08affrdt

July 25, 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 [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

from sklearn.model_selection import train_test_split
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import roc_auc_score
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.metrics import precision_score
from sklearn.metrics import f1_score
from sklearn.metrics import recall_score
from prettytable import PrettyTable
```

```
from sklearn.tree import DecisionTreeClassifier
from sklearn import tree
import pydotplus
from IPython.display import Image
from IPython.display import SVG
from graphviz import Source
from IPython.display import display
```

```
C:\Users\ACER\Anaconda3\lib\site-packages\gensim\utils.py:860: UserWarning: detected Windows; a
  warnings.warn("detected Windows; aliasing chunkize to chunkize_serial")
```

In [2]:
```
# 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 100

# 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 (100000, 10)
```

Out[2]:
```
   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 [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  \
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 [5]:
```
display[display['UserId']=='AZY10LLTJ71NX']
```

Out[5]:
```
              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 [6]:
```
display['COUNT(*)'].sum()
```

Out[6]:
```
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 [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  \
        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 [8]:
```
#Sorting data according to ProductId in ascending order
sorted_data=filtered_data.sort_values('ProductId', axis=0, ascending=True, inplace=Fals
```

In [9]:
```
#Deduplication of entries
final=sorted_data.drop_duplicates(subset={"UserId","ProfileName","Time","Text"}, keep=
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  \
0  64422  B000MIDROQ  A161DK06JJMCYF  J. E. Stephens "Jeanne"
1  44737  B001EQ55RW  A2V0I904FH7ABY                      Ram

   HelpfulnessNumerator  HelpfulnessDenominator  Score        Time  \
0                     3                       1      5  1224892800
1                     3                       2      4  1212883200

                                       Summary  \
0            Bought This for My Son at College
1  Pure cocoa taste with crunchy almonds inside

                                               Text
0  My son loves spaghetti so I didn't hesitate or...
1  It was almost a 'love at first bite' - the per...
```

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

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

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

```
(87773, 10)
```

Out[13]:
```
1    73592
0    14181
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 [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 it anymore.  Its
==================================================
The Candy Blocks were a nice visual for the Lego Birthday party but the candy has little taste
==================================================
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 if you are afraid
==================================================
```

In [15]:
```python
# 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 it anymore.  Its
```

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

```
My dogs loves this chicken but its a product from China, so we wont be buying it anymore.  Its
==================================================
The Candy Blocks were a nice visual for the Lego Birthday party but the candy has little taste
==================================================
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 if you are afraic
```

```
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/18082370/4084039
         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 it anymore.  Its
```

```
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 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', '
            'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'a
            '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 [22]: # Combining all the above stundents
         from tqdm import tqdm
         preprocessed_reviews_dt = []
         # 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_dt.append(sentance.strip())
```

```
100%|| 87773/87773 [01:04<00:00, 1353.97it/s]
```

```
In [23]: preprocessed_reviews_dt[1500]
```

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

4.2 [4] Splitting the data

```
In [24]: X = preprocessed_reviews_dt
         Y = final['Score'].values
```

```
In [25]: # Here we are splitting the data(X ,Y) into train and test data
         # X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.33, shuffle=F
         X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.30) # this is r
```

5 [4] Featurization

5.1 [4.1] BAG OF WORDS

```
In [26]: #BoW
         vectorizer = CountVectorizer(min_df = 10)
         vectorizer.fit(X_train) # fit has to happen only on train data
         print(vectorizer.get_feature_names()[:20])# printing some feature names
         print("="*50)

         # we use the fitted CountVectorizer to convert the text to vector
         X_train_bow = vectorizer.transform(X_train)
         X_test_bow = vectorizer.transform(X_test)

         print("After vectorizations")
         print(X_train_bow.shape, Y_train.shape)
         print(X_test_bow.shape, Y_test.shape)
```

```
['aa', 'abandoned', 'abdominal', 'ability', 'able', 'abroad', 'absence', 'absent', 'absolute',
==================================================
After vectorizations
(61441, 9615) (61441,)
(26332, 9615) (26332,)
```

5.2 [4.3] TF-IDF

```
In [27]: tfidf_vect = TfidfVectorizer(min_df=10)
         tfidf_vect.fit(X_train)
         print("some sample features ",tfidf_vect.get_feature_names()[0:10])
         print('='*50)

         # we use the fitted CountVectorizer to convert the text to vector
         X_train_tfidf = tfidf_vect.transform(X_train)
         X_test_tfidf = tfidf_vect.transform(X_test)

         print("After vectorizations")
         print(X_train_tfidf.shape, Y_train.shape)
         print(X_test_tfidf.shape, Y_test.shape)
```

```
some sample features  ['aa', 'abandoned', 'abdominal', 'ability', 'able', 'abroad', 'absence',
==================================================
After vectorizations
(61441, 9615) (61441,)
(26332, 9615) (26332,)
```

5.3 [4.4] Word2Vec

In [28]:
```
# Train your own Word2Vec model using your own text corpus
list_of_sentance_train=[]
for sentance in X_train:
    list_of_sentance_train.append(sentance.split())
```

In [29]:
```
# this line of code trains your w2v model on the give list of sentances, fitting the 
w2v_model=Word2Vec(list_of_sentance_train,min_count=5,size=50, workers=-1)
```

In [30]:
```
w2v_words = list(w2v_model.wv.vocab)
print("number of words that occured minimum 5 times ",len(w2v_words))
print("sample words ", w2v_words[0:50])
```

```
number of words that occured minimum 5 times  14838
sample words  ['elegant', 'treat', 'unique', 'tiny', 'pieces', 'crystallized', 'ginger', 'pape
```

5.4 [4.4.1] Converting text into vectors using Avg W2V, TFIDF-W2V

[4.4.1.1] Avg W2v

5.4.1 Converting Train data set

In [31]:
```
# average Word2Vec
# compute average word2vec for each review.
sent_vectors_train = []; # the avg-w2v for each sentence/review is stored in this lis
for sent in tqdm(list_of_sentance_train): # for each review/sentence
    sent_vec = np.zeros(50) # as word vectors are of zero length 50, you might need t
    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_train.append(sent_vec)
sent_vectors_train = np.array(sent_vectors_train)
print(sent_vectors_train.shape)
print(sent_vectors_train[0])
```

```
100%|| 61441/61441 [02:59<00:00, 341.42it/s]


(61441, 50)
[-1.52331924e-03 -6.72609045e-04 -1.16279504e-06  1.54955658e-03
 -2.44743625e-04 -1.34313607e-03  6.51958653e-04  8.34782274e-04
 -6.89543737e-04 -1.94559560e-03  2.69455990e-03 -1.57745466e-04
```

```
 -3.36392239e-04 -2.48113894e-03  2.32932686e-03  1.17400606e-03
 -1.45757615e-03  3.66421816e-04  1.62687580e-03  1.57651301e-03
 -8.54632654e-04 -1.57016028e-03  1.34887942e-03  2.03365526e-04
  9.05575537e-04  2.04606804e-03  1.88504259e-03 -1.66922798e-03
 -7.92932755e-05  2.02520568e-04  2.33976118e-03 -4.32965695e-04
  1.48973602e-03 -1.79969327e-03  4.19588061e-05 -6.72537929e-04
 -1.74678222e-03  4.54549464e-04 -6.89351737e-04 -1.86725346e-03
  1.24476521e-04 -2.63654248e-04  2.36952642e-03 -6.51951992e-04
  2.13944926e-04  5.12260946e-04  3.91353900e-04  1.64038315e-03
 -3.17203890e-04  1.61795191e-03]
```

5.4.2 Converting Test data set

```
In [32]: list_of_sentance_test=[]
         for sentance in X_test:
             list_of_sentance_test.append(sentance.split())
```

```
In [33]: # average Word2Vec
         # compute average word2vec for each review.
         sent_vectors_test = []; # the avg-w2v for each sentence/review is stored in this list
         for sent in tqdm(list_of_sentance_test): # for each review/sentence
             sent_vec = np.zeros(50) # as word vectors are of zero length 50, you might need t
             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_test.append(sent_vec)
         sent_vectors_test = np.array(sent_vectors_test)
         print(sent_vectors_test.shape)
         print(sent_vectors_test[0])
```

```
100%|| 26332/26332 [01:18<00:00, 334.35it/s]


(26332, 50)
[ 1.16814716e-03  1.65648162e-03  1.34781573e-03  9.57494521e-04
 -2.11389616e-03 -9.77314180e-04  2.13661531e-03 -1.63896219e-03
 -1.12652907e-03  1.42470353e-03  4.11780828e-04  5.12305641e-04
 -3.40808348e-04 -6.88369745e-04  7.76484251e-04 -1.91222383e-04
  3.19553471e-04  1.60628137e-04  7.30847401e-04 -1.10237829e-03
  2.87855381e-04  1.28066931e-03  2.27156355e-03 -1.47085852e-03
  5.72064261e-04  1.13073081e-03 -1.18434567e-04 -3.15557236e-03
 -1.04568132e-03 -2.12658704e-04  6.11644985e-05 -6.86965823e-04
  3.16509801e-03 -1.04895481e-04 -1.37526119e-03  6.75385813e-04
```

```
-3.94201172e-04 -5.46043250e-04  2.84100117e-03 -1.53799535e-04
 3.51271716e-04 -6.97062521e-04 -1.21251263e-03 -1.82103339e-03
-4.29098091e-04  7.55673118e-05  7.45144060e-04 -1.01646979e-03
 3.62979109e-04 -3.22206398e-03]
```

[4.4.1.2] TFIDF weighted W2v

5.4.3 Converting Train Data

```
In [34]: # S = ["abc def pqr", "def def def abc", "pqr pqr def"]
         model = TfidfVectorizer()
         tf_idf_matrix_train = model.fit_transform(X_train)
         # we are converting a dictionary with word as a key, and the idf as a value
         dictionary = dict(zip(model.get_feature_names(), list(model.idf_)))
```

```
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

         tfidf_sent_vectors_train = []; # the tfidf-w2v for each sentence/review is stored in
         row=0;
         for sent in tqdm(list_of_sentance_train): # 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_train.append(sent_vec)
             row += 1
```

```
100%|| 61441/61441 [29:25<00:00, 34.80it/s]
```

5.4.4 Converting Test Data

```
In [36]: # 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
```

```python
tfidf_sent_vectors_test = []; # the tfidf-w2v for each sentence/review is stored in th
row=0;
for sent in tqdm(list_of_sentance_test): # 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_test.append(sent_vec)
    row += 1
```

```
100%|| 26332/26332 [06:45<00:00, 64.97it/s]
```

6 [5] Assignment 8: Decision Trees

```
<li><strong>Apply Decision Trees 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
        <li><font color='red'>SET 3:</font>Review text, preprocessed one converted into vectors
        <li><font color='red'>SET 4:</font>Review text, preprocessed one converted into vectors
    </ul>
</li>
<br>
<li><strong>The hyper paramter tuning (best `depth` in range [1, 5, 10, 50, 100, 500, 100], and
    <ul>
<li>Find the best hyper parameter which will give the maximum <a href='https://www.appliedaicou
<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>Graphviz</strong>
    <ul>
<li>Visualize your decision tree with Graphviz. It helps you to understand how a decision is be
<li>Since feature names are not obtained from word2vec related models, visualize only BOW & TFI
```

```
<li>Make sure to print the words in each node of the decision tree instead of printing its inde
<li>Just for visualization purpose, limit max_depth to 2 or 3 and either embed the generated ir
    </ul>
</li>
<br>
<li><strong>Feature importance</strong>
    <ul>
<li>Find the top 20 important features from both feature sets <font color='red'>Set 1</font> ar
    </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.

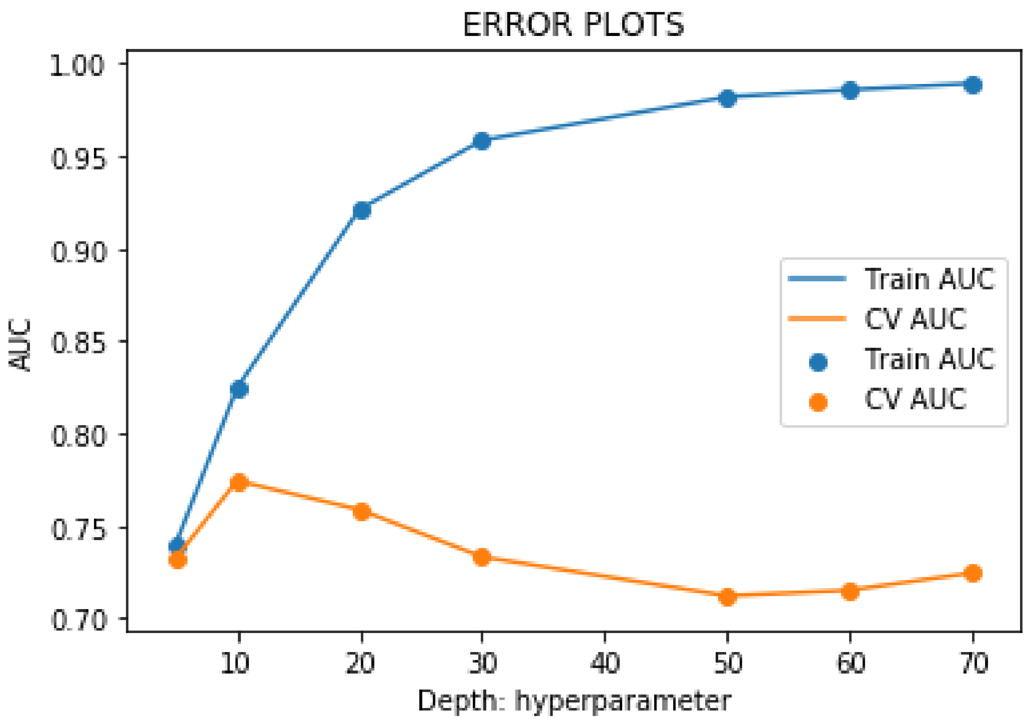
7 Applying Decision Trees

7.1 [5.1] Applying Decision Trees on BOW, SET 1

7.1.1 Hyperparameter tuning using GridSearch

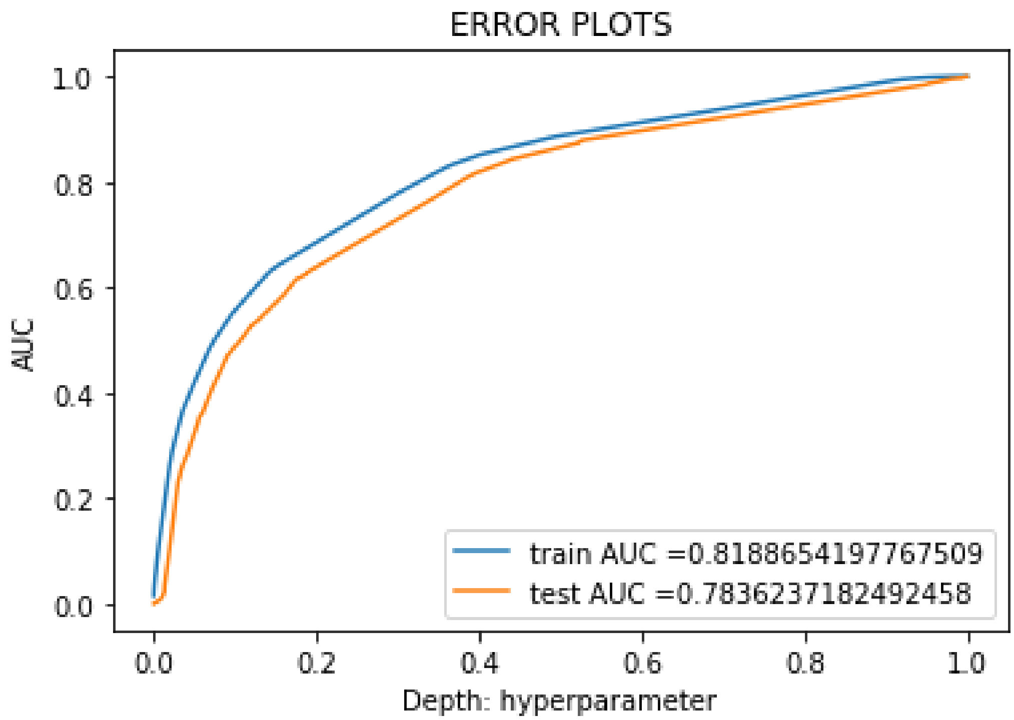
In [41]:
```
# clf = DecisionTreeClassifier()
# for Best Depth in Decision Tree
depth = [5,10,20,30,50,60,70]
parameters = {'max_depth':[5,10,20,30,50,60,70]}
grid = GridSearchCV(DecisionTreeClassifier(class_weight ='balanced',min_samples_split=
grid.fit(X_train_bow, Y_train)

