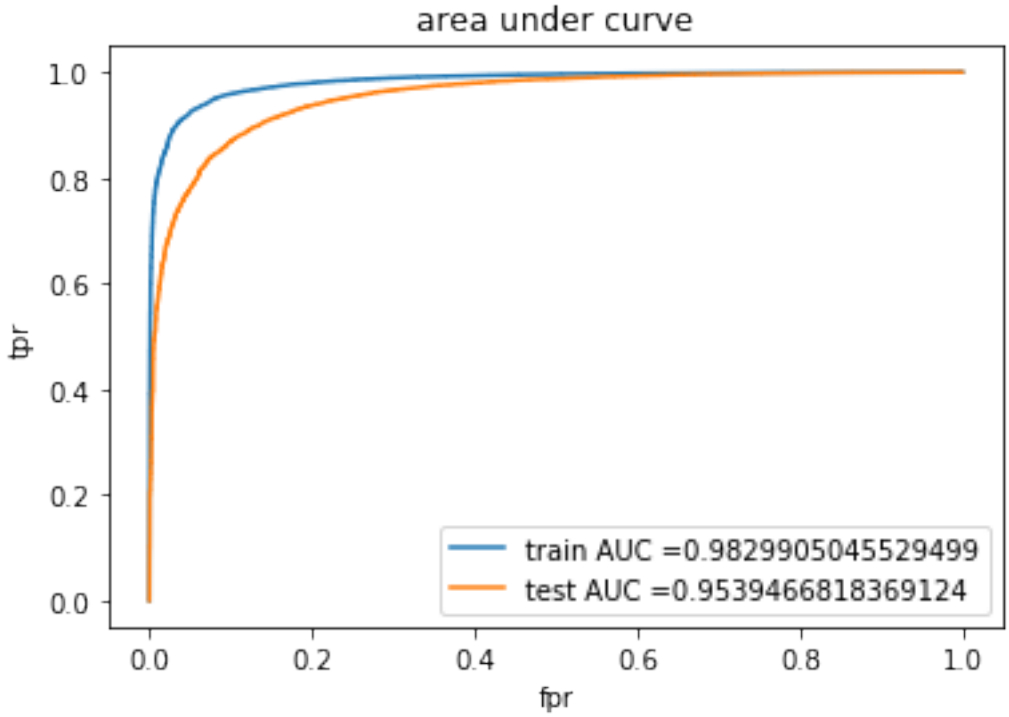
```
y_true : array, shape = [n_samples] or [n_samples, n_classes]
True binary labels or binary label indicators.

y_score : array, shape = [n_samples] or [n_samples, n_classes]
Target scores, can either be probability estimates of the positive class, confidence
decisions (as returned by decision_function on some classifiers).
For binary y_true, y_score is supposed to be the score of the class with greater labe

"""
#print(final_countsXtrain.toarray().shape)
train_auc = []
cv_auc = []
A = [0.000001,0.00001,0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000,10000]
for i in A:
    model = MultinomialNB(alpha=i)
    model.fit(x_train, y_train)
    # roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimate
    # not the predicted outputs
    y_train_pred =  model.predict_proba(x_train)[:,1]
    y_cv_pred =  model.predict_proba(x_cv)[:,1]

    train_auc.append(roc_auc_score(y_train,y_train_pred))
    cv_auc.append(roc_auc_score(y_cv, y_cv_pred))

plt.plot(A, train_auc, label='Train AUC')
plt.plot(A, cv_auc, label='CV AUC')
plt.legend()
plt.xlabel("Alpha: hyperparameter")
plt.ylabel("AUC")
plt.title("ERROR PLOTS")
plt.show()
```