print("best depth = ", grid.best_params_)

train_auc_bow = grid.cv_results_['mean_train_score']
cv_auc_bow = grid.cv_results_['mean_test_score']

plt.plot(depth, train_auc_bow, label='Train AUC')
plt.scatter(depth, train_auc_bow, label='Train AUC')
plt.plot(depth, cv_auc_bow, label='CV AUC')
plt.scatter(depth, cv_auc_bow, label='CV AUC')

plt.legend()
#plt.xscale('log')
plt.xlabel("Depth: hyperparameter")
plt.ylabel("AUC")
plt.title("ERROR PLOTS")
plt.show()
```

```
best depth =  {'max_depth': 10}
```



7.1.2 Testing with Test Data

In [115]:
```
clf = DecisionTreeClassifier(max_depth = 10, class_weight ='balanced')
clf.fit(X_train_bow, Y_train)

# roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates o
# not the predicted outputs

train_fpr_bow, train_tpr_bow, thresholds_bow = roc_curve(Y_train, clf.predict_proba(X
test_fpr_bow, test_tpr_bow, thresholds_bow = roc_curve(Y_test, clf.predict_proba(X_te

plt.plot(train_fpr_bow, train_tpr_bow, label="train AUC ="+str(auc(train_fpr_bow, tra
plt.plot(test_fpr_bow, test_tpr_bow, label="test AUC ="+str(auc(test_fpr_bow, test_tp
plt.legend()
plt.xlabel("Depth: hyperparameter")
plt.ylabel("AUC")
plt.title("ERROR PLOTS")
plt.show()
```



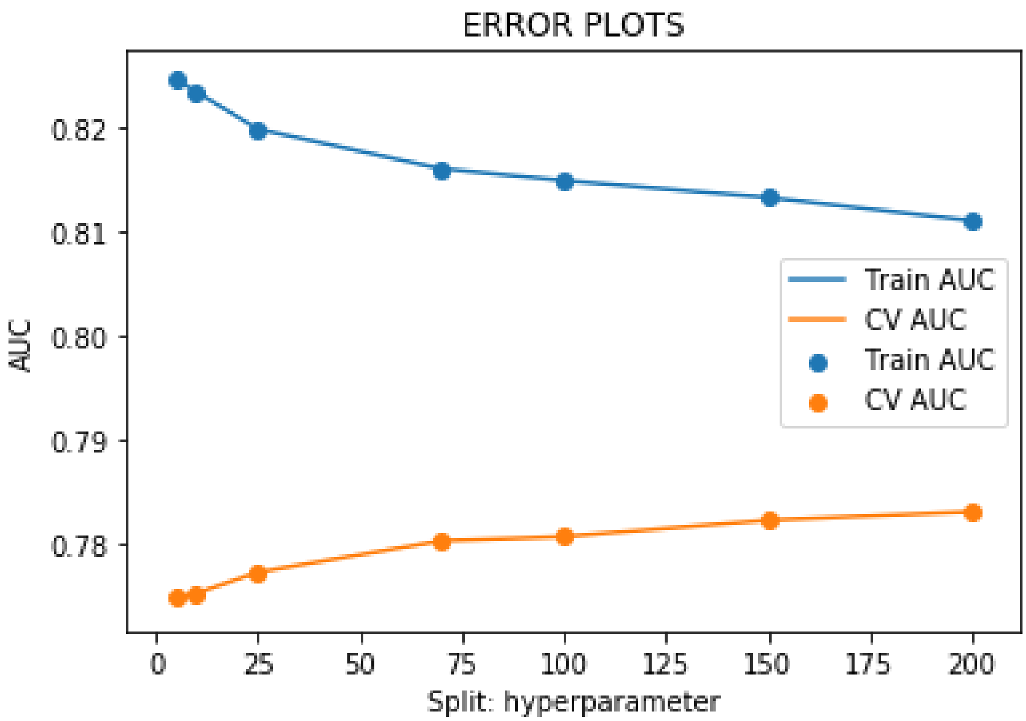
In [116]:
```
bow_depth = auc(test_fpr_bow, test_tpr_bow)
print(bow_depth)
```

0.7836237182492458

In [44]:
```
# clf = DecisionTreeClassifier()
# for Minimum samples split in Decision Tree
split = [5,10,25,70,100,150,200]
parameters = {'min_samples_split':[5,10,25,70,100,150,200]}
grid = GridSearchCV(DecisionTreeClassifier(class_weight ='balanced',max_depth=10), pa
grid.fit(X_train_bow, Y_train)

print("best samples split = ", grid.best_params_)

train_auc_bow = grid.cv_results_['mean_train_score']
cv_auc_bow = grid.cv_results_['mean_test_score']

plt.plot(split, train_auc_bow, label='Train AUC')
plt.scatter(split, train_auc_bow, label='Train AUC')
plt.plot(split, cv_auc_bow, label='CV AUC')
plt.scatter(split, cv_auc_bow, label='CV AUC')

plt.legend()
```

```
    #plt.xscale('log')
    plt.xlabel("Split: hyperparameter")
    plt.ylabel("AUC")
    plt.title("ERROR PLOTS")
    plt.show()
```

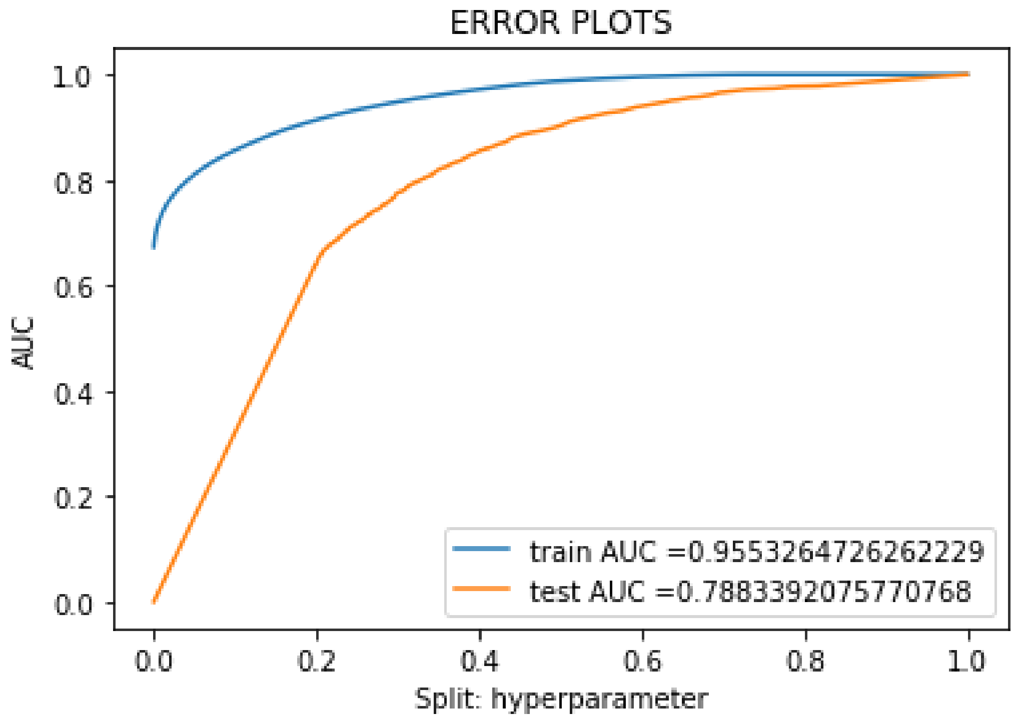
```
best samples split =  {'min_samples_split': 200}
```



7.1.3 Testing with Test Data

```
In [117]: clf = DecisionTreeClassifier(min_samples_split = 200, class_weight ='balanced')
          clf.fit(X_train_bow, Y_train)

          # roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates o
          # not the predicted outputs

          train_fpr_bow, train_tpr_bow, thresholds_bow = roc_curve(Y_train, clf.predict_proba(X
          test_fpr_bow, test_tpr_bow, thresholds_bow = roc_curve(Y_test, clf.predict_proba(X_te

          plt.plot(train_fpr_bow, train_tpr_bow, label="train AUC ="+str(auc(train_fpr_bow, tra
          plt.plot(test_fpr_bow, test_tpr_bow, label="test AUC ="+str(auc(test_fpr_bow, test_tp
          plt.legend()
          plt.xlabel("Split: hyperparameter")
```

```
plt.ylabel("AUC")
plt.title("ERROR PLOTS")
plt.show()
```




```
In [118]: bow_split = auc(test_fpr_bow, test_tpr_bow)
           print(bow_split)
```

```
0.7883392075770768
```

```
In [47]: split = [5,10,25,70,100,150,200]
         depth = [5,10,20,30,50,60,70]

         parameters = {'min_samples_split': split,'max_depth': depth}
         grid = GridSearchCV(DecisionTreeClassifier(class_weight ='balanced'), parameters, cv=3
         grid.fit(X_train_bow, Y_train)
```

```
Out[47]: GridSearchCV(cv=3, error_score='raise',
                estimator=DecisionTreeClassifier(class_weight='balanced', criterion='gini',
                     max_depth=None, max_features=None, max_leaf_nodes=None,
                     min_impurity_decrease=0.0, min_impurity_split=None,
                     min_samples_leaf=1, min_samples_split=2,
                     min_weight_fraction_leaf=0.0, presort=False, random_state=None,
                     splitter='best'),
```

```
             fit_params=None, iid=True, n_jobs=-1,
             param_grid={'min_samples_split': [5, 10, 25, 70, 100, 150, 200], 'max_depth':
             pre_dispatch='2*n_jobs', refit=True, return_train_score='warn',
             scoring='roc_auc', verbose=0)
```

```
In [49]: optimal_split = grid.best_estimator_.min_samples_split
         print("The optimal number of samples split is : ",optimal_split)

         optimal_depth = grid.best_estimator_.max_depth
         print("The optimal number of depth is : ",optimal_depth)
```

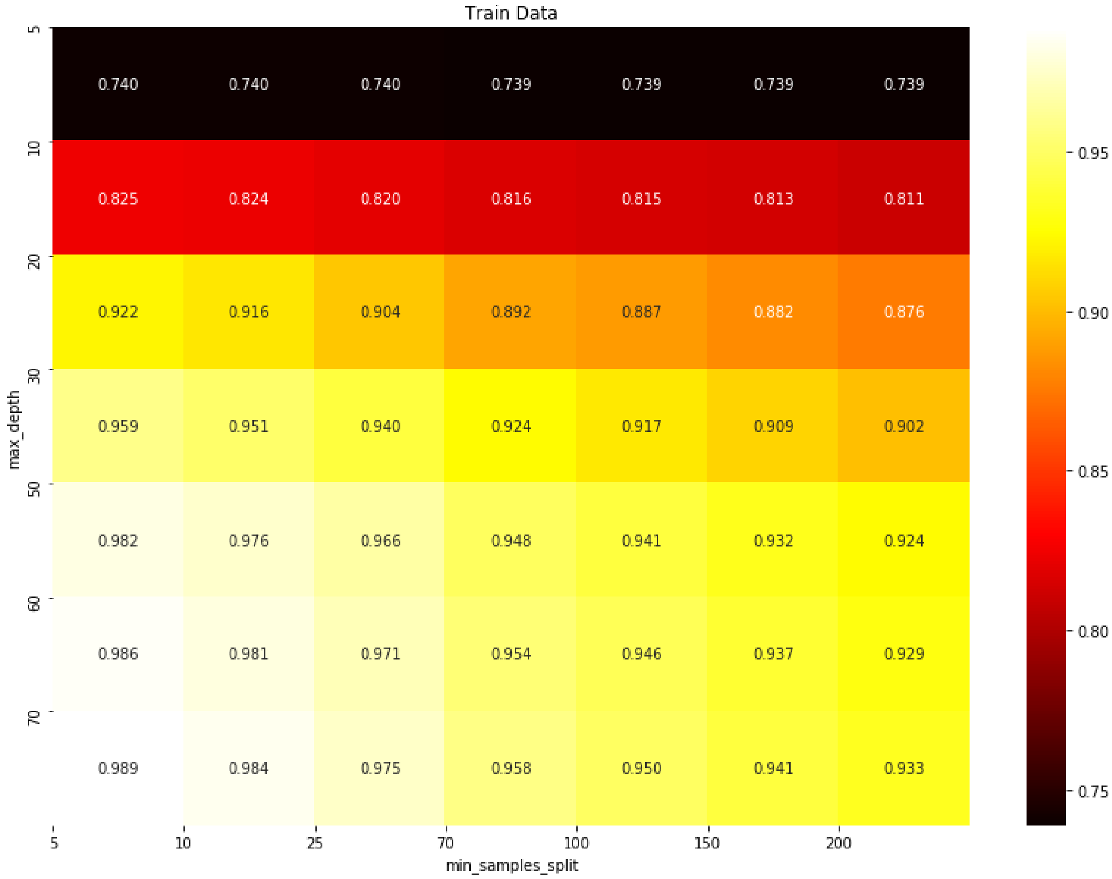
```
The optimal number of samples split is :  200
The optimal number of depth is :  30
```

```
In [50]: clf = DecisionTreeClassifier(min_samples_split = optimal_split, max_depth = optimal_de
         clf.fit(X_train_bow, Y_train)
         predb = clf.predict(X_test_bow)

         accb = accuracy_score(Y_test, predb) * 100
         preb = precision_score(Y_test, predb) * 100
         recb = recall_score(Y_test, predb) * 100
         f1b = f1_score(Y_test, predb) * 100

         print('\nAccuracy=%f%%' % (accb))
         print('\nprecision=%f%%' % (preb))
         print('\nrecall=%f%%' % (recb))
         print('\nF1-Score=%f%%' % (f1b))
```

```
Accuracy=85.796749%

precision=89.059844%

recall=94.728752%

F1-Score=91.806870%
```

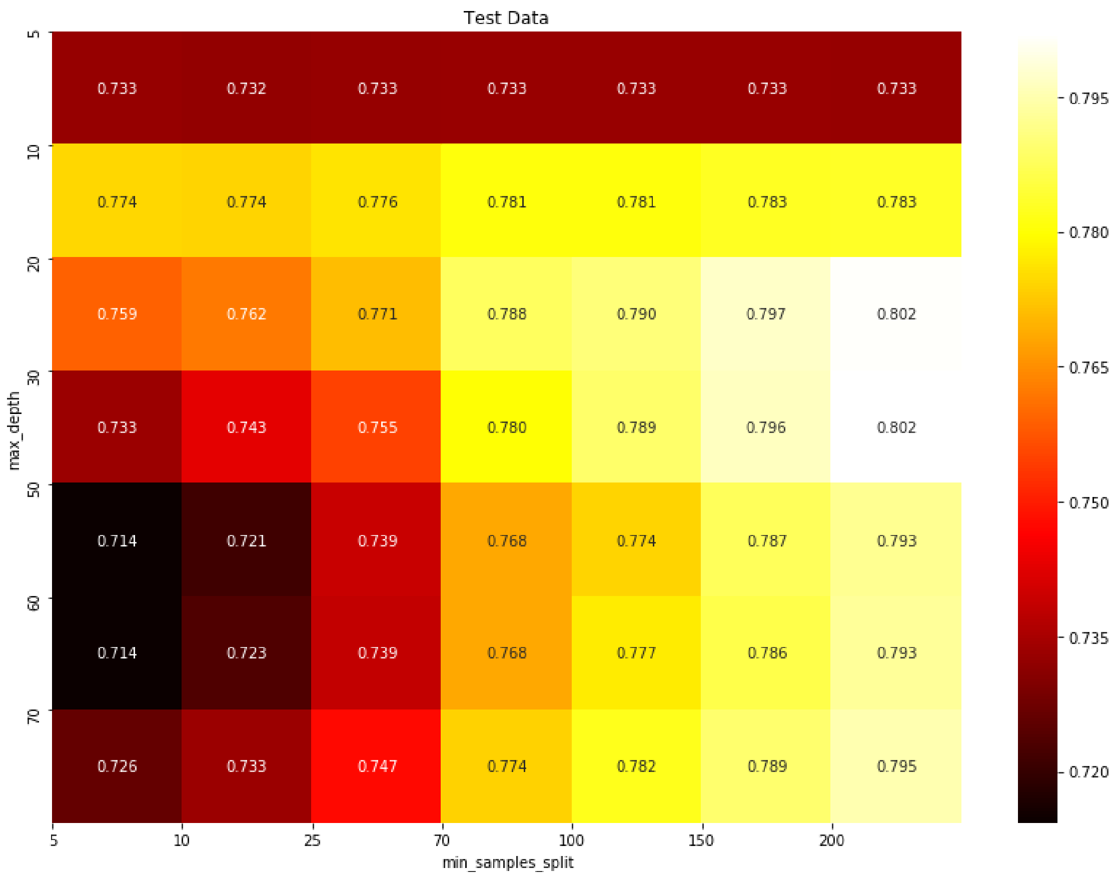
7.1.4 Heatmap on Train Data

```
In [53]: scores = grid.cv_results_['mean_train_score'].reshape(len(split),len(depth))
         plt.figure(figsize=(14,10))
         sns.heatmap(scores, annot=True, cmap=plt.cm.hot, fmt=".3f", xticklabels=split, ytickla
         plt.xlabel('min_samples_split')
         plt.ylabel('max_depth')
         plt.xticks(np.arange(len(split)), split)
         plt.yticks(np.arange(len(depth)), depth)
         plt.title('Train Data')
         plt.show()
```



7.1.5 Heatmap on Test Data

```
In [54]: scores = grid.cv_results_['mean_test_score'].reshape(len(split),len(depth))
         plt.figure(figsize=(14,10))
         sns.heatmap(scores, annot=True, cmap=plt.cm.hot, fmt=".3f", xticklabels=split, yticklab
         plt.xlabel('min_samples_split')
         plt.ylabel('max_depth')
         plt.xticks(np.arange(len(split)), split)
         plt.yticks(np.arange(len(depth)), depth)
         plt.title('Test Data')
         plt.show()
```



7.1.6 [5.1.1] Top 20 important features from SET 1

In [104]:
```
# Calculate feature importances from decision trees
importances = clf.feature_importances_

# Sort feature importances in descending order
indices = list(np.argsort(importances)[::-1][:50])
print(indices)
```

```
[15, 13, 41, 11, 22, 43, 2, 31, 47, 17, 37, 24, 21, 8, 49, 35, 44, 30, 16, 5, 45, 14, 42, 12, 
```

In [105]:
```
names = np.array(vectorizer.get_feature_names())
print(names[indices])
```

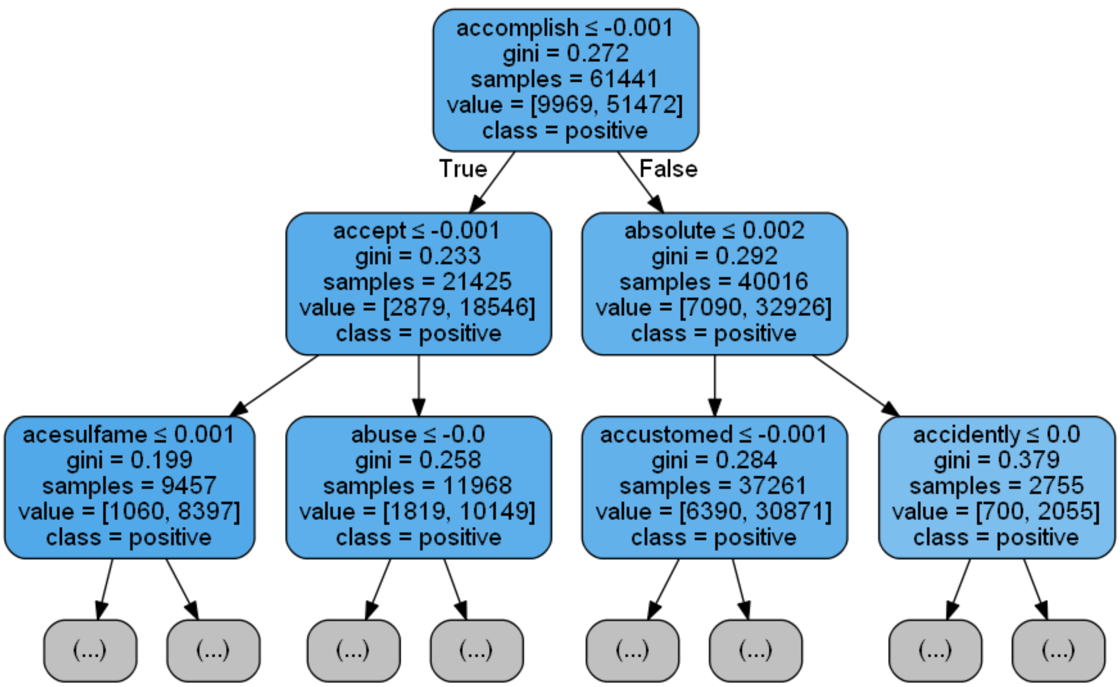
```
['absorption' 'absorbing' 'accurately' 'absorb' 'accent' 'acerola'
 'abdominal' 'accidents' 'achieve' 'abundance' 'accordingly' 'acceptable'
 'acana' 'absolute' 'acid' 'accomplish' 'acesulfame' 'accidently' 'absurd'
 'abroad' 'ache' 'absorbs' 'accustomed' 'absorbed' 'absolutly'
 'absolutely' 'accompany' 'absent' 'accurate' 'able' 'ability' 'aches'
```

```
'abandoned' 'absence' 'accounts' 'accompaniment' 'abundant' 'abuse'
'acai' 'account' 'according' 'accept' 'achieved' 'accepted' 'access'
'accessible' 'accident' 'accidentally' 'accompanied' 'aa']
```

7.1.7 [5.1.2] Graphviz visualization of Decision Tree on BOW, SET 1

In [106]:
```
target = ['negative','positive']
# Create DOT data
data = tree.export_graphviz(clf, max_depth=2,feature_names= names[indices], out_file=

# Draw graph
graph = pydotplus.graph_from_dot_data(data)

# Show graph
Image(graph.create_png())
```

Out[106]:



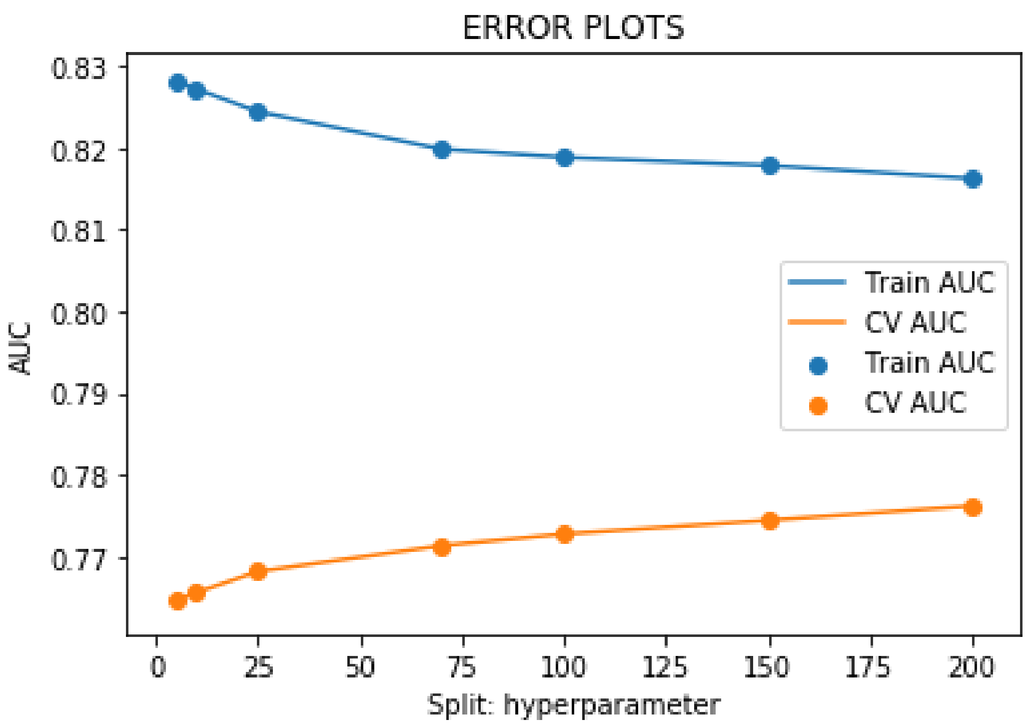
7.2 [5.2] Applying Decision Trees on TFIDF, SET 2

7.2.1 Hyperparameter tuning using GridSearch

In [56]:
```
# clf = DecisionTreeClassifier()
# for Minimum samples split in Decision Tree
split = [5,10,25,70,100,150,200]
```

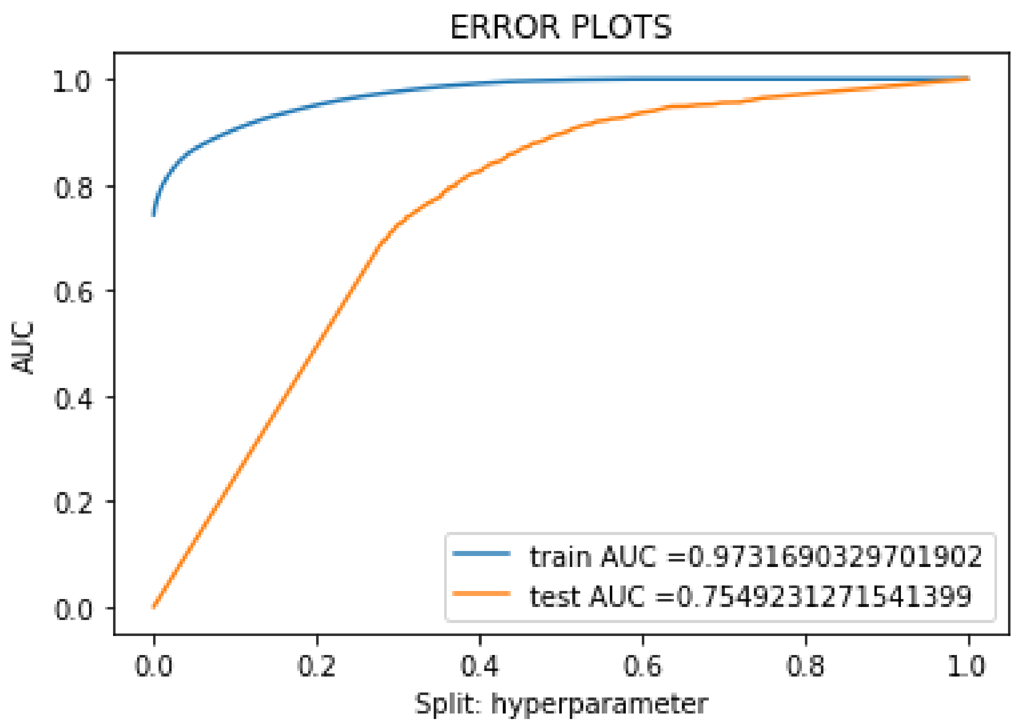
```
parameters = {'min_samples_split':[5,10,25,70,100,150,200]}
grid = GridSearchCV(DecisionTreeClassifier(class_weight ='balanced',max_depth=10), pa
grid.fit(X_train_tfidf, Y_train)

print("best samples split = ", grid.best_params_)

train_auc_tfidf = grid.cv_results_['mean_train_score']
cv_auc_tfidf = grid.cv_results_['mean_test_score']

plt.plot(split, train_auc_tfidf, label='Train AUC')
plt.scatter(split, train_auc_tfidf, label='Train AUC')
plt.plot(split, cv_auc_tfidf, label='CV AUC')
plt.scatter(split, cv_auc_tfidf, label='CV AUC')

plt.legend()
plt.xlabel("Split: hyperparameter")
plt.ylabel("AUC")
plt.title("ERROR PLOTS")
plt.show()
```