In [35]:
```
from sklearn.metrics import roc_curve, auc

best_a = A[cv_auc.index(max(cv_auc))]
print(best_a)
model = MultinomialNB(alpha=best_a)
model.fit(x_train, y_train)
# roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates of
# not the predicted outputs

train_fpr, train_tpr, thresholds = roc_curve(y_train, model.predict_proba(x_train)[:,
test_fpr, test_tpr, thresholds = roc_curve(y_test, model.predict_proba(x_test)[:,1])

plt.plot(train_fpr, train_tpr, label="train AUC ="+str(auc(train_fpr, train_tpr)))
plt.plot(test_fpr, test_tpr, label="test AUC ="+str(auc(test_fpr, test_tpr)))
plt.legend()
plt.xlabel("fpr")
plt.ylabel("tpr")
plt.title("area under curve")
plt.show()

print("="*100)

from sklearn.metrics import confusion_matrix
print("Train confusion matrix")
```

```
print(confusion_matrix(y_train, model.predict(x_train)))
print("Test confusion matrix")
print(confusion_matrix(y_test, model.predict(x_test)))
```

```
0.1
```




```
==========================================================================================
Train confusion matrix
[[ 4453  2562]
 [  291 37584]]
Test confusion matrix
[[ 2311  2723]
 [  282 27684]]
```

7.2.1 [5.2.1] Top 10 important features of positive class from SET 2

```
In [36]: # Please write all the code with proper documentation
         fi=np.array(model.feature_log_prob_)
         print(fi.shape)
         top=np.array(sorted(fi[1,:]))
         toppos=top[len(top)-10:]
         top_pos_index=[]
```

```
        top_pos_feat=[]
        for i in toppos:
            top_pos_feat.append(tf_idf_vect.get_feature_names()[fi[1,:].tolist().index(i)])
            top_pos_index.append(fi[1,:].tolist().index(i))
        print(top_pos_index)
        print("Top 10 important features of positive class are")
        print(top_pos_feat)
```

```
(2, 26171)
[22558, 16171, 18032, 4009, 13098, 22886, 12296, 9526, 9957, 15296]
Top 10 important features of positive class are
['taste', 'one', 'product', 'coffee', 'love', 'tea', 'like', 'good', 'great', 'not']
```

7.2.2 [5.2.2] Top 10 important features of negative class from SET 2

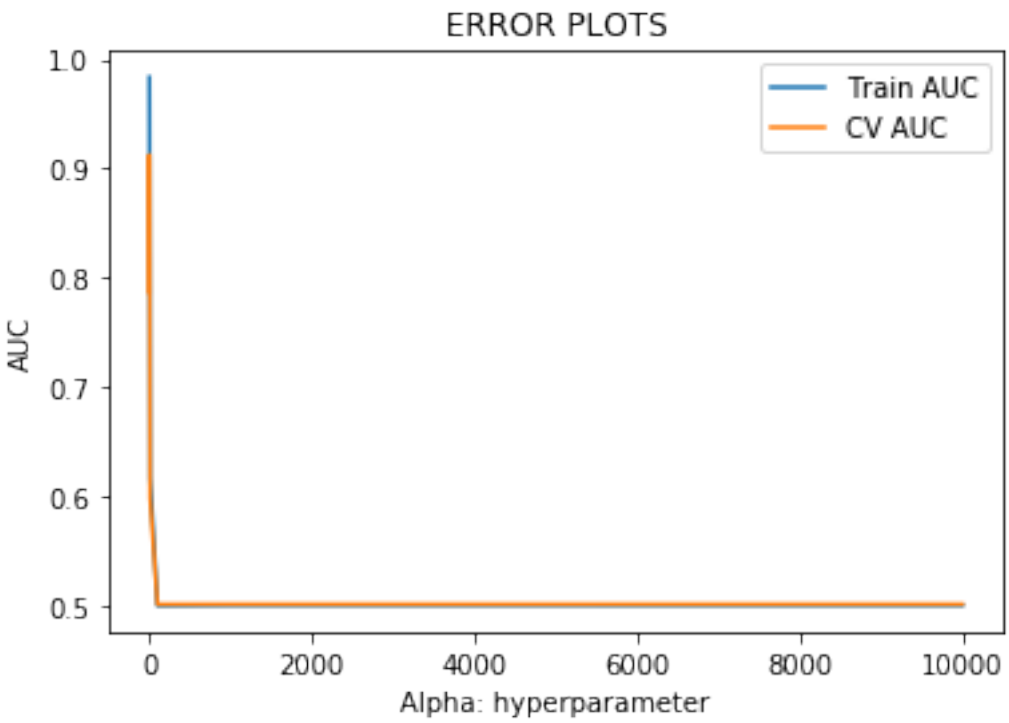
```
In [37]: # Please write all the code with proper documentation
        fi=np.array(model.feature_log_prob_)
        print(fi.shape)
        top=np.array(sorted(fi[0,:]))
        topneg=top[len(top)-10:]
        top_neg_index=[]
        top_neg_feat=[]
        for i in topneg:
            top_neg_feat.append(tf_idf_vect.get_feature_names()[fi[0,:].tolist().index(i)])
            top_neg_index.append(fi[0,:].tolist().index(i))
        print(top_neg_index)
        print("Top 10 important features of negitive class are")
        print(top_neg_feat)
```

```
(2, 26171)
[9526, 8258, 4009, 15099, 16171, 22558, 25822, 18032, 12296, 15296]
Top 10 important features of negitive class are
['good', 'food', 'coffee', 'no', 'one', 'taste', 'would', 'product', 'like', 'not']
```