```
best samples split =  {'min_samples_split': 200}
```



7.2.2 Testing with Test Data

```
In [119]: clf = DecisionTreeClassifier(min_samples_split = 200, class_weight ='balanced')
          clf.fit(X_train_tfidf, Y_train)

          # roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates o
          # not the predicted outputs

          train_fpr_tfidf, train_tpr_tfidf, thresholds_tfidf = roc_curve(Y_train, clf.predict_
          test_fpr_tfidf, test_tpr_tfidf, thresholds_tfidf = roc_curve(Y_test, clf.predict_pro

          plt.plot(train_fpr_tfidf, train_tpr_tfidf, label="train AUC ="+str(auc(train_fpr_tfi
          plt.plot(test_fpr_tfidf, test_tpr_tfidf, label="test AUC ="+str(auc(test_fpr_tfidf, 
          plt.legend()
          plt.xlabel("Split: hyperparameter")
          plt.ylabel("AUC")
          plt.title("ERROR PLOTS")
          plt.show()
```

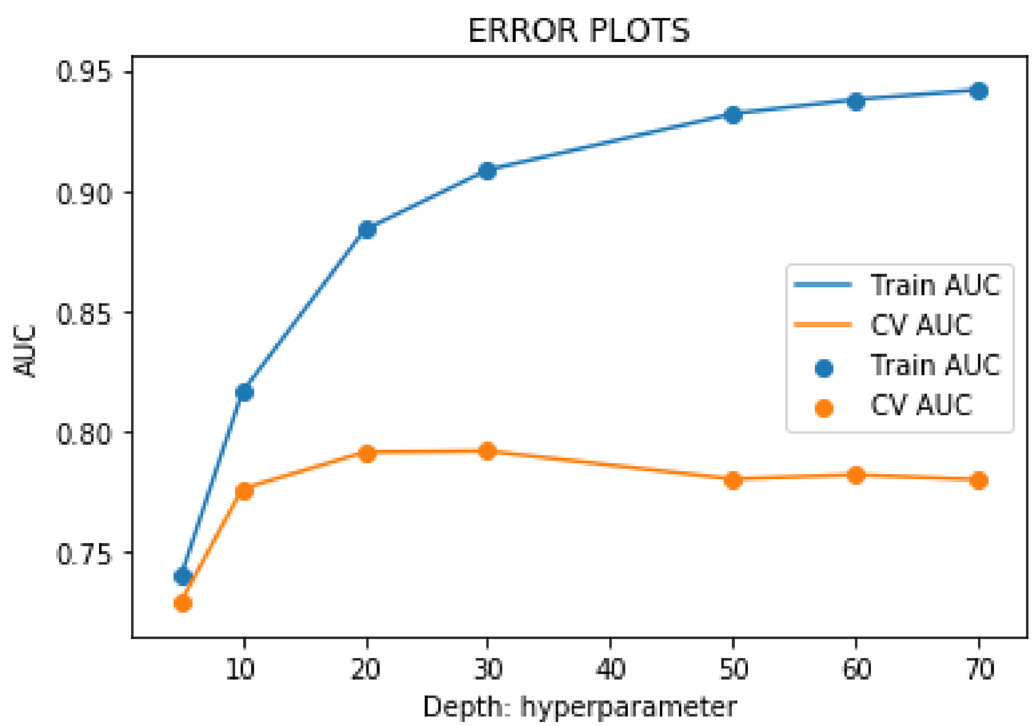



```
In [121]: tfidf_split = auc(test_fpr_tfidf, test_tpr_tfidf)
          print(tfidf_split)
```

0.7549231271541399

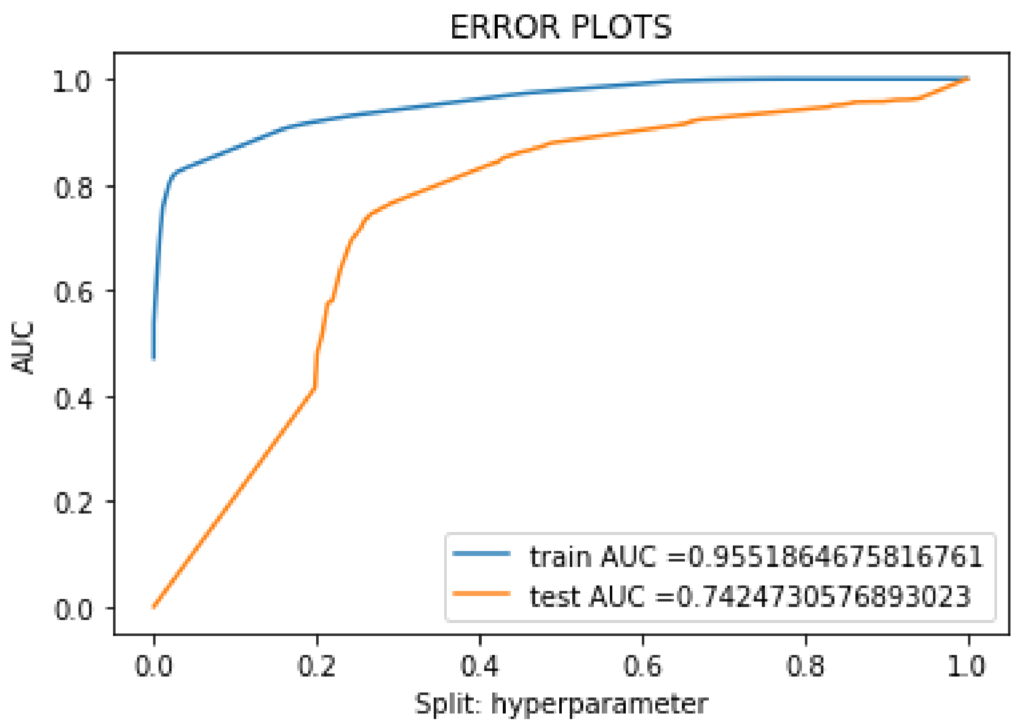
```
In [58]: # clf = DecisionTreeClassifier()
         # for Best Depth in Decision Tree
         depth = [5,10,20,30,50,60,70]
         parameters = {'max_depth':[5,10,20,30,50,60,70]}
         grid = GridSearchCV(DecisionTreeClassifier(class_weight ='balanced',min_samples_split=
         grid.fit(X_train_tfidf, Y_train)

         print("best depth = ", grid.best_params_)

         train_auc_tfidf = grid.cv_results_['mean_train_score']
         cv_auc_tfidf = grid.cv_results_['mean_test_score']

         plt.plot(depth, train_auc_tfidf, label='Train AUC')
         plt.scatter(depth, train_auc_tfidf, label='Train AUC')
         plt.plot(depth, cv_auc_tfidf, label='CV AUC')
         plt.scatter(depth, cv_auc_tfidf, label='CV AUC')

         plt.legend()
         plt.xlabel("Depth: hyperparameter")
         plt.ylabel("AUC")
         plt.title("ERROR PLOTS")
         plt.show()
```

```
best depth =  {'max_depth': 30}
```



7.2.3 Testing with Test Data

In [122]:
```
clf = DecisionTreeClassifier(max_depth = 30, class_weight ='balanced')
clf.fit(X_train_tfidf, Y_train)

# roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates o
# not the predicted outputs

train_fpr_tfidf, train_tpr_tfidf, thresholds_tfidf = roc_curve(Y_train, clf.predict_
test_fpr_tfidf, test_tpr_tfidf, thresholds_tfidf = roc_curve(Y_test, clf.predict_pro

plt.plot(train_fpr_tfidf, train_tpr_tfidf, label="train AUC ="+str(auc(train_fpr_tfi
plt.plot(test_fpr_tfidf, test_tpr_tfidf, label="test AUC ="+str(auc(test_fpr_tfidf, 
plt.legend()
plt.xlabel("Split: hyperparameter")
plt.ylabel("AUC")
plt.title("ERROR PLOTS")
plt.show()
```



In [123]:
```
tfidf_depth = auc(test_fpr_tfidf, test_tpr_tfidf)
print(tfidf_depth)
```

```
0.7424730576893023
```

```
In [61]: split = [5,10,25,70,100,150,200]
         depth = [5,10,20,30,50,60,70]

         parameters = {'min_samples_split': split,'max_depth': depth}
         grid = GridSearchCV(DecisionTreeClassifier(class_weight ='balanced'), parameters, cv=3
         grid.fit(X_train_tfidf, Y_train)
```

```
Out[61]: GridSearchCV(cv=3, error_score='raise',
                estimator=DecisionTreeClassifier(class_weight='balanced', criterion='gini',
                     max_depth=None, max_features=None, max_leaf_nodes=None,
                     min_impurity_decrease=0.0, min_impurity_split=None,
                     min_samples_leaf=1, min_samples_split=2,
                     min_weight_fraction_leaf=0.0, presort=False, random_state=None,
                     splitter='best'),
                fit_params=None, iid=True, n_jobs=-1,
                param_grid={'min_samples_split': [5, 10, 25, 70, 100, 150, 200], 'max_depth':
                pre_dispatch='2*n_jobs', refit=True, return_train_score='warn',
                scoring='roc_auc', verbose=0)
```

```
In [67]: optimal_split = grid.best_estimator_.min_samples_split
         print("The optimal number of samples split is : ",optimal_split)

         optimal_depth = grid.best_estimator_.max_depth
         print("The optimal number of depth is : ",optimal_depth)
```

```
The optimal number of samples split is :  200
The optimal number of depth is :  30
```

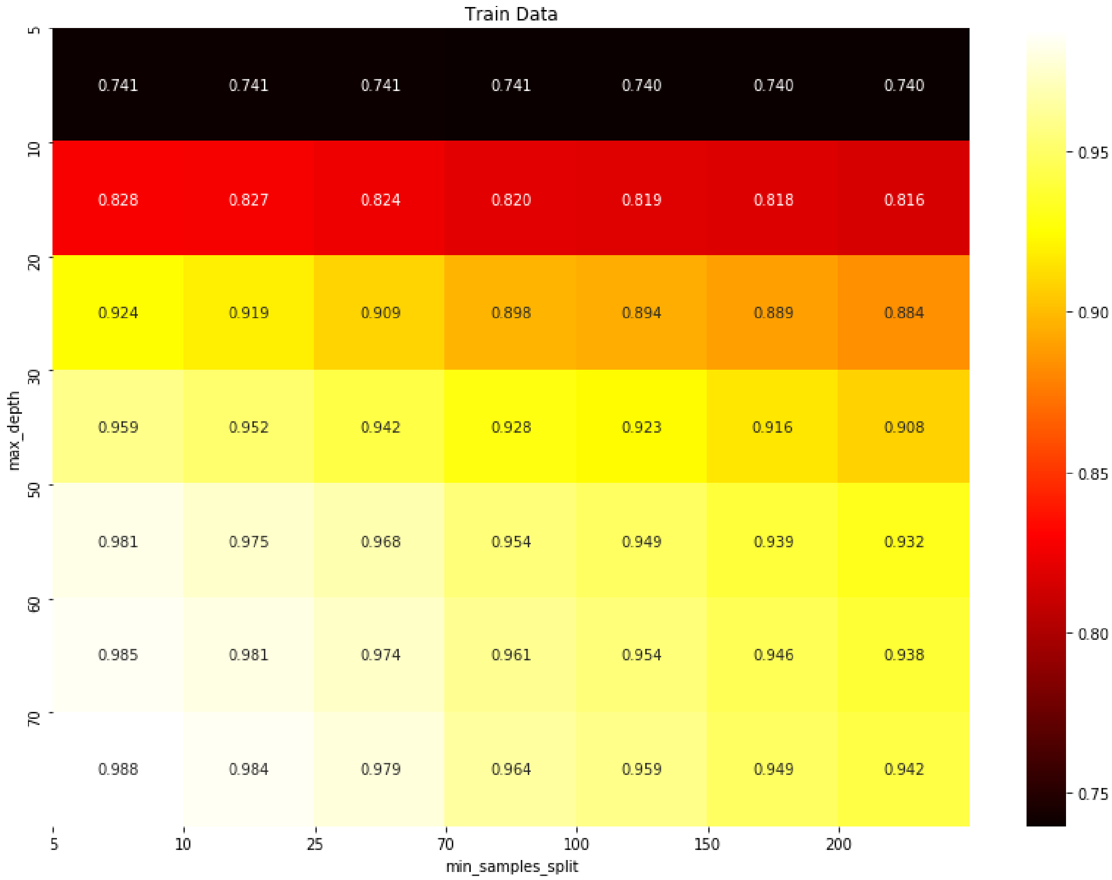
```
In [63]: clf = DecisionTreeClassifier(min_samples_split = optimal_split, max_depth = optimal_de
         clf.fit(X_train_tfidf, Y_train)
         predt = clf.predict(X_test_tfidf)

         acct = accuracy_score(Y_test, predt) * 100
         pret = precision_score(Y_test, predt) * 100
         rect = recall_score(Y_test, predt) * 100
         f1t = f1_score(Y_test, predt) * 100

         print('\nAccuracy=%f%%' % (acct))
         print('\nprecision=%f%%' % (pret))
         print('\nrecall=%f%%' % (rect))
         print('\nF1-Score=%f%%' % (f1t))
```

```
Accuracy=85.568890%

precision=88.182576%

recall=95.637432%
```

```
F1-Score=91.758838%
```

7.2.4 Heatmap on Train data

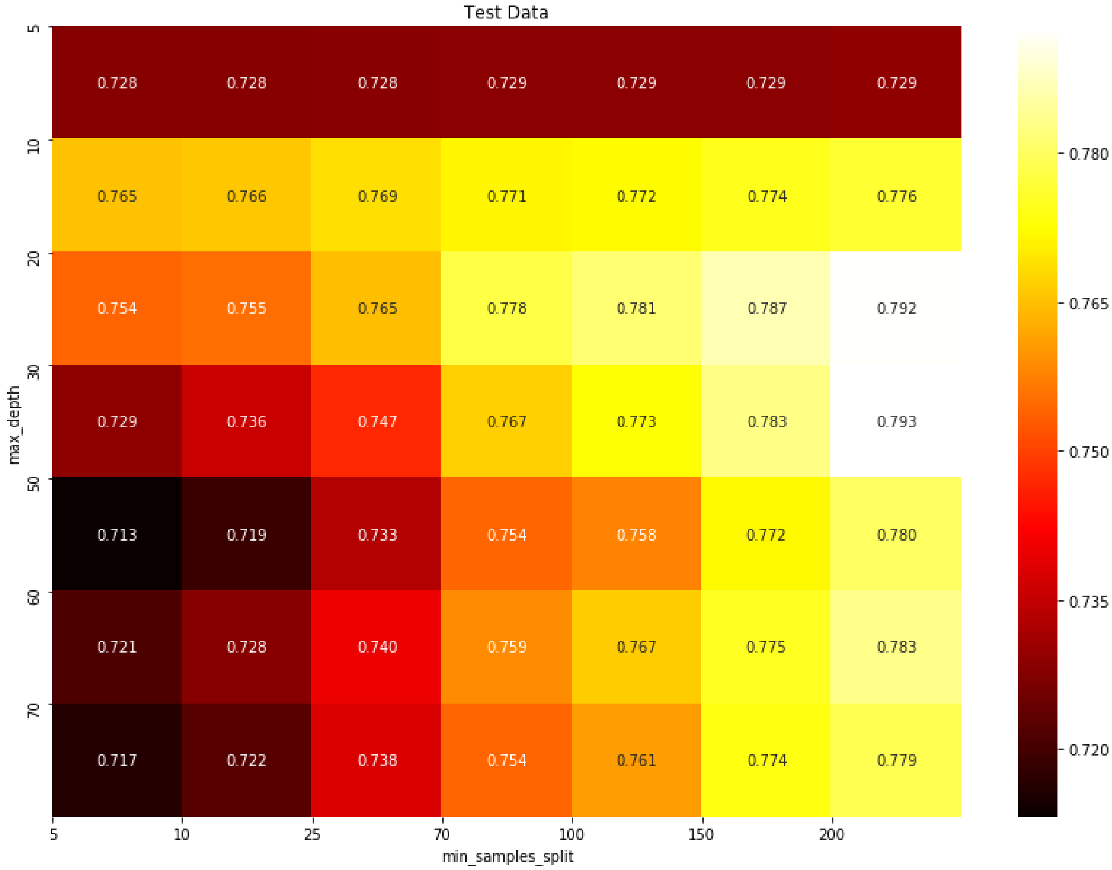
```
In [64]: scores = grid.cv_results_['mean_train_score'].reshape(len(split),len(depth))
         plt.figure(figsize=(14,10))
         sns.heatmap(scores, annot=True, cmap=plt.cm.hot, fmt=".3f", xticklabels=split, ytickla
         plt.xlabel('min_samples_split')
         plt.ylabel('max_depth')
         plt.xticks(np.arange(len(split)), split)
         plt.yticks(np.arange(len(depth)), depth)
         plt.title('Train Data')
         plt.show()
```



7.2.5 Heatmap on Test Data

```
In [65]: scores = grid.cv_results_['mean_test_score'].reshape(len(split),len(depth))
         plt.figure(figsize=(14,10))
```

```
sns.heatmap(scores, annot=True, cmap=plt.cm.hot, fmt=".3f", xticklabels=split, ytickl
plt.xlabel('min_samples_split')
plt.ylabel('max_depth')
plt.xticks(np.arange(len(split)), split)
plt.yticks(np.arange(len(depth)), depth)
plt.title('Test Data')
plt.show()
```



7.2.6 [5.2.1] Top 20 important features from SET 2

```
In [101]: # Calculate feature importances from decision trees
          importances = clf.feature_importances_

          # Sort feature importances in descending order
          indices = list(np.argsort(importances)[::-1][:50])
          print(indices)
```

```
[15, 13, 41, 11, 22, 43, 2, 31, 47, 17, 37, 24, 21, 8, 49, 35, 44, 30, 16, 5, 45, 14, 42, 12, 
```

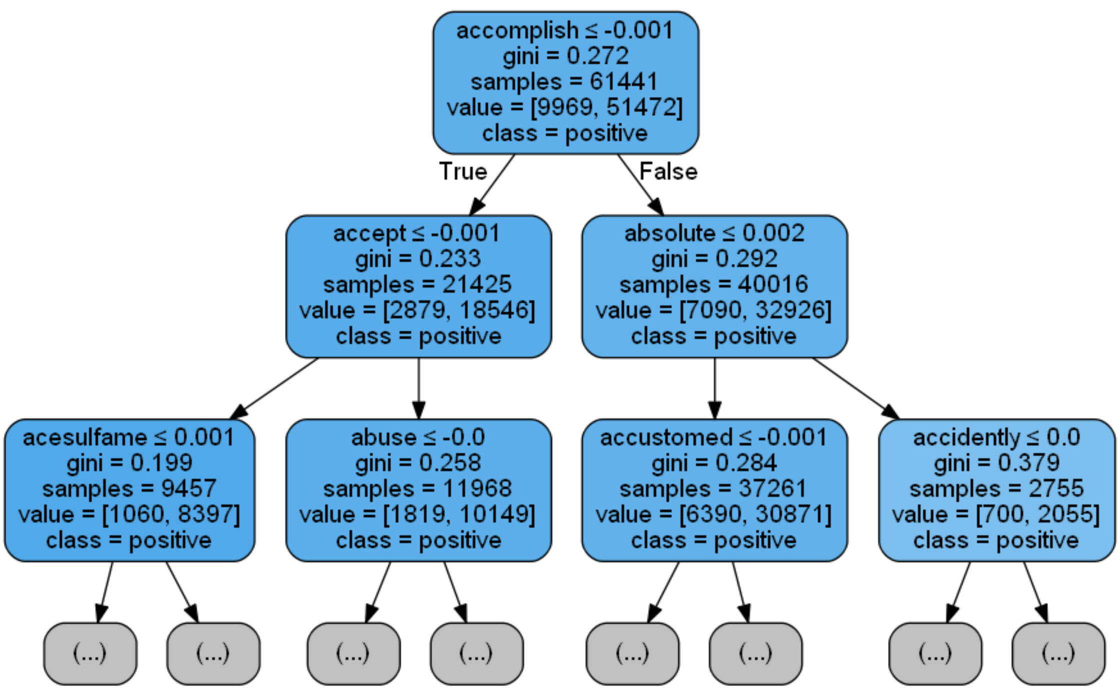
```
In [102]: names = np.array(vectorizer.get_feature_names())
          print(names[indices])
```

```
['absorption' 'absorbing' 'accurately' 'absorb' 'accent' 'acerola'
 'abdominal' 'accidents' 'achieve' 'abundance' 'accordingly' 'acceptable'
 'acana' 'absolute' 'acid' 'accomplish' 'acesulfame' 'accidently' 'absurd'
 'abroad' 'ache' 'absorbs' 'accustomed' 'absorbed' 'absolutly'
 'absolutely' 'accompany' 'absent' 'accurate' 'able' 'ability' 'aches'
 'abandoned' 'absence' 'accounts' 'accompaniment' 'abundant' 'abuse'
 'acai' 'account' 'according' 'accept' 'achieved' 'accepted' 'access'
 'accessible' 'accident' 'accidentally' 'accompanied' 'aa']
```

7.2.7 [5.2.2] Graphviz visualization of Decision Tree on TFIDF, SET 2

```
In [103]: target = ['negative','positive']
          # Create DOT data
          data = tree.export_graphviz(clf, max_depth=2,feature_names= names[indices], out_file=

          # Draw graph
          graph = pydotplus.graph_from_dot_data(data)

          # Show graph
          Image(graph.create_png())
```

Out[103]:

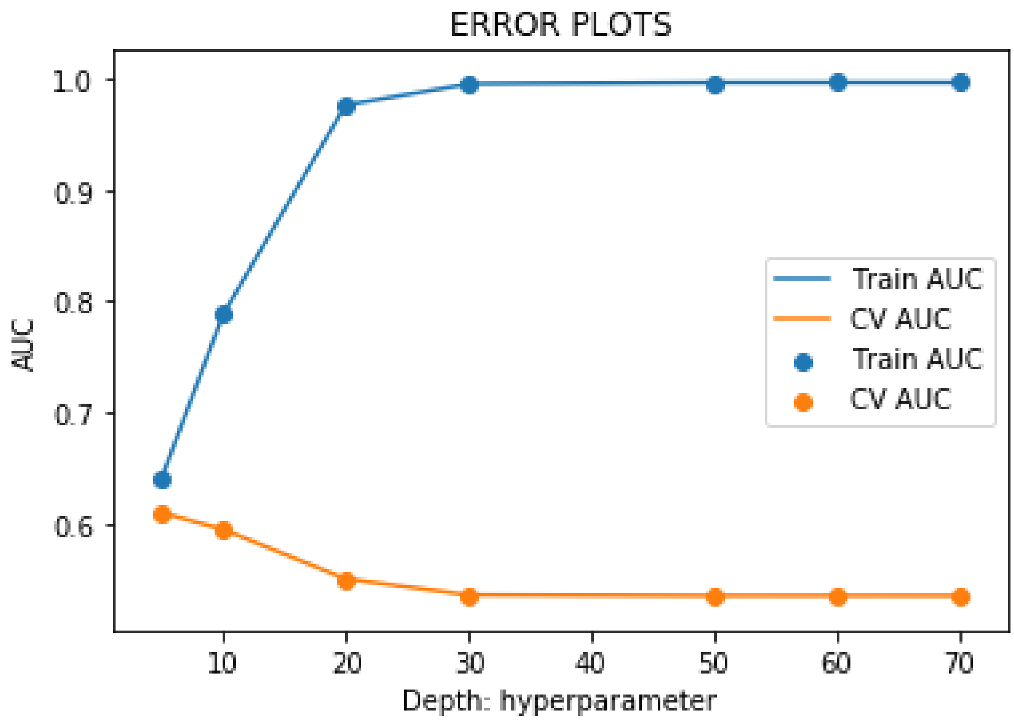


7.3 [5.3] Applying Decision Trees on AVG W2V, SET 3

7.3.1 Hyperparameter tuning using GridSearch

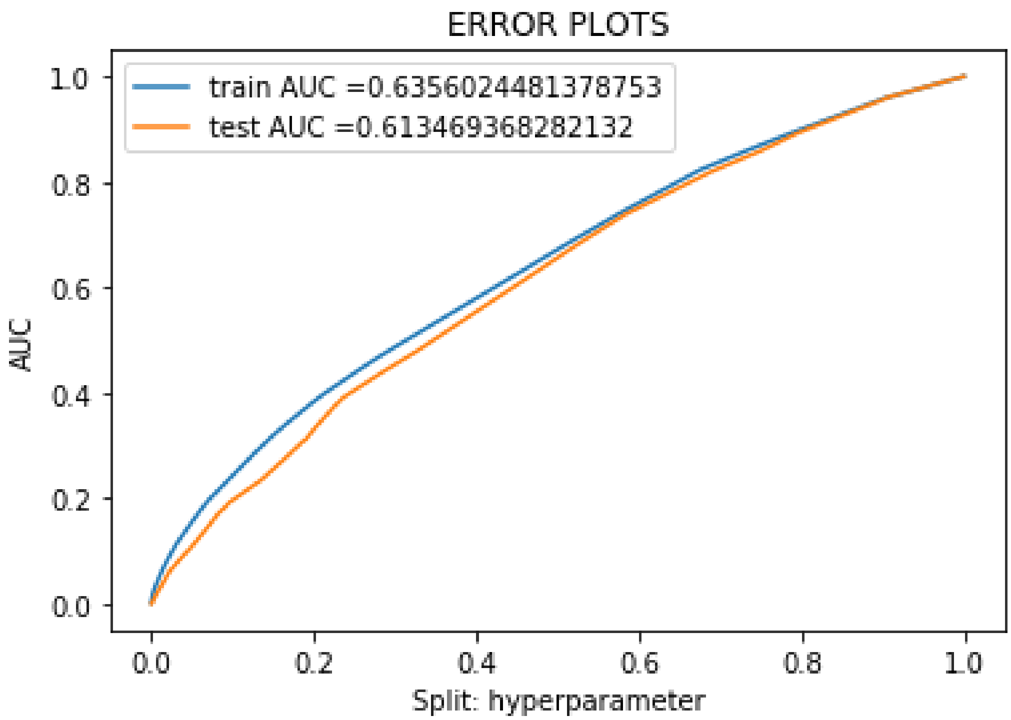
```
In [69]: # clf = DecisionTreeClassifier()
         # for Best Depth in Decision Tree
         depth = [5,10,20,30,50,60,70]
         parameters = {'max_depth':[5,10,20,30,50,60,70]}
         grid = GridSearchCV(DecisionTreeClassifier(class_weight ='balanced',min_samples_split=
         grid.fit(sent_vectors_train, Y_train)

         print("best depth = ", grid.best_params_)

         train_auc_aw2v = grid.cv_results_['mean_train_score']
         cv_auc_aw2v = grid.cv_results_['mean_test_score']

         plt.plot(depth, train_auc_aw2v, label='Train AUC')
         plt.scatter(depth, train_auc_aw2v, label='Train AUC')
         plt.plot(depth, cv_auc_aw2v, label='CV AUC')
         plt.scatter(depth, cv_auc_aw2v, label='CV AUC')

         plt.legend()
         plt.xlabel("Depth: hyperparameter")
         plt.ylabel("AUC")
         plt.title("ERROR PLOTS")
         plt.show()
```

```
best depth =  {'max_depth': 5}
```



7.3.2 Testing with Test Data

In [124]:
```
clf = DecisionTreeClassifier(max_depth = 5, class_weight ='balanced')
clf.fit(sent_vectors_train, Y_train)

# roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates o
# not the predicted outputs

train_fpr_aw2v, train_tpr_aw2v, thresholds_aw2v = roc_curve(Y_train, clf.predict_prob
test_fpr_aw2v, test_tpr_aw2v, thresholds_aw2v = roc_curve(Y_test, clf.predict_proba(s

plt.plot(train_fpr_aw2v, train_tpr_aw2v, label="train AUC ="+str(auc(train_fpr_aw2v,
plt.plot(test_fpr_aw2v, test_tpr_aw2v, label="test AUC ="+str(auc(test_fpr_aw2v, tes
plt.legend()
plt.xlabel("Split: hyperparameter")
plt.ylabel("AUC")
plt.title("ERROR PLOTS")
plt.show()
```



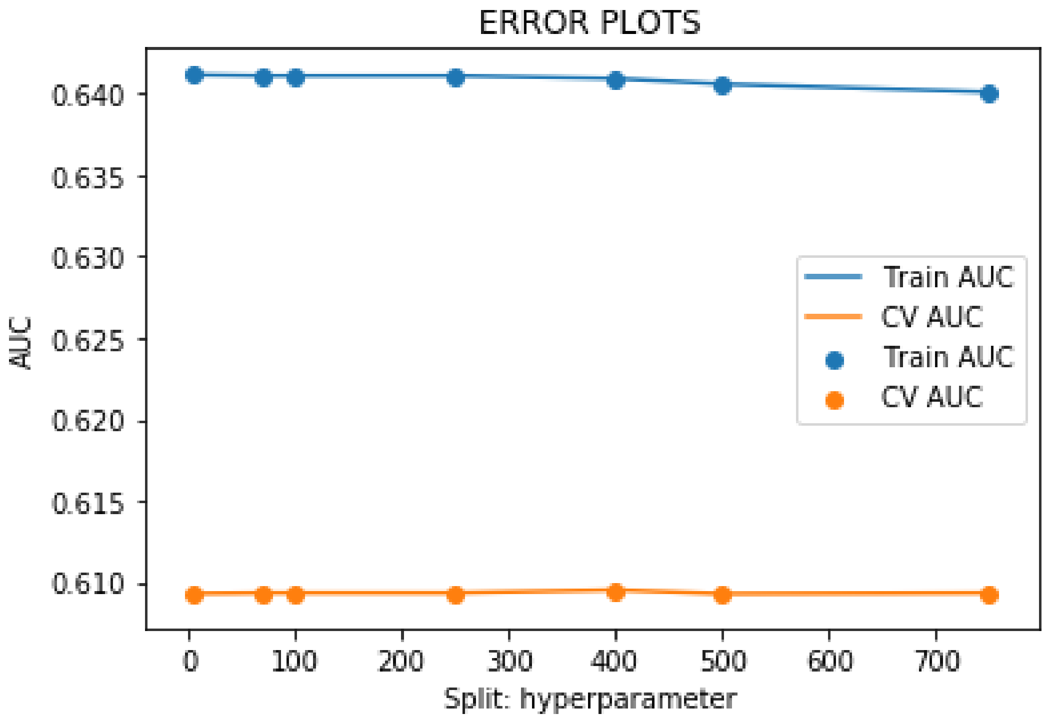

```
In [126]: aw2v_depth = auc(test_fpr_aw2v, test_tpr_aw2v)
          print(aw2v_depth)
```

```
0.613469368282132
```

```
In [73]: # clf = DecisionTreeClassifier()
         # for Minimum samples split in Decision Tree
         split = [5,70,100,250,400,500,750]
         parameters = {'min_samples_split':[5,70,100,250,400,500,750]}
         grid = GridSearchCV(DecisionTreeClassifier(class_weight ='balanced',max_depth=5), para
         grid.fit(sent_vectors_train, Y_train)

         print("best split = ", grid.best_params_)

         train_auc_aw2v = grid.cv_results_['mean_train_score']
         cv_auc_aw2v = grid.cv_results_['mean_test_score']

         plt.plot(split, train_auc_aw2v, label='Train AUC')
         plt.scatter(split, train_auc_aw2v, label='Train AUC')
         plt.plot(split, cv_auc_aw2v, label='CV AUC')
         plt.scatter(split, cv_auc_aw2v, label='CV AUC')

         plt.legend()
```

```
    plt.xlabel("Split: hyperparameter")
    plt.ylabel("AUC")
    plt.title("ERROR PLOTS")
    plt.show()
```

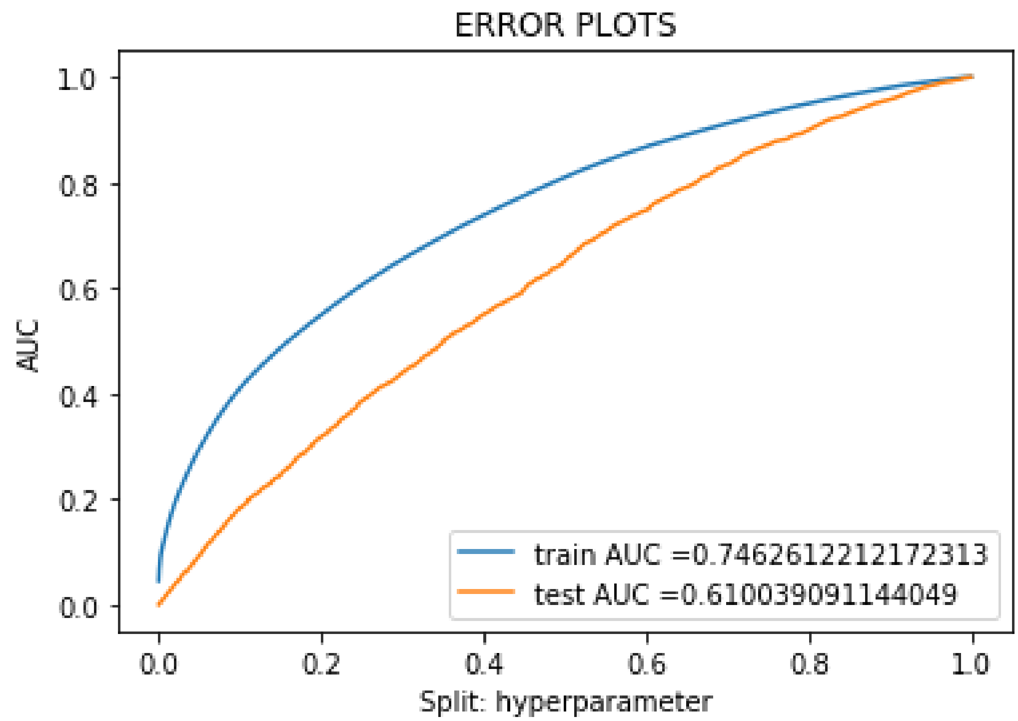
```
best split =  {'min_samples_split': 400}
```



7.3.3 Testing with Test Data

```
In [127]: clf = DecisionTreeClassifier(min_samples_split = 400, class_weight ='balanced')
          clf.fit(sent_vectors_train, Y_train)

          # roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates o
          # not the predicted outputs

          train_fpr_aw2v, train_tpr_aw2v, thresholds_aw2v = roc_curve(Y_train, clf.predict_prol
          test_fpr_aw2v, test_tpr_aw2v, thresholds_aw2v = roc_curve(Y_test, clf.predict_proba(

          plt.plot(train_fpr_aw2v, train_tpr_aw2v, label="train AUC ="+str(auc(train_fpr_aw2v,
          plt.plot(test_fpr_aw2v, test_tpr_aw2v, label="test AUC ="+str(auc(test_fpr_aw2v, tes
          plt.legend()
          plt.xlabel("Split: hyperparameter")
          plt.ylabel("AUC")
```

```
plt.title("ERROR PLOTS")
plt.show()
```



In [128]:
```
aw2v_split = auc(test_fpr_aw2v, test_tpr_aw2v)
print(aw2v_split)
```

```
0.610039091144049
```

In [76]:
```
split = [5,70,100,250,400,500,750]
depth = [5,10,20,30,50,60,70]

parameters = {'min_samples_split': split,'max_depth': depth}
grid = GridSearchCV(DecisionTreeClassifier(class_weight ='balanced'), parameters, cv=3
grid.fit(sent_vectors_train, Y_train)
```

Out[76]:
```
GridSearchCV(cv=3, error_score='raise',
       estimator=DecisionTreeClassifier(class_weight='balanced', criterion='gini',
            max_depth=None, max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=None,
            splitter='best'),
       fit_params=None, iid=True, n_jobs=-1,
```

```
             param_grid={'min_samples_split': [5, 70, 100, 250, 400, 500, 750], 'max_depth'
             pre_dispatch='2*n_jobs', refit=True, return_train_score='warn',
             scoring='roc_auc', verbose=0)
```

```
In [77]: optimal_split = grid.best_estimator_.min_samples_split
         print("The optimal number of samples split is : ",optimal_split)

         optimal_depth = grid.best_estimator_.max_depth
         print("The optimal number of depth is : ",optimal_depth)
```

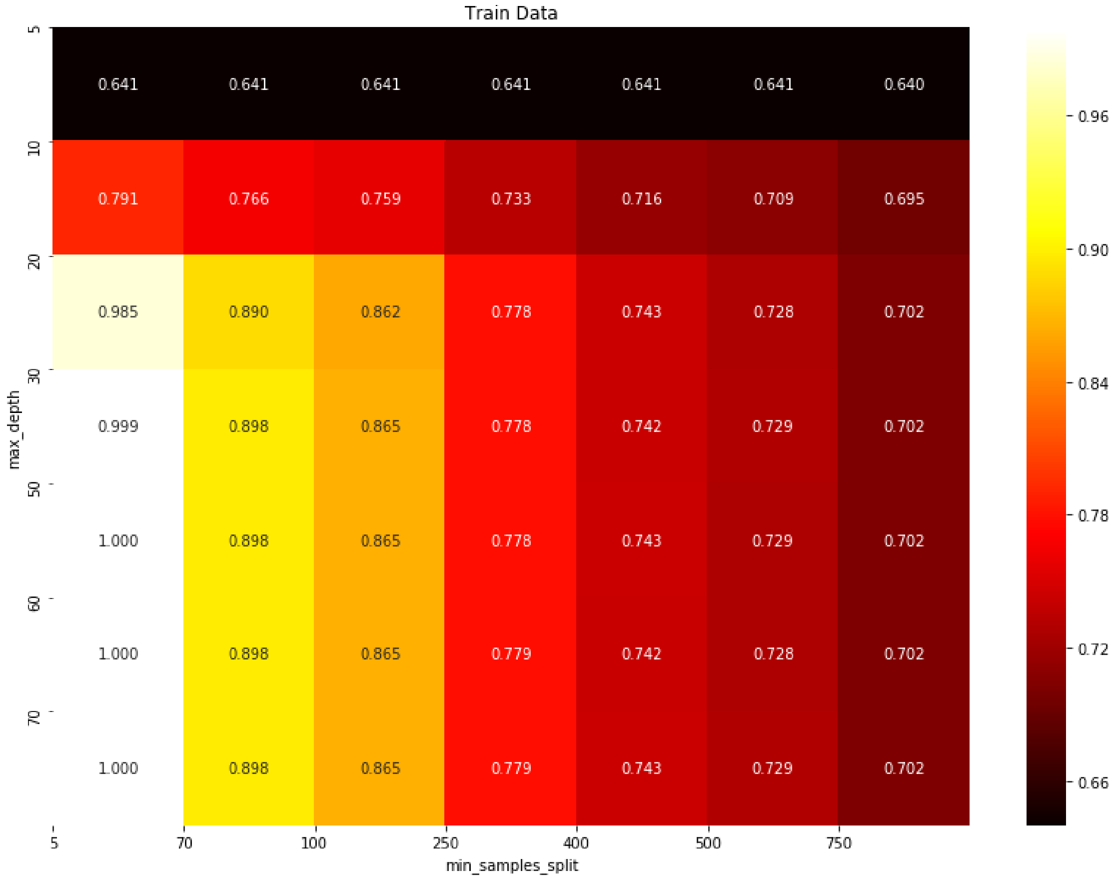
```
The optimal number of samples split is :  750
The optimal number of depth is :  10
```

```
In [78]: clf = DecisionTreeClassifier(min_samples_split = optimal_split, max_depth = optimal_de
         clf.fit(sent_vectors_train, Y_train)
         preda = clf.predict(sent_vectors_test)

         acca = accuracy_score(Y_test, preda) * 100
         prea = precision_score(Y_test, preda) * 100
         reca = recall_score(Y_test, preda) * 100
         f1a = f1_score(Y_test, preda) * 100

         print('\nAccuracy=%f%%' % (acca))
         print('\nprecision=%f%%' % (prea))
         print('\nrecall=%f%%' % (reca))
         print('\nF1-Score=%f%%' % (f1a))
```

```

Accuracy=83.692845%

precision=84.272860%

recall=99.077758%

F1-Score=91.077588%
```

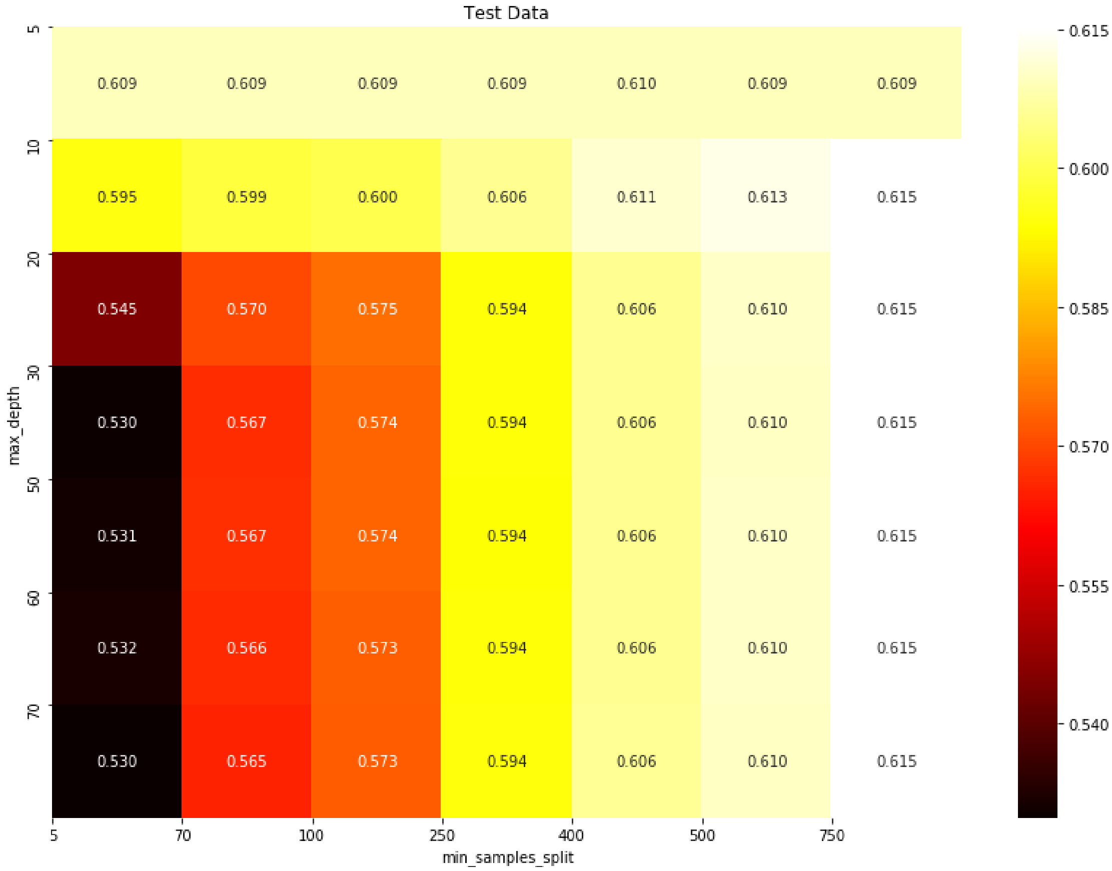
7.3.4 Heatmap on Train Data

```
In [79]: scores = grid.cv_results_['mean_train_score'].reshape(len(split),len(depth))
         plt.figure(figsize=(14,10))
         sns.heatmap(scores, annot=True, cmap=plt.cm.hot, fmt=".3f", xticklabels=split, ytickla
         plt.xlabel('min_samples_split')
         plt.ylabel('max_depth')
         plt.xticks(np.arange(len(split)), split)
         plt.yticks(np.arange(len(depth)), depth)
         plt.title('Train Data')
         plt.show()
```



7.3.5 Heatmap on Test Data

```
In [80]: scores = grid.cv_results_['mean_test_score'].reshape(len(split),len(depth))
         plt.figure(figsize=(14,10))
         sns.heatmap(scores, annot=True, cmap=plt.cm.hot, fmt=".3f", xticklabels=split, yticklabe
         plt.xlabel('min_samples_split')
         plt.ylabel('max_depth')
         plt.xticks(np.arange(len(split)), split)
         plt.yticks(np.arange(len(depth)), depth)
         plt.title('Test Data')
         plt.show()
```



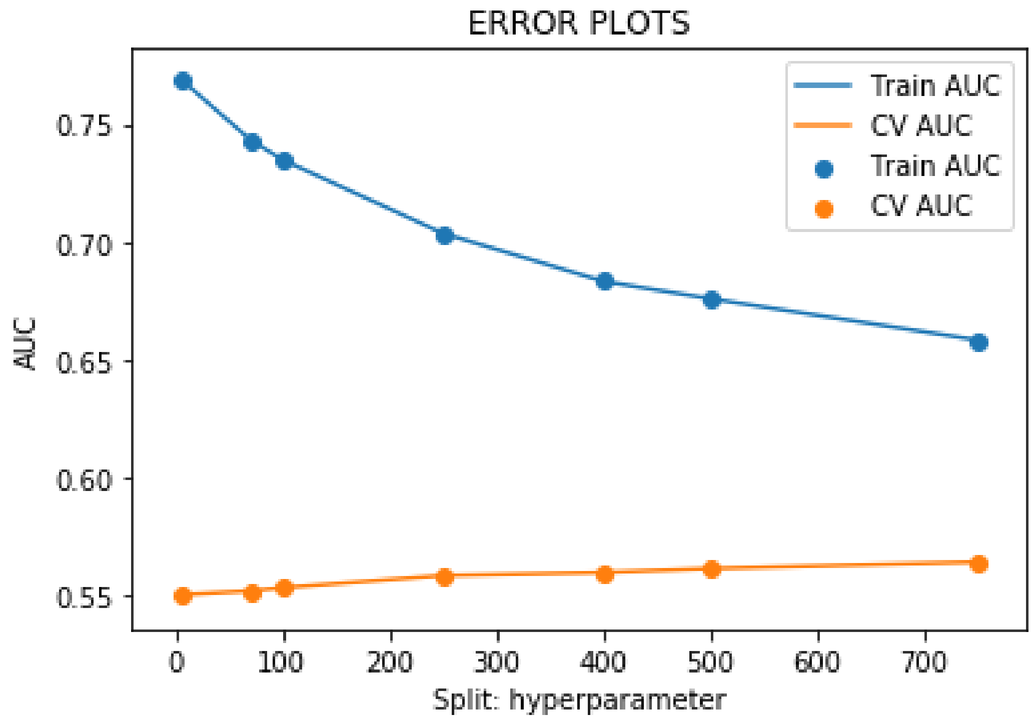
7.4 [5.4] Applying Decision Trees on TFIDF W2V, SET 4

7.4.1 Hyperparameter tuning using GridSearch

In [81]:
```
# clf = DecisionTreeClassifier()
# for Minimum samples split in Decision Tree
split = [5,70,100,250,400,500,750]
parameters = {'min_samples_split':[5,70,100,250,400,500,750]}
grid = GridSearchCV(DecisionTreeClassifier(class_weight ='balanced',max_depth=10), pa
grid.fit(tfidf_sent_vectors_train, Y_train)

print("best split = ", grid.best_params_)

train_auc_tfw2v = grid.cv_results_['mean_train_score']
cv_auc_tfw2v = grid.cv_results_['mean_test_score']

plt.plot(split, train_auc_tfw2v, label='Train AUC')
plt.scatter(split, train_auc_tfw2v, label='Train AUC')
plt.plot(split, cv_auc_tfw2v, label='CV AUC')
plt.scatter(split, cv_auc_tfw2v, label='CV AUC')
```

```
plt.legend()
plt.xlabel("Split: hyperparameter")
plt.ylabel("AUC")
plt.title("ERROR PLOTS")
plt.show()
```

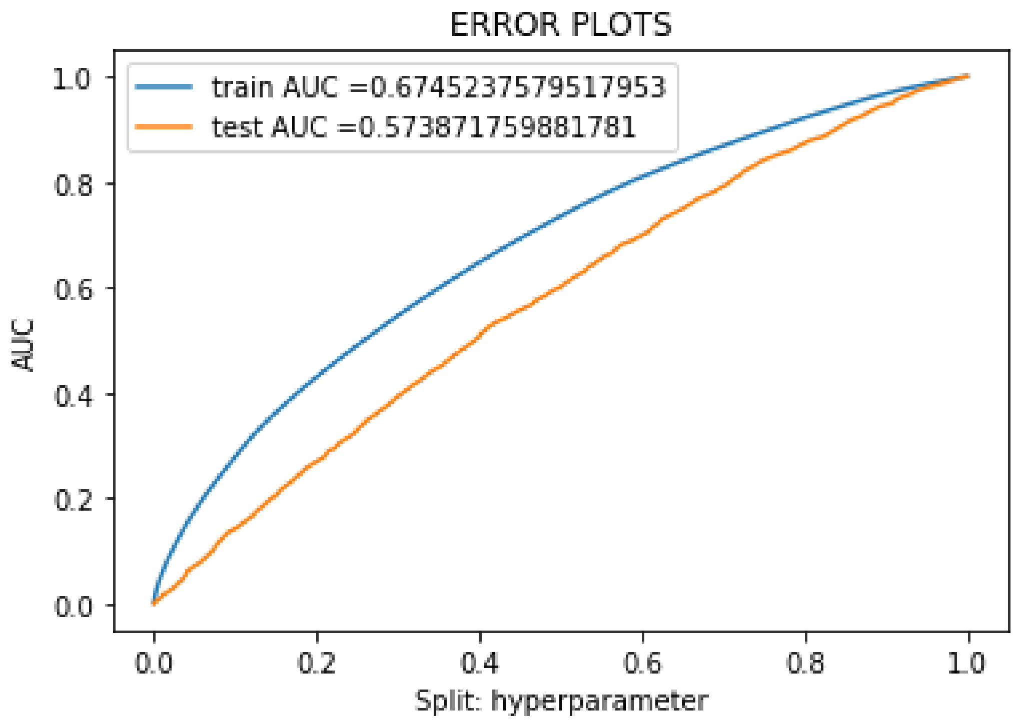
```
best split =  {'min_samples_split': 750}
```



7.4.2 Testing with Test Data

```
In [129]: clf = DecisionTreeClassifier(min_samples_split = 750, class_weight ='balanced')
          clf.fit(tfidf_sent_vectors_train, Y_train)

          # roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates o
          # not the predicted outputs

          train_fpr_tfw2v, train_tpr_tfw2v, thresholds_tfw2v = roc_curve(Y_train, clf.predict_
          test_fpr_tfw2v, test_tpr_tfw2v, thresholds_tfw2v = roc_curve(Y_test, clf.predict_pro

          plt.plot(train_fpr_tfw2v, train_tpr_tfw2v, label="train AUC ="+str(auc(train_fpr_tfw2
          plt.plot(test_fpr_tfw2v, test_tpr_tfw2v, label="test AUC ="+str(auc(test_fpr_tfw2v, 
          plt.legend()
```

```
plt.xlabel("Split: hyperparameter")
plt.ylabel("AUC")
plt.title("ERROR PLOTS")
plt.show()
```



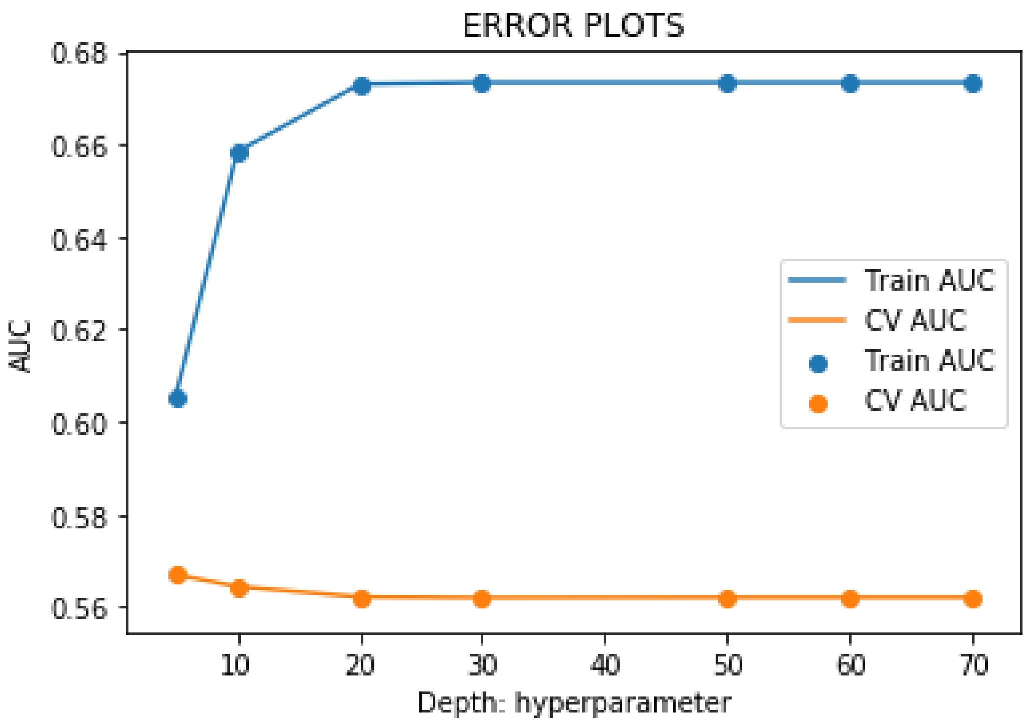

```
In [130]: tfw2v_split = auc(test_fpr_tfw2v, test_tpr_tfw2v)
          print(tfw2v_split)
```

```
0.573871759881781
```

```
In [85]: # clf = DecisionTreeClassifier()
         # for Maximum Depth in Decision Tree
         depth = [5,10,20,30,50,60,70]
         parameters = {'max_depth':[5,10,20,30,50,60,70]}
         grid = GridSearchCV(DecisionTreeClassifier(class_weight ='balanced',min_samples_split=
         grid.fit(tfidf_sent_vectors_train, Y_train)

         print("best depth = ", grid.best_params_)

         train_auc_tfw2v = grid.cv_results_['mean_train_score']
         cv_auc_tfw2v = grid.cv_results_['mean_test_score']

         plt.plot(depth, train_auc_tfw2v, label='Train AUC')
```

```
plt.scatter(depth, train_auc_tfw2v, label='Train AUC')
plt.plot(depth, cv_auc_tfw2v, label='CV AUC')
plt.scatter(depth, cv_auc_tfw2v, label='CV AUC')

plt.legend()
plt.xlabel("Depth: hyperparameter")
plt.ylabel("AUC")
plt.title("ERROR PLOTS")
plt.show()
```

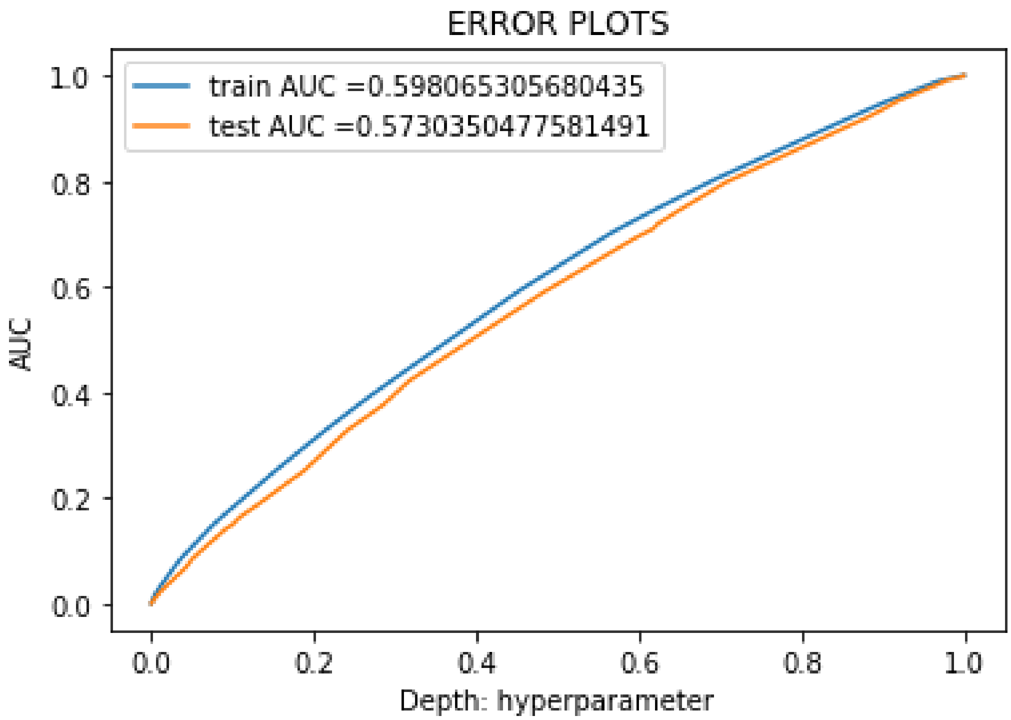
```
best depth =  {'max_depth': 5}
```



7.4.3 Testing with Test Data

```
In [131]: clf = DecisionTreeClassifier(max_depth = 5, class_weight ='balanced')
          clf.fit(tfidf_sent_vectors_train, Y_train)

          # roc_auc_score(y_true, y_score) the 2nd parameter should be probability estimates o
          # not the predicted outputs

          train_fpr_tfw2v, train_tpr_tfw2v, thresholds_tfw2v = roc_curve(Y_train, clf.predict_
          test_fpr_tfw2v, test_tpr_tfw2v, thresholds_tfw2v = roc_curve(Y_test, clf.predict_prol
```

```
plt.plot(train_fpr_tfw2v, train_tpr_tfw2v, label="train AUC ="+str(auc(train_fpr_tfw2
plt.plot(test_fpr_tfw2v, test_tpr_tfw2v, label="test AUC ="+str(auc(test_fpr_tfw2v, t
plt.legend()
plt.xlabel("Depth: hyperparameter")
plt.ylabel("AUC")
plt.title("ERROR PLOTS")
plt.show()
```



In [132]:
```
tfw2v_depth = auc(test_fpr_tfw2v, test_tpr_tfw2v)
print(tfw2v_depth)
```

```
0.5730350477581491
```

In [87]:
```
split = [5,70,100,250,400,500,750]
depth = [5,10,20,30,50,60,70]

parameters = {'min_samples_split': split,'max_depth': depth}
grid = GridSearchCV(DecisionTreeClassifier(class_weight ='balanced'), parameters, cv=3
grid.fit(tfidf_sent_vectors_train, Y_train)
```

Out[87]:
```
GridSearchCV(cv=3, error_score='raise',
       estimator=DecisionTreeClassifier(class_weight='balanced', criterion='gini',
            max_depth=None, max_features=None, max_leaf_nodes=None,
```

```
                     min_impurity_decrease=0.0, min_impurity_split=None,
                     min_samples_leaf=1, min_samples_split=2,
                     min_weight_fraction_leaf=0.0, presort=False, random_state=None,
                     splitter='best'),
                fit_params=None, iid=True, n_jobs=-1,
                param_grid={'min_samples_split': [5, 70, 100, 250, 400, 500, 750], 'max_depth'
                pre_dispatch='2*n_jobs', refit=True, return_train_score='warn',
                scoring='roc_auc', verbose=0)
```

```
In [88]: optimal_split = grid.best_estimator_.min_samples_split
         print("The optimal number of samples split is : ",optimal_split)

         optimal_depth = grid.best_estimator_.max_depth
         print("The optimal number of depth is : ",optimal_depth)
```

```
The optimal number of samples split is :  750
The optimal number of depth is :  5
```

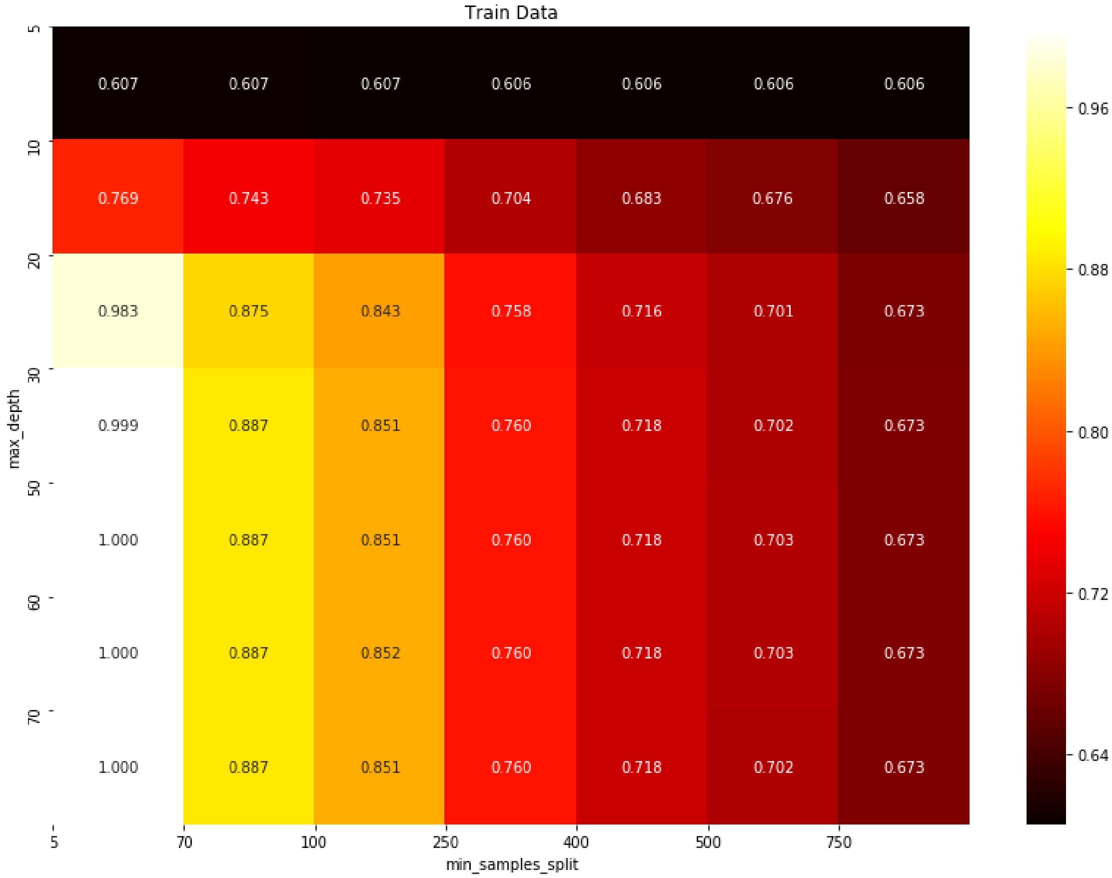
```
In [89]: clf = DecisionTreeClassifier(min_samples_split = optimal_split, max_depth = optimal_de
         clf.fit(tfidf_sent_vectors_train, Y_train)
         predw = clf.predict(tfidf_sent_vectors_test)

         accw = accuracy_score(Y_test, predw) * 100
         prew = precision_score(Y_test, predw) * 100
         recw = recall_score(Y_test, predw) * 100
         f1w = f1_score(Y_test, predw) * 100

         print('\nAccuracy=%f%%' % (accw))
         print('\nprecision=%f%%' % (prew))
         print('\nrecall=%f%%' % (recw))
         print('\nF1-Score=%f%%' % (f1w))
```

```

Accuracy=83.970074%

precision=84.001368%

recall=99.954792%

F1-Score=91.286307%
```

7.4.4 Heatmap for Train Data

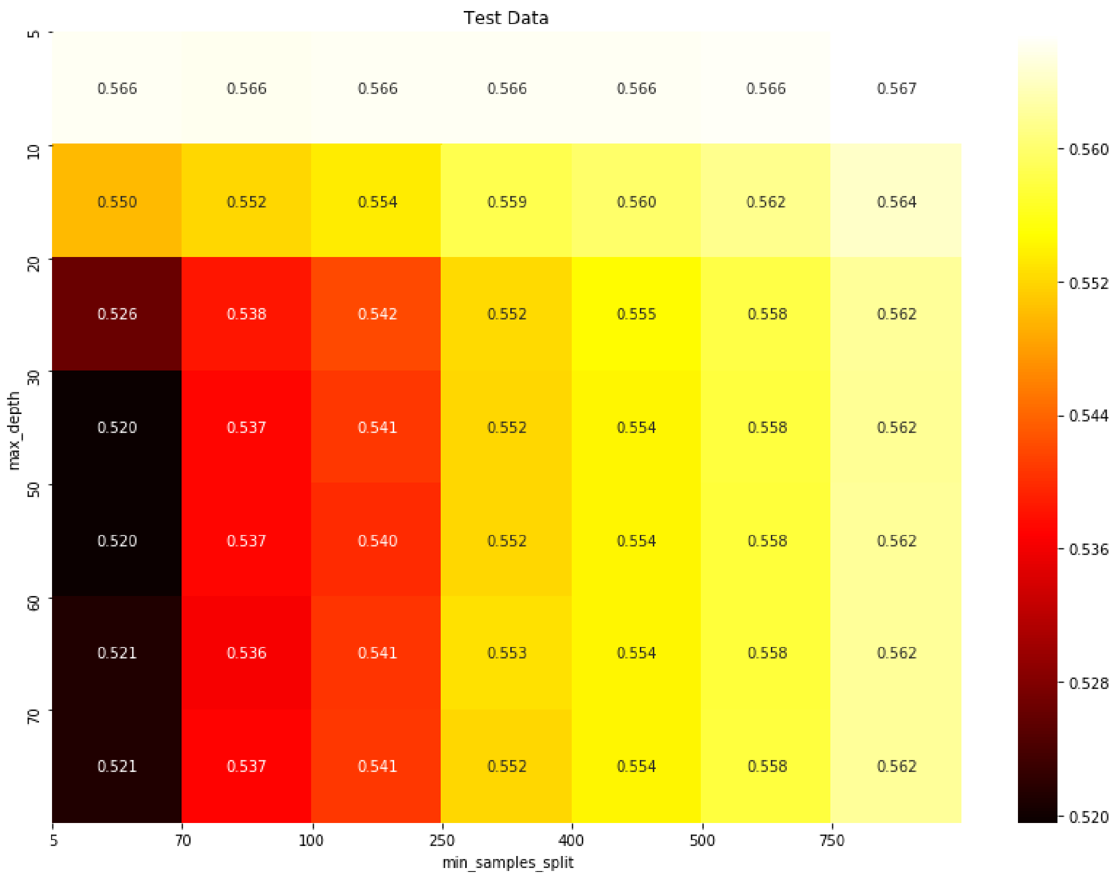
```
In [90]: scores = grid.cv_results_['mean_train_score'].reshape(len(split),len(depth))
         plt.figure(figsize=(14,10))
         sns.heatmap(scores, annot=True, cmap=plt.cm.hot, fmt=".3f", xticklabels=split, ytickla
         plt.xlabel('min_samples_split')
```

```
plt.ylabel('max_depth')
plt.xticks(np.arange(len(split)), split)
plt.yticks(np.arange(len(depth)), depth)
plt.title('Train Data')
plt.show()
```



7.4.5 Heatmap for Test Data

```
In [91]: scores = grid.cv_results_['mean_test_score'].reshape(len(split),len(depth))
         plt.figure(figsize=(14,10))
         sns.heatmap(scores, annot=True, cmap=plt.cm.hot, fmt=".3f", xticklabels=split, yticklabels=depth)
         plt.xlabel('min_samples_split')
         plt.ylabel('max_depth')
         plt.xticks(np.arange(len(split)), split)
         plt.yticks(np.arange(len(depth)), depth)
         plt.title('Test Data')
         plt.show()
```



8 [6] Conclusions

```
In [135]: # Please compare all your models using Prettytable library
          number= [1,2,3,4,5,6,7,8]
          name= ["Bow", "Bow", "Tfidf", "Tfidf", "Avg W2v", "Avg W2v", "Tfidf W2v", "Tfidf W2v
          metric= ["Depth", "Split", "Split", "Depth", "Depth", "Split", "Split", "Depth"]
          auc = [bow_depth, bow_split, tfidf_split, tfidf_depth, aw2v_depth, aw2v_split, tfw2v_
          acc= [accb,"SAME",acct,"SAME",acca,"SAME",accw,"SAME"]
          pre= [preb,"SAME",pret,"SAME",prea,"SAME",prew,"SAME"]

          #Initialize Prettytable
          ptable = PrettyTable()
          ptable.add_column("Index", number)
          ptable.add_column("Model", name)
          ptable.add_column("Hyperparameter", metric)
          ptable.add_column("AUC Score", auc)
          ptable.add_column("Accuracy%", acc)
          ptable.add_column("Precision%", pre)
```

```python
print(ptable)
```

Index	Model	Hyperparameter	AUC Score	Accuracy%	Precision%
1	Bow	Depth	0.7836237182492458	85.79674920249126	89.059843590615
2	Bow	Split	0.7883392075770768	SAME	SAME
3	Tfidf	Split	0.7549231271541399	85.56888956402857	88.18257607336
4	Tfidf	Depth	0.7424730576893023	SAME	SAME
5	Avg W2v	Depth	0.613469368282132	83.69284520735228	84.272860109205
6	Avg W2v	Split	0.610039091144049	SAME	SAME
7	Tfidf W2v	Split	0.573871759881781	83.97007443414857	84.001367729189
8	Tfidf W2v	Depth	0.5730350477581491	SAME	SAME

1. We have considered 100k data points
2. BOW and TFIDF have more accuracy value
3. BOW and TFIDF models have more AUC Score value.
4. BOW and TFIDF models are better than AW2V and TFIDF-W2V.