Applying Naive Bayes on BOW after adding feature

```
In [39]: from sklearn.naive_bayes import MultinomialNB
        from sklearn.metrics import roc_auc_score
        import matplotlib.pyplot as plt
        from sklearn.preprocessing import StandardScaler

        x_train=bowtrain
        x_test=bowtest
        x_cv=bowcv
        """
        y_true : array, shape = [n_samples] or [n_samples, n_classes]
        True binary labels or binary label indicators.
```

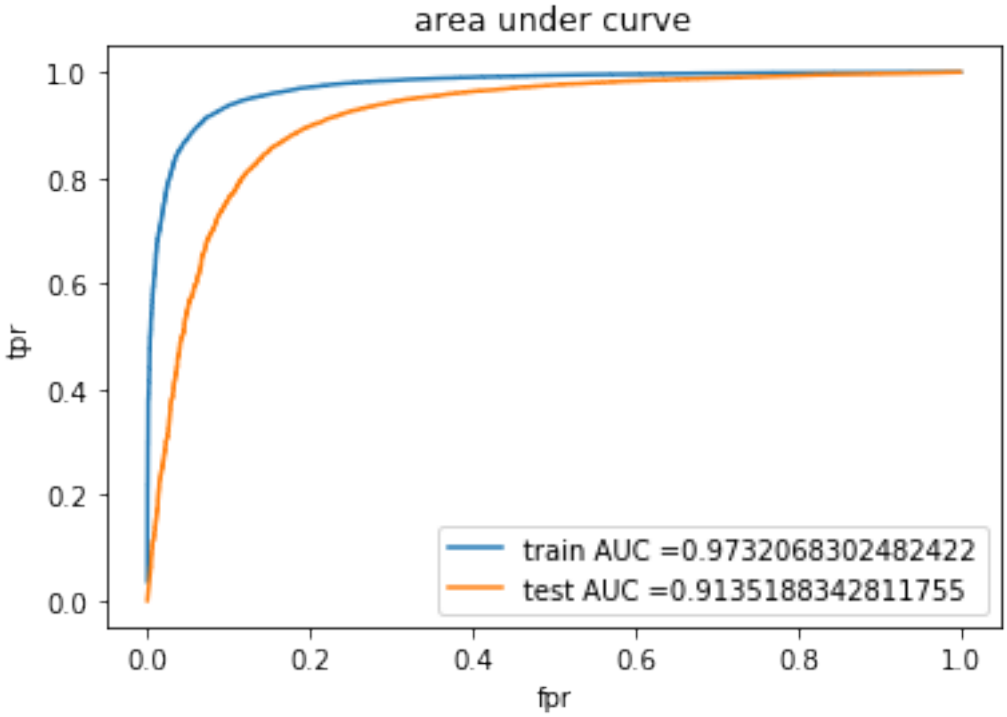
```
y_score : array, shape = [n_samples] or [n_samples, n_classes]
Target scores, can either be probability estimates of the positive class, confidence
decisions (as returned by decision_function on some classifiers).
For binary y_true, y_score is supposed to be the score of the class with greater labe

"""
#print(final_countsXtrain.toarray().shape)
train_auc = []
cv_auc = []
A = [0.000001,0.00001,0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000,10000]
for i in A:
    model = MultinomialNB(alpha=i)
    model.fit(x_train, y_train)
    # roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimate
    # not the predicted outputs
    y_train_pred =  model.predict_proba(x_train)[:,1]
    y_cv_pred =  model.predict_proba(x_cv)[:,1]

    train_auc.append(roc_auc_score(y_train,y_train_pred))
    cv_auc.append(roc_auc_score(y_cv, y_cv_pred))

plt.plot(A, train_auc, label='Train AUC')
plt.plot(A, cv_auc, label='CV AUC')
plt.legend()
plt.xlabel("Alpha: hyperparameter")
plt.ylabel("AUC")
plt.title("ERROR PLOTS")
plt.show()
```



In [40]:
```
from sklearn.metrics import roc_curve, auc

best_a = A[cv_auc.index(max(cv_auc))]
print(best_a)
model = MultinomialNB(alpha=best_a)
model.fit(x_train, y_train)
# roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates of
# not the predicted outputs

train_fpr, train_tpr, thresholds = roc_curve(y_train, model.predict_proba(x_train)[:,
test_fpr, test_tpr, thresholds = roc_curve(y_test, model.predict_proba(x_test)[:,1])

plt.plot(train_fpr, train_tpr, label="train AUC ="+str(auc(train_fpr, train_tpr)))
plt.plot(test_fpr, test_tpr, label="test AUC ="+str(auc(test_fpr, test_tpr)))
plt.legend()
plt.xlabel("fpr")
plt.ylabel("tpr")
plt.title("area under curve")
plt.show()

print("="*100)

from sklearn.metrics import confusion_matrix
print("Train confusion matrix")
```

```
        print(confusion_matrix(y_train, model.predict(x_train)))
        print("Test confusion matrix")
        print(confusion_matrix(y_test, model.predict(x_test)))
```

```
0.1
```

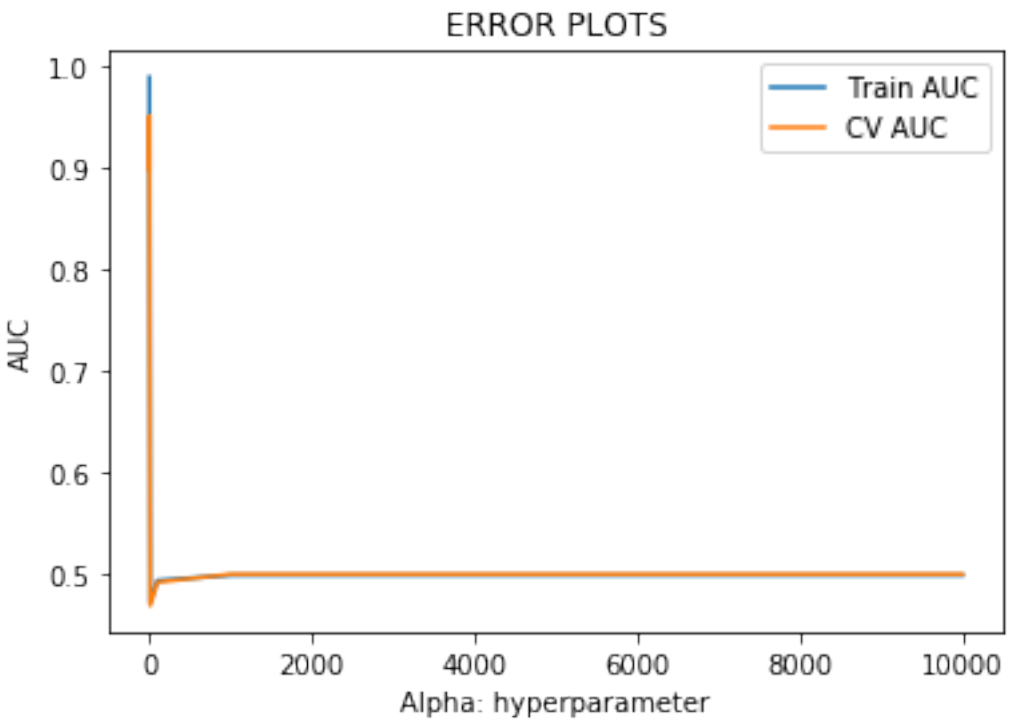



```
==========================================================================================
Train confusion matrix
[[ 5821  1194]
 [ 1351 36524]]
Test confusion matrix
[[ 3389  1645]
 [ 1401 26565]]
```

Applying Naive Bayes on TFIDF after adding feature

```
In [41]: from sklearn.neighbors import KNeighborsClassifier
         from sklearn.metrics import roc_auc_score
         import matplotlib.pyplot as plt
         from sklearn.preprocessing import StandardScaler

         x_train=tfidftrain
```

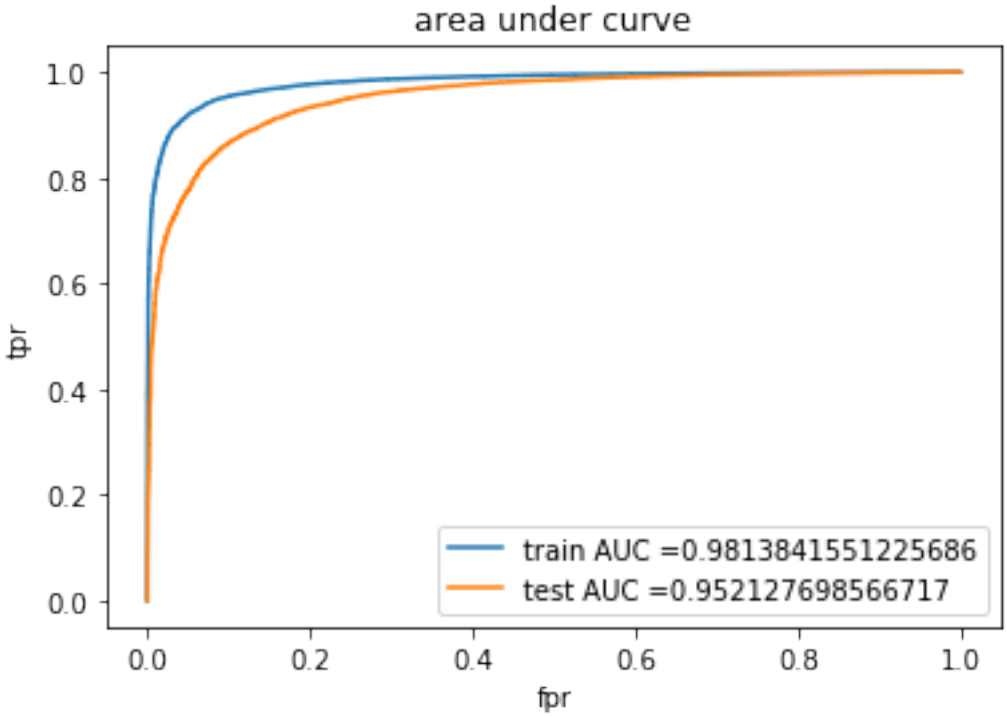
```
x_test=tfidftest
x_cv=tfidfcv
"""
y_true : array, shape = [n_samples] or [n_samples, n_classes]
True binary labels or binary label indicators.

y_score : array, shape = [n_samples] or [n_samples, n_classes]
Target scores, can either be probability estimates of the positive class, confidence
decisions (as returned by decision_function on some classifiers).
For binary y_true, y_score is supposed to be the score of the class with greater labe

"""
#print(final_countsXtrain.toarray().shape)
train_auc = []
cv_auc = []
A = [0.000001,0.00001,0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000,10000]
for i in A:
    model = MultinomialNB(alpha=i)
    model.fit(x_train, y_train)
    # roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimate
    # not the predicted outputs
    y_train_pred =  model.predict_proba(x_train)[:,1]
    y_cv_pred =  model.predict_proba(x_cv)[:,1]

    train_auc.append(roc_auc_score(y_train,y_train_pred))
    cv_auc.append(roc_auc_score(y_cv, y_cv_pred))

plt.plot(A, train_auc, label='Train AUC')
plt.plot(A, cv_auc, label='CV AUC')
plt.legend()
plt.xlabel("Alpha: hyperparameter")
plt.ylabel("AUC")
plt.title("ERROR PLOTS")
plt.show()
```



In [42]:
```
from sklearn.metrics import roc_curve, auc

best_a = A[cv_auc.index(max(cv_auc))]
print(best_a)
model = MultinomialNB(alpha=best_a)
model.fit(x_train, y_train)
# roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates of
# not the predicted outputs

train_fpr, train_tpr, thresholds = roc_curve(y_train, model.predict_proba(x_train)[:,
test_fpr, test_tpr, thresholds = roc_curve(y_test, model.predict_proba(x_test)[:,1])

plt.plot(train_fpr, train_tpr, label="train AUC ="+str(auc(train_fpr, train_tpr)))
plt.plot(test_fpr, test_tpr, label="test AUC ="+str(auc(test_fpr, test_tpr)))
plt.legend()
plt.xlabel("fpr")
plt.ylabel("tpr")
plt.title("area under curve")
plt.show()

print("="*100)

from sklearn.metrics import confusion_matrix
print("Train confusion matrix")
```

```
print(confusion_matrix(y_train, model.predict(x_train)))
print("Test confusion matrix")
print(confusion_matrix(y_test, model.predict(x_test)))
```

```
0.1
```




```
==========================================================================================
Train confusion matrix
[[ 4455  2560]
 [  371 37504]]
Test confusion matrix
[[ 2353  2681]
 [  345 27621]]
```

8 [6] Conclusions

```
In [44]: from prettytable import PrettyTable

         x = PrettyTable()

         x.field_names = ["Vectorizer", "Model", "Hyperparameter -alpha ", "AUC"]
```

```
x.add_row(["BOW", "Multinomial NaiveBayes", 0.1, 0.909])
x.add_row(["TFIDF", "Multinomial NaiveBayes", 0.1, 0.952])
x.add_row(["BOW with feature eng ", "Multinomial NaiveBayes", 0.1, 0.911])
x.add_row(["TFIDF with feature eng  ","Multinomial NaiveBayes", 0.1, 0.950])

print(x)
```

```
+--------------------------+-----------------------+-------------------------+-------+
|        Vectorizer        |         Model         | Hyperparameter -alpha   |  AUC  |
+--------------------------+-----------------------+-------------------------+-------+
|           BOW            | Multinomial NaiveBayes |          0.1           | 0.909 |
|          TFIDF           | Multinomial NaiveBayes |          0.1           | 0.952 |
|  BOW with feature eng    | Multinomial NaiveBayes |          0.1           | 0.911 |
| TFIDF with feature eng   | Multinomial NaiveBayes |          0.1           |  0.95 |
+--------------------------+-----------------------+-------------------------+-------+
```

In [ ]: