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]:

```
#mounting the dataset from drive
# from google.colab import drive
# drive.mount('/content/gdrive')

#connecting to sqlite db
con = sqlite3.connect('database.sqlite')

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

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

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

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

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

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

Out[3]:

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

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

In [6]:

```
# Removing duplicate reviews
final=filtered_data.drop_duplicates(subset={"UserId","ProfileName","Time","Text"}, keep='fi
print(final.shape)
```

```
(364173, 10)
```

In [7]:

```
(final['Id'].size*1.0)/(filtered_data['Id'].size*1.0)*100
```

Out[7]:

```
69.25890143662969
```

In [8]:

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

In [9]:

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

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

```
(364171, 10)
```

Out[9]:

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

In [10]:

```
final["cleanReview"] = final["Summary"].map(str) + ". " + final["Text"]
final['cleanReview'].head()
```

Out[10]:

```
0    Good Quality Dog Food. I have bought several o...
1    Not as Advertised. Product arrived labeled as ...
2    "Delight" says it all. This is a confection th...
3    Cough Medicine. If you are looking for the sec...
4    Great taffy. Great taffy at a great price.  Th...
Name: cleanReview, dtype: object
```

In [11]:

```
final['lengthOfReview'] = final['cleanReview'].str.split().str.len()
final['lengthOfReview'].head()
```

Out[11]:

```
0    52
1    34
2    98
3    43
4    29
Name: lengthOfReview, dtype: int64
```

In [10]:

```
#remove urls from text python
from tqdm import tqdm
lst = []
removed_urls_list = []
for text in tqdm(final['Text']):
  removed_urls_text = re.sub(r"http\S+", "", text)
  lst.append(removed_urls_text)
```

100%|██████████| 364171/364171 [00:00<00:00, 447313.57it/s]

In [11]:

```
#remove urls from text python
removed_urls_list = []
for text in tqdm(lst):
  removed_urls_text = re.sub(r"http\S+", "", text)
  removed_urls_list.append(removed_urls_text)
```

100%|██████████| 364171/364171 [00:00<00:00, 452270.97it/s]

In [12]:

```
from bs4 import BeautifulSoup
text_lst = []
for text in tqdm(removed_urls_list):
  soup = BeautifulSoup(text, 'lxml')
  text = soup.get_text()
  text_lst.append(text)
# print(text)
# print("="*50)
```

100%|██████████| 364171/364171 [01:49<00:00, 3330.00it/s]

In [13]:

```
print(len(final['Text']))
```

364171

In [14]:

```
# 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 [15]:

```
decat_lst = []
for decat_text in tqdm(text_lst):
  text = decontracted(decat_text)
  decat_lst.append(text)
```

100%|██████████| 364171/364171 [00:05<00:00, 65510.16it/s]

In [16]:

```
strip_list = []
for to_strip in tqdm(decat_lst):
  text = re.sub("\S*\d\S*", "", to_strip).strip()
  strip_list.append(text)
```

100%|██████████| 364171/364171 [00:22<00:00, 16465.51it/s]

In [17]:

```
spatial_list = []
for to_spatial in tqdm(strip_list):
  text = re.sub('[^A-Za-z0-9]+', ' ', to_spatial)
  spatial_list.append(text)
```

100%|██████████| 364171/364171 [00:12<00:00, 29401.19it/s]

In [18]:

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

In [19]:

```
# Combining all the above stundents
preprocessed_reviews = []
# tqdm is for printing the status bar
for sentance in tqdm(spatial_list):
    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 stopwords)
    preprocessed_reviews.append(sentance.strip())
```

100%|██████████| 364171/364171 [02:44<00:00, 2216.92it/s]

In [20]:

```
print(len(preprocessed_reviews))
preprocessed_reviews[-1]
```

364171

Out[20]:

'satisfied product advertised use cereal raw vinegar general sweetner'

In [21]:

```
final['Preprocessed_text'] = preprocessed_reviews
```

In [22]:

```
print(len(final))
final.tail(5)
```

364171

Out[22]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator
525809	568450	B001EO7N10	A28KG5XORO54AY	Lettie D. Carter	0
525810	568451	B003S1WTCU	A3I8AFVPEE8KI5	R. Sawyer	0
525811	568452	B004I613EE	A121AA1GQV751Z	pksd "pk_007"	2
525812	568453	B004I613EE	A3IBEVCTXKNOH	Kathy A. Welch "katwel"	1
525813	568454	B001LR2CU2	A3LGQPJCZVL9UC	srfell17	0

In [3]:

```
dir_path = os.getcwd()
conn = sqlite3.connect(os.path.join(dir_path, 'final.sqlite'))
# final.to_sql('Reviews', conn, if_exists='replace', index=False)
```

In [4]:

```
review_3 = pd.read_sql_query(""" SELECT count(*) FROM Reviews""", conn)
print(review_3)
```

```
   count(*)
0    364171
```

In [5]:

```
filtered_data = pd.read_sql_query(""" SELECT * FROM Reviews""", conn)
```

In [6]:

```
filtered_data.shape
```

Out[6]:

```
(364171, 12)
```

In [7]:

```
filtered_data["Time"] = pd.to_datetime(filtered_data["Time"], unit = "s")
filtered_data = filtered_data.sort_values(by = "Time")
```

In [8]:

```
filtered_data.head(5)
```

Out[8]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator
117924	150524	0006641040	ACITT7DI6IDDL	shari zychinski	0
117901	150501	0006641040	AJ46FKXOVC7NR	Nicholas A Mesiano	2
298792	451856	B00004CXX9	AIUWLEQ1ADEG5	Elizabeth Medina	0
169281	230285	B00004RYGX	A344SMIA5JECGM	Vincent P. Ross	1
298791	451855	B00004CXX9	AJH6LUC1UT1ON	The Phantom of the Opera	0

In [9]:

```
print(len(filtered_data))
filtered_data.info()
filtered_data = filtered_data.head(100000)
print(len(filtered_data))
```

```
364171
<class 'pandas.core.frame.DataFrame'>
Int64Index: 364171 entries, 117924 to 107253
Data columns (total 12 columns):
Id                      364171 non-null int64
ProductId               364171 non-null object
UserId                  364171 non-null object
ProfileName             364171 non-null object
HelpfulnessNumerator    364171 non-null int64
HelpfulnessDenominator  364171 non-null int64
Score                   364171 non-null int64
Time                    364171 non-null datetime64[ns]
Summary                 364171 non-null object
Text                    364171 non-null object
cleanReview             364171 non-null object
lengthOfReview          364171 non-null int64
dtypes: datetime64[ns](1), int64(5), object(6)
memory usage: 36.1+ MB
100000
```

In [10]:

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

Out[10]:

```
1    87729
0    12271
Name: Score, dtype: int64
```

In [11]:

```
X = filtered_data["cleanReview"]
print(print("shape of X:", X.head(5)))
y = filtered_data["Score"]
print("shape of y:", y.head(5))
X_len = filtered_data['lengthOfReview']
```

```
shape of X: 117924    every book educational witty little book makes...
117901    whole series great way spend time child rememb...
298792    entertainingl funny beetlejuice well written m...
169281    modern day fairy tale twist rumplestiskin capt...
298791    fantastic beetlejuice excellent funny movie ke...
Name: cleanReview, dtype: object
None
shape of y: 117924    1
117901    1
298792    1
169281    1
298791    1
Name: Score, dtype: int64
```

In [12]:

```
len(filtered_data['lengthOfReview'])
```

Out[12]:

```
100000
```

In [13]:

```
X_train = X[0:60000]
Y_train = y[0:60000]
X_val = X[60000:80000]
Y_val = y[60000:80000]
X_test = X[80000:100000]
Y_test = y[80000:100000]
```

In [14]:

```
print(len(X_train), len(X_test), len(X_val))
print(len(Y_train), len(Y_test), len(Y_val))
```

```
60000 20000 20000
60000 20000 20000
```

[4.1] BAG OF WORDS

In [87]:

```
from sklearn.feature_extraction.text import CountVectorizer

count_vect = CountVectorizer()
X_train_vect = count_vect.fit_transform(X_train)
X_test_vect = count_vect.transform(X_test)
X_val_vect = count_vect.transform(X_val)
feature_names = count_vect.get_feature_names()
# BoW_dict = {'X_train_vect':X_train_vect, 'X_test_vect': X_test_vect, 'X_val_vect': X_val_
print(X_train_vect.shape)
# print(feature_names)
```

```
(60000, 47535)
```

In [25]:

```
X_train_vect.shape
```

Out[25]:

```
(60000, 47535)
```

In [26]:

```
len(final['lengthOfReview'])
```

Out[26]:

```
364171
```

In [27]:

```
from scipy.sparse import hstack
# len_review = final['lengthOfReview'].to_sparse()
concat_data = hstack((X_train_vect,np.array(final['lengthOfReview'][0:60000])[:,None]))
concat_data_val = hstack((X_val_vect,np.array(final['lengthOfReview'][60000:80000])[:,None]
concat_data_test = hstack((X_test_vect,np.array(final['lengthOfReview'][80000:100000])[:,No
```

In [28]:

```
print(concat_data.shape)
print(concat_data_val.shape)
print(concat_data_test.shape)
```

```
(60000, 47536)
(20000, 47536)
(20000, 47536)
```

In [29]:

```
print(len(feature_names))
```

```
47535
```

In [30]:

```
BoW_dict = {'X_train_vect':concat_data, 'X_test_vect': concat_data_test, 'X_val_vect': conc
print(BoW_dict['X_train_vect'].shape)
```

```
(60000, 47536)
```

In [ ]:

```
import pickle
with open('BoW.pkl', 'wb') as handle:
    pickle.dump(BoW_dict, handle, protocol=pickle.HIGHEST_PROTOCOL)
```

[4.3] TF-IDF

In [149]:

```
tf_idf_vect = TfidfVectorizer(ngram_range=(1,2), min_df=10)
train_tf_idf = tf_idf_vect.fit_transform(X_train)
cv_tf_idf = tf_idf_vect.transform(X_val)
test_tf_idf = tf_idf_vect.transform(X_test)

print("the shape of out text TFIDF vectorizer ",train_tf_idf.get_shape())
print("the type of count vectorizer ",type(train_tf_idf))
print("the number of unique words including both unigrams and bigrams ", train_tf_idf.get_s
```

```
the shape of out text TFIDF vectorizer  (60000, 35873)
the type of count vectorizer  <class 'scipy.sparse.csr.csr_matrix'>
the number of unique words including both unigrams and bigrams  35873
```

In [32]:

```
tfidf_concat_data_train = hstack((train_tf_idf,np.array(final['lengthOfReview'][0:60000])[:
tfidf_concat_data_val = hstack((cv_tf_idf,np.array(final['lengthOfReview'][60000:80000])[:,
tfidf_concat_data_test = hstack((test_tf_idf,np.array(final['lengthOfReview'][80000:100000]
```

In [33]:

```
tf_idf_dict = {'train_tf_idf': tfidf_concat_data_train, 'cv_tf_idf': tfidf_concat_data_val,
```

In [ ]:

```
import pickle
with open('tf_idf.pkl', 'wb') as handle:
    pickle.dump(tf_idf_dict, handle, protocol=pickle.HIGHEST_PROTOCOL)
```

[4.4] Word2Vec

In [34]:

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

In [35]:

```
is_your_ram_gt_16g=False
want_to_use_google_w2v = False
want_to_train_w2v = True

if want_to_train_w2v:
    # min_count = 5 considers only words that occured atleast 5 times
    w2v_model=Word2Vec(list_of_sen,min_count=5,size=50, workers=4)
    print(w2v_model.wv.most_similar('great'))
    print('='*50)
    print(w2v_model.wv.most_similar('worst'))

elif want_to_use_google_w2v and is_your_ram_gt_16g:
    if os.path.isfile('GoogleNews-vectors-negative300.bin'):
        w2v_model=KeyedVectors.load_word2vec_format('GoogleNews-vectors-negative300.bin', b
        print(w2v_model.wv.most_similar('great'))
        print(w2v_model.wv.most_similar('worst'))
    else:
        print("you don't have gogole's word2vec file, keep want_to_train_w2v = True, to tra
```

```
[('terrific', 0.8565828204154968), ('excellent', 0.8381140828132629), ('fant
astic', 0.8366681337356567), ('awesome', 0.7857832908630371), ('wonderful',
0.7829444408416748), ('good', 0.742619514465332), ('perfect', 0.717479586601
2573), ('nice', 0.6593438386917114), ('fabulous', 0.6570981740951538), ('inc
redible', 0.6524804830551147)]
==================================================
[('greatest', 0.7822151780128479), ('best', 0.7523022294044495), ('tasties
t', 0.6484744548797607), ('coolest', 0.6170215606689453), ('terrible', 0.612
8978729248047), ('awful', 0.6031897664070129), ('nicest', 0.598495066165924
1), ('nastiest', 0.5957451462745667), ('closest', 0.5847468376159668), ('sof
test', 0.5774857401847839)]
```

In [36]:

```
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  15289
sample words  ['flat', 'mater', 'elements', 'crock', 'tripe', 'reversed', 'l
actaid', 'capsule', 'easiest', 'clarify', 'pees', 'swore', 'similiar', 'powd
ery', 'cement', 'deb', 'burned', 'seasonally', 'stove', 'reinforcement', 'co
nfusion', 'sky', 'mama', 'evil', 'contrast', 'start', 'booklet', 'moves', 'c
hestnuts', 'virtuous', 'monitors', 'twain', 'liquified', 'recommendations',
'quinoa', 'micro', 'corned', 'celebrated', 'pitcher', 'clip', 'movie', 'hfc
s', 'single', 'leftover', 'inhaled', 'impulse', 'leak', 'gag', 'farming', 'b
razilian']
```

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

[4.4.1.1] Avg W2v

In [37]:

```
print(X_train[117924])
print(len(X_val))
print(len(X_test))
```

```
every book educational witty little book makes son laugh loud recite car dri
ving along always sing refrain learned whales india drooping roses love new
words book introduces silliness classic book willing bet son still able reci
te memory college
20000
20000
```

In [38]:

```
# average Word2Vec
# compute average word2vec for each review.
def avg_w2vec(sentences_received):
    sent_vectors = []; # the avg-w2v for each sentence/review is stored in this list
    for sent in tqdm(sentences_received): # for each review/sentence
        sent_vec = np.zeros(50) # as word vectors are of zero length 50, you might need to
        cnt_words =0; # num of words with a valid vector in the sentence/review
        for word in sent: # for each word in a review/sentence
            if word in w2v_words:
                vec = w2v_model.wv[word]
                sent_vec += vec
                cnt_words += 1
        if cnt_words != 0:
            sent_vec /= cnt_words
        sent_vectors.append(sent_vec)

    print(len(sent_vectors))
    print(len(sent_vectors[0]))
    return sent_vectors
```

In [39]:

```
print(len([sent.split() for sent in X_test]))
```

```
20000
```

In [22]:

```
avg_w2v_train = avg_w2vec([sent.split() for sent in X_train])
avg_w2v_cv = avg_w2vec([sent.split() for sent in X_val])
avg_w2v_test = avg_w2vec([sent.split() for sent in X_test])
```

In [ ]:

```
Avg_w2v_dict = {'X_train_avgw2v':avg_w2v_train, 'Y_train_avgw2v': Y_train,
                    'X_val_avgw2v': avg_w2v_cv, 'Y_val_avgw2v': Y_val,
                   'X_test_avgw2v': avg_w2v_test, 'Y_test_avgw2v': Y_test}
```

In [ ]:

```
import pickle
with open('/content/gdrive/My Drive/Colab Notebooks/Assignment 3/avg_w2v.pkl', 'wb') as har
    pickle.dump(Avg_w2v_dict, handle, protocol=pickle.HIGHEST_PROTOCOL)
```

[4.4.1.2] TFIDF weighted W2v

In [79]:

```
# S = ["abc def pqr", "def def def abc", "pqr pqr def"]
model = TfidfVectorizer()
tf_idf_matrix = 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 [ ]:

```
# 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

def tfidf_w2v(sentences_received):
    tfidf_sent_vectors = []; # the tfidf-w2v for each sentence/review is stored in this lis
    row=0;
    for sent in tqdm(sentences_received): # for each review/sentence
        sent_vec = np.zeros(50) # as word vectors are of zero length
        weight_sum =0; # num of words with a valid vector in the sentence/review
        for word in sent: # for each word in a review/sentence
            if word in w2v_words and word in tfidf_feat:
                vec = w2v_model.wv[word]
    #             tf_idf = tf_idf_matrix[row, tfidf_feat.index(word)]
                # to reduce the computation we are
                # dictionary[word] = idf value of word in whole courpus
                # sent.count(word) = tf valeus of word in this review
                tf_idf = dictionary[word]*(sent.count(word)/len(sent))
                sent_vec += (vec * tf_idf)
                weight_sum += tf_idf
        if weight_sum != 0:
            sent_vec /= weight_sum
        tfidf_sent_vectors.append(sent_vec)
        row += 1

    return tfidf_sent_vectors
```

In [73]:

```
tfidf_w2v_train = tfidf_w2v([sent.split() for sent in X_train])
tfidf_w2v_cv = tfidf_w2v([sent.split() for sent in X_val])
tfidf_w2v_test = tfidf_w2v([sent.split() for sent in X_test])
```

In [74]:

```
tfidf_w2v_dict = {'X_train_tfidfw2v':tfidf_w2v_train, 'Y_train_tfidfw2v': Y_train,
                    'X_val_tfidfw2v': tfidf_w2v_cv, 'Y_val_tfidfw2v': Y_val,
                   'X_test_tfidfw2v': tfidf_w2v_test, 'Y_test_tfidfw2v': Y_test}
```

In [75]:

```
with open('tfidf_w2v.pkl', 'wb') as handle:
    pickle.dump(tfidf_w2v_dict, handle, protocol=pickle.HIGHEST_PROTOCOL)
```

In [104]:

```
from sklearn import tree
from tqdm import tqdm
```

Decision Trees on BoW

In [105]:

```
import pickle
# with open(r"/content/gdrive/My Drive/Colab Notebooks/Assignment 4/BoW.pkl", "rb") as inpu
with open(r"BoW.pkl", "rb") as input_file:
    BoW_dict = pickle.load(input_file)
```

In [106]:

```
print(type(BoW_dict['X_train_vect']))
print(type(BoW_dict['X_val_vect']))
print(type(X_train))
print(type(X_val))
```

```
<class 'scipy.sparse.coo.coo_matrix'>
<class 'scipy.sparse.coo.coo_matrix'>
<class 'pandas.core.series.Series'>
<class 'pandas.core.series.Series'>
```

In [107]:

```
from scipy.sparse import vstack
X_train_val = vstack((BoW_dict['X_train_vect'], BoW_dict['X_val_vect']))
```

In [108]:

```
Y_train_val = pd.concat([Y_train, Y_val], axis= 0)
```

In [109]:

```
Y_train_val.shape
```

Out[109]:

```
(80000,)
```

In [110]:

```
fpr_val = dict()
tpr_val = dict()
roc_auc_val = dict()
fpr_train = dict()
tpr_train = dict()
roc_auc_train = dict()
from sklearn.model_selection import GridSearchCV
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import roc_auc_score

param_grid = {'max_depth': [1, 5, 10, 50, 100, 500, 1000], 'min_samples_split' : [5, 10, 10

grid_search_cv = GridSearchCV(DecisionTreeClassifier(), param_grid, scoring = 'roc_auc')
grid_search_cv.fit(X_train_val,Y_train_val)
```

Out[110]:

```
GridSearchCV(cv='warn', error_score='raise-deprecating',
       estimator=DecisionTreeClassifier(class_weight=None, 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='warn', n_jobs=None,
       param_grid={'max_depth': [1, 5, 10, 50, 100, 500, 1000], 'min_samples
_split': [5, 10, 100, 500]},
       pre_dispatch='2*n_jobs', refit=True, return_train_score='warn',
       scoring='roc_auc', verbose=0)
```

In [111]:

```
f = grid_search_cv.cv_results_
f
```

Out[111]:

```
{'mean_fit_time': array([ 0.96633339,  0.92899998,  0.93066676,  1.0773332
9,  1.52200007,
         1.65700006,  1.421     ,  1.52700003,  3.5893333 ,  3.20966665,
         2.89333336,  2.16833337, 18.04499992, 18.14166665, 13.48433328,
         9.43799996, 24.76233331, 23.1813333 , 19.09399994, 14.09299986,
        28.45566662, 27.6346666 , 25.32966669, 18.69333339, 27.81133326,
        28.17633335, 23.01066669, 18.18166669]),
 'mean_score_time': array([0.03733341, 0.03200006, 0.03066667, 0.03433339,
0.205     ,
        0.0326666 , 0.03066667, 0.05566669, 0.04100005, 0.0316666 ,
        0.04766663, 0.0333333 , 0.03766672, 0.03866673, 0.03766672,
        0.03733325, 0.03966665, 0.04000004, 0.03700002, 0.0356667 ,
        0.04233336, 0.045     , 0.04133336, 0.04066666, 0.04166667,
        0.04166667, 0.04100005, 0.04033335]),
 'mean_test_score': array([0.63281207, 0.63281207, 0.63281207, 0.63281207,
0.70339984,
        0.70391822, 0.70437949, 0.70495083, 0.78224554, 0.78540815,
        0.7931347 , 0.79645109, 0.69934434, 0.72244643, 0.8091168 ,
```

In [112]:

```
import plotly.offline as offline
import plotly.graph_objs as go
offline.init_notebook_mode()
import numpy as np
```

In [121]:

```
x1_list = []
x2_list = []
for c1 in grid_search_cv.cv_results_['params']:
    x1_list.append(c1['max_depth'])
for c2 in grid_search_cv.cv_results_['params']:
    x2_list.append(c2['min_samples_split'])
print(x1_list, x2_list)
```

```
[1, 1, 1, 1, 5, 5, 5, 5, 10, 10, 10, 10, 50, 50, 50, 50, 100, 100, 100, 100,
500, 500, 500, 500, 1000, 1000, 1000, 1000] [5, 10, 100, 500, 5, 10, 100, 50
0, 5, 10, 100, 500, 5, 10, 100, 500, 5, 10, 100, 500, 5, 10, 100, 500, 5, 1
0, 100, 500]
```

In [122]:

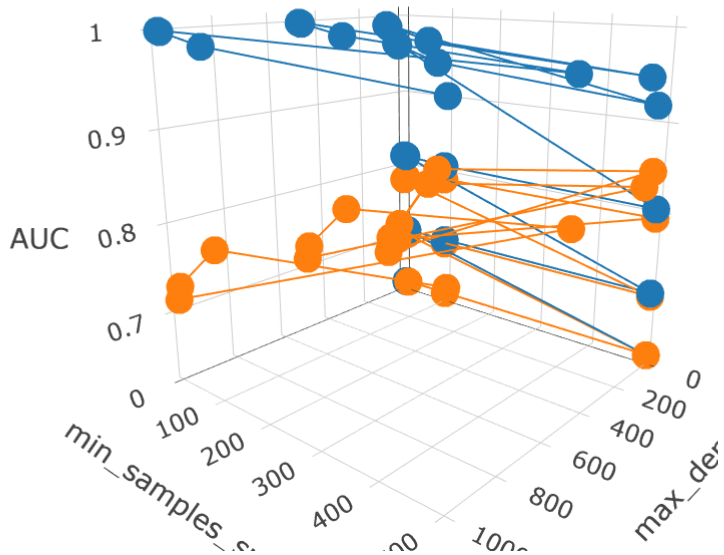
```
x1 = x1_list
y1 = x2_list
z1 = grid_search_cv.cv_results_['mean_train_score'].tolist()
x2 = x1_list
y2 = x2_list
z2 = grid_search_cv.cv_results_['mean_test_score'].tolist()
```

In [123]:

```
# https://plot.ly/python/3d-axes/
trace1 = go.Scatter3d(x=x1,y=y1,z=z1, name = 'train')
trace2 = go.Scatter3d(x=x2,y=y2,z=z2, name = 'Cross validation')
data = [trace1, trace2]

layout = go.Layout(scene = dict(
        xaxis = dict(title='max_depth'),
        yaxis = dict(title='min_samples_split'),
        zaxis = dict(title='AUC'),))

fig = go.Figure(data=data, layout=layout)
offline.iplot(fig, filename='3d-scatter-colorscale')
```



In [124]:

```
grid_search_cv.best_params_
```

Out[124]:

```
{'max_depth': 50, 'min_samples_split': 500}
```

In [125]:

```
best_max_depth = grid_search_cv.best_params_['max_depth']
best_min_samples_split = grid_search_cv.best_params_['min_samples_split']
```

In [126]:

```
dt_clf = tree.DecisionTreeClassifier(max_depth = best_max_depth, min_samples_split = best_m
dt_clf.fit(BoW_dict['X_train_vect'],Y_train)
bow_test_proba = dt_clf.predict_proba(BoW_dict['X_test_vect'])
bow_train_proba = dt_clf.predict_proba(BoW_dict['X_train_vect'])
bow_test_proba
```

Out[126]:

```
array([[0.00854399, 0.99145601],
       [0.07377883, 0.92622117],
       [0.03005115, 0.96994885],
       ...,
       [0.07103825, 0.92896175],
       [0.00854399, 0.99145601],
       [0.65240642, 0.34759358]])
```

In [127]:

```
print("Top 20 Important Features")
d = sorted(list(zip(count_vect.get_feature_names(), dt_clf.feature_importances_ )), key=lam
d
```

```
Top 20 Important Features
```

Out[127]:

```
[('not', 0.10687457837134103),
 ('great', 0.06338434822136996),
 ('best', 0.04854930314041147),
 ('disappointed', 0.04352055705128352),
 ('worst', 0.042072337203084946),
 ('delicious', 0.02987117822067617),
 ('awful', 0.0298089783097103),
 ('terrible', 0.028527502519896178),
 ('horrible', 0.027389934845085017),
 ('good', 0.024573346893783484),
 ('disappointing', 0.02112008865333756),
 ('money', 0.02111834328229383),
 ('love', 0.02014915914398918),
 ('yuck', 0.018208573613810192),
 ('excellent', 0.013169458248609245),
 ('nice', 0.012411715119656264),
 ('perfect', 0.011649620398568651),
 ('bad', 0.011574530085451049),
 ('threw', 0.011504957397600973),
 ('favorite', 0.011101257269373492)]
```

In [128]:

```
bow_fpr_train, bow_tpr_train, _ = roc_curve(Y_train, bow_train_proba[:, 1])
bow_fpr_test, bow_tpr_test, _ = roc_curve(Y_test, bow_test_proba[:, 1])
bow_test_auc = auc(bow_fpr_test, bow_tpr_test)
bow_train_auc = auc(bow_fpr_train, bow_tpr_train)
print(bow_test_auc)
print(bow_train_auc)
```

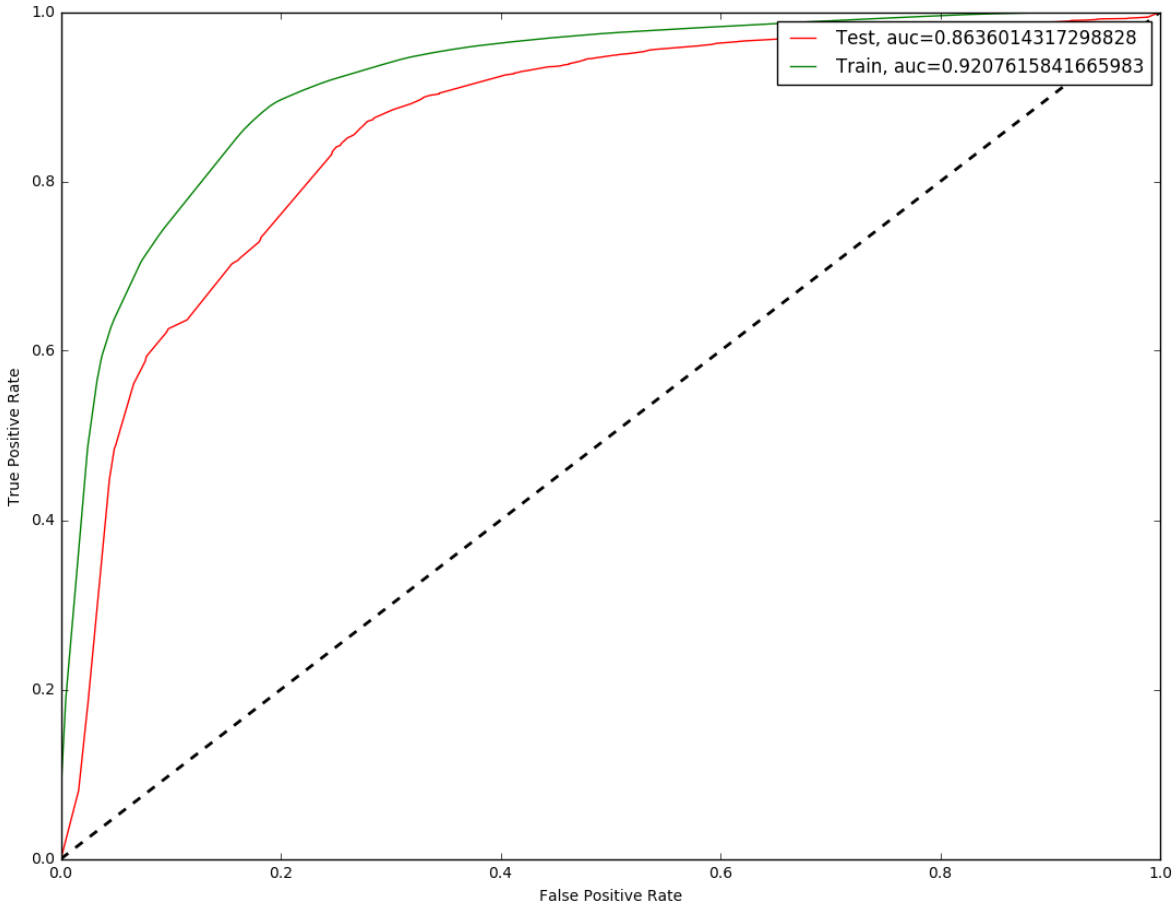
```
0.8636014317298828
0.9207615841665983
```

In [129]:

```
import pylab
plt.figure(figsize=(13, 10))
plt.plot([0,1], [0,1], color='black', lw=2, linestyle='--')
plt.plot(bow_fpr_test, bow_tpr_test, label="Test, auc="+str(bow_test_auc), color = 'red')
plt.plot(bow_fpr_train, bow_tpr_train, label="Train, auc="+str(bow_train_auc), color = 'gre

plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.legend()

plt.show()
```



In [130]:

```
bow_test_conf = dt_clf.predict(BoW_dict['X_test_vect'])
bow_train_conf = dt_clf.predict(BoW_dict['X_train_vect'])
```

In [131]:

```
from sklearn.metrics import classification_report, confusion_matrix
bow_train_conf_matrix = confusion_matrix(Y_train, bow_train_conf)
bow_test_conf_matrix = confusion_matrix(Y_test, bow_test_conf)
class_report = classification_report(Y_test, bow_test_conf)
print(bow_test_conf_matrix)
print(class_report)
```

```
[[ 1257  1419]
 [  809 16515]]
              precision    recall  f1-score   support

           0       0.61      0.47      0.53      2676
           1       0.92      0.95      0.94     17324

   micro avg       0.89      0.89      0.89     20000
   macro avg       0.76      0.71      0.73     20000
weighted avg       0.88      0.89      0.88     20000
```

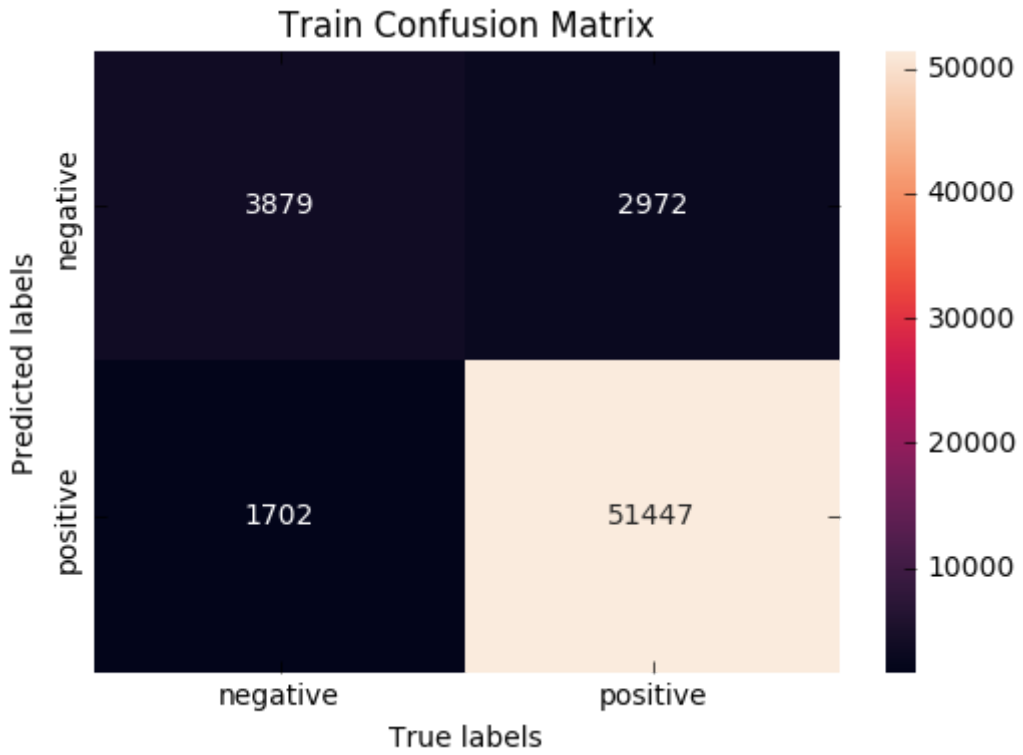
In [132]:

```
ax= plt.subplot()
sns.heatmap(bow_train_conf_matrix, annot=True, ax = ax, fmt='g')

ax.set_ylabel('Predicted labels')
ax.set_xlabel('True labels')
ax.set_title('Train Confusion Matrix')
ax.xaxis.set_ticklabels(['negative', 'positive'])
ax.yaxis.set_ticklabels(['negative', 'positive'])
```

Out[132]:

```
[<matplotlib.text.Text at 0x30a20ba8>, <matplotlib.text.Text at 0x30a41828>]
```



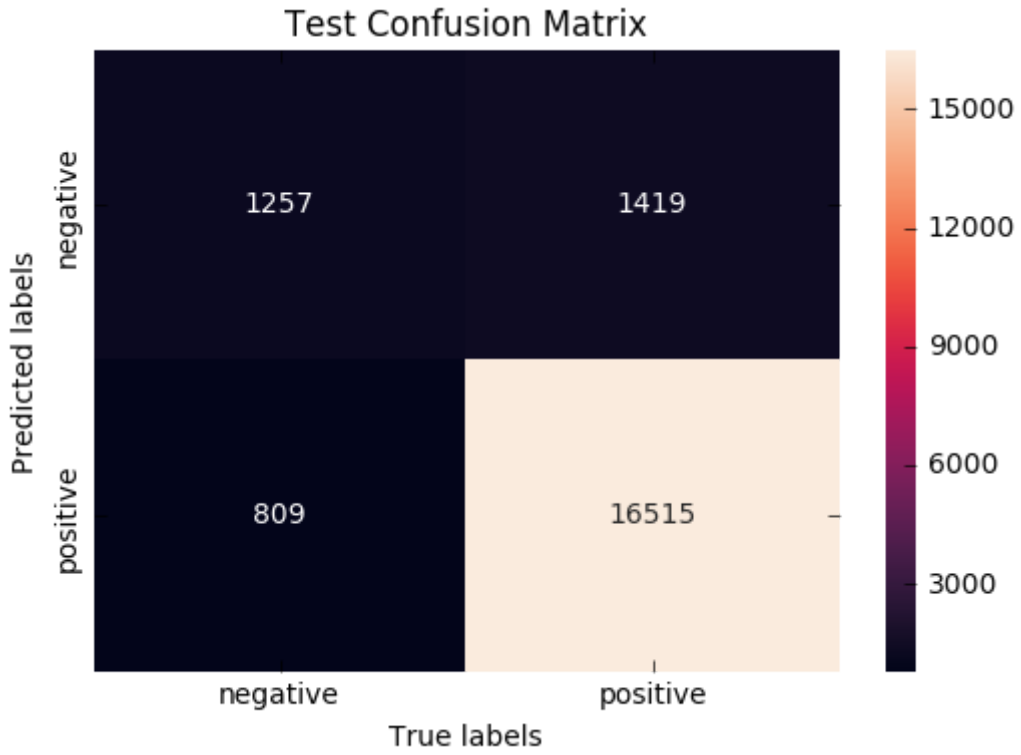
In [133]:

```
ax= plt.subplot()
sns.heatmap(bow_test_conf_matrix, annot=True, ax = ax, fmt='g')

ax.set_ylabel('Predicted labels')
ax.set_xlabel('True labels')
ax.set_title('Test Confusion Matrix')
ax.xaxis.set_ticklabels(['negative', 'positive'])
ax.yaxis.set_ticklabels(['negative', 'positive'])
```

Out[133]:

```
[<matplotlib.text.Text at 0x326e6470>, <matplotlib.text.Text at 0x285b8a20>]
```



Exporting Decision Tree

In [76]:

```
tclassifier = tree.DecisionTreeClassifier(max_depth = 3)
tclassifier.fit(BoW_dict['X_train_vect'],Y_train)
```

Out[76]:

```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=3,
            max_features=None, max_leaf_nodes=None, min_samples_leaf=1,
            min_samples_split=2, min_weight_fraction_leaf=0.0,
            presort=False, random_state=None, splitter='best')
```

In [ ]:

```
from graphviz import Source
import os
os.environ["PATH"] += os.pathsep + 'C:/Program Files (x86)/Graphviz2.38/bin/'
graph = Source(tree.export_graphviz(tclassifier, out_file='tree.dot', feature_names=count_v
png_bytes = graph.pipe(format='png')
with open('bow.png','wb') as f:
    f.write(png_bytes)
```

Decision Tree on Tf-IDF

In [134]:

```
import pickle
with open(r"tf_idf.pkl", "rb") as input_file:
    tfidf_dict = pickle.load(input_file)
```

In [136]:

```
from scipy.sparse import vstack
X_train_val_tfidf = vstack((tfidf_dict['train_tf_idf'], tfidf_dict['cv_tf_idf']))
```

In [137]:

```
print(X_train_val_tfidf.shape)
```

```
(80000, 35874)
```

In [138]:

```
fpr_val = dict()
tpr_val = dict()
roc_auc_val = dict()
fpr_train = dict()
tpr_train = dict()
roc_auc_train = dict()
from sklearn.model_selection import GridSearchCV
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import roc_auc_score

param_grid = {'max_depth': [1, 5, 10, 50, 100, 500, 1000], 'min_samples_split' : [5, 10, 10

grid_search_cv = GridSearchCV(DecisionTreeClassifier(), param_grid, scoring = 'roc_auc')
grid_search_cv.fit(X_train_val_tfidf,Y_train_val)
```

Out[138]:

```
GridSearchCV(cv='warn', error_score='raise-deprecating',
       estimator=DecisionTreeClassifier(class_weight=None, 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='warn', n_jobs=None,
       param_grid={'max_depth': [1, 5, 10, 50, 100, 500, 1000], 'min_samples
_split': [5, 10, 100, 500]},
       pre_dispatch='2*n_jobs', refit=True, return_train_score='warn',
       scoring='roc_auc', verbose=0)
```

In [139]:

```
tfidf_results = grid_search_cv.cv_results_
tfidf_results
```

```
        16.98333335, 34.21866663, 33.49666667, 31.33933338, 27.4000001 ,
        46.61499993, 44.38000011, 41.52200007, 37.21833332, 44.89000003,
        44.5370001 , 41.86866665, 38.72133327]),
 'mean_score_time': array([0.03966665, 0.03900003, 0.03799995, 0.03766672,
0.04233329,
        0.03833334, 0.04133328, 0.04766671, 0.04033327, 0.03966665,
        0.04133336, 0.04833325, 0.04733332, 0.04633339, 0.04733332,
        0.04499992, 0.04866672, 0.046     , 0.04833333, 0.04733324,
        0.05333336, 0.04966664, 0.05033326, 0.04866672, 0.04933325,
        0.04966664, 0.05033334, 0.04933341]),
 'mean_test_score': array([0.62235324, 0.62235324, 0.62235324, 0.62235324,
0.69014115,
        0.69019916, 0.69080329, 0.69275649, 0.79262171, 0.79642745,
        0.80523174, 0.80768587, 0.72815712, 0.74631251, 0.82040417,
        0.84364496, 0.71763471, 0.73329391, 0.79897981, 0.82920351,
        0.73094104, 0.74264047, 0.78786884, 0.80949783, 0.73286431,
        0.74282804, 0.78757356, 0.80936875]),
 'mean_train_score': array([0.62490502, 0.62490502, 0.62490502, 0.6249050
2, 0.69896613,
        0.69895793, 0.69863899, 0.69779619, 0.83069246, 0.82993477
```

In [141]:

```
tfidf_x_list = []
tfidf_y_list = []
for c1 in grid_search_cv.cv_results_['params']:
    tfidf_x_list.append(c1['max_depth'])
for c2 in grid_search_cv.cv_results_['params']:
    tfidf_y_list.append(c2['min_samples_split'])
print(tfidf_x_list, tfidf_y_list)
```

```
[1, 1, 1, 1, 5, 5, 5, 5, 10, 10, 10, 10, 50, 50, 50, 50, 100, 100, 100, 100,
500, 500, 500, 500, 1000, 1000, 1000, 1000] [5, 10, 100, 500, 5, 10, 100, 50
0, 5, 10, 100, 500, 5, 10, 100, 500, 5, 10, 100, 500, 5, 10, 100, 500, 5, 1
0, 100, 500]
```

In [142]:

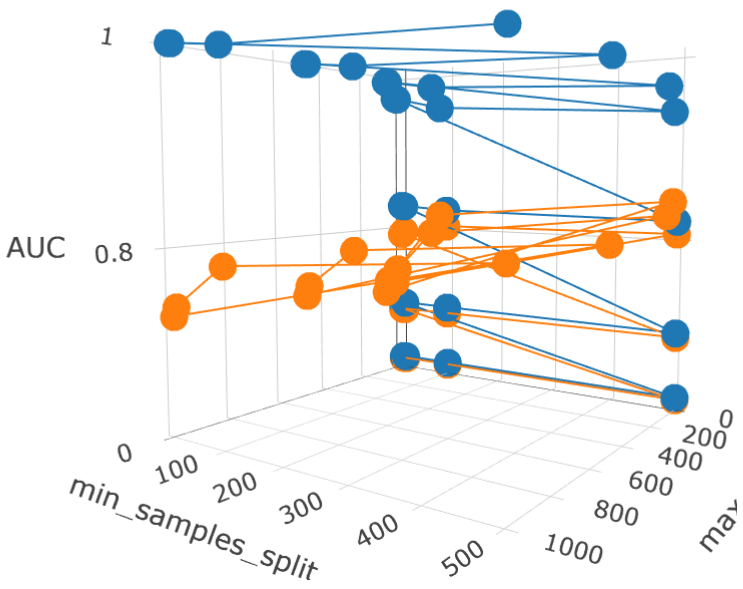
```
x1 = tfidf_x_list
y1 = tfidf_y_list
z1 = grid_search_cv.cv_results_['mean_train_score'].tolist()
x2 = tfidf_x_list
y2 = tfidf_y_list
z2 = grid_search_cv.cv_results_['mean_test_score'].tolist()
```

In [143]:

```
# https://plot.ly/python/3d-axes/
trace1 = go.Scatter3d(x=x1,y=y1,z=z1, name = 'train')
trace2 = go.Scatter3d(x=x2,y=y2,z=z2, name = 'Cross validation')
data = [trace1, trace2]

layout = go.Layout(scene = dict(
        xaxis = dict(title='max_depth'),
        yaxis = dict(title='min_samples_split'),
        zaxis = dict(title='AUC'),))

fig = go.Figure(data=data, layout=layout)
offline.iplot(fig, filename='3d-scatter-colorscale')
```



In [144]:

```
grid_search_cv.best_params_
```

Out[144]:

```
{'max_depth': 50, 'min_samples_split': 500}
```

In [145]:

```
tfidf_best_max_depth = grid_search_cv.best_params_['max_depth']
tfidf_best_min_samples_split = grid_search_cv.best_params_['min_samples_split']
```

In [152]:

```
tfidf_clf = tree.DecisionTreeClassifier(max_depth = tfidf_best_max_depth, min_samples_split
tfidf_clf.fit(tfidf_dict['train_tf_idf'],Y_train)
tfidf_train_proba = tfidf_clf.predict_proba(tfidf_dict['train_tf_idf'])
tfidf_test_proba = tfidf_clf.predict_proba(tfidf_dict['test_tf_idf'])
tfidf_test_proba
```

Out[152]:

```
array([[0.0115989 , 0.9884011 ],
       [0.07651568, 0.92348432],
       [0.        , 1.        ],
       ...,
       [0.08144796, 0.91855204],
       [0.0115989 , 0.9884011 ],
       [0.70638298, 0.29361702]])
```

In [150]:

```
print("Top 20 Important Features")
d = sorted(list(zip(tf_idf_vect.get_feature_names(), tfidf_clf.feature_importances_ )), key
d
```

Top 20 Important Features

Out[150]:

```
[('not', 0.1296138926399141),
 ('great', 0.05231338165228872),
 ('awful', 0.03719177884398404),
 ('disappointed', 0.03517821183427467),
 ('best', 0.03338195949642611),
 ('horrible', 0.03128741452304388),
 ('worst', 0.030493781340716446),
 ('terrible', 0.026043480326384608),
 ('delicious', 0.022185571838568684),
 ('disappointing', 0.019670939228679133),
 ('not worth', 0.017155293324281652),
 ('love', 0.015536469454621445),
 ('not disappointed', 0.014694171082176837),
 ('waste money', 0.014217266723955653),
 ('not buy', 0.014178367103307324),
 ('good', 0.013679512890012032),
 ('yuck', 0.013235576748896928),
 ('poor', 0.012219830938931241),
 ('threw', 0.009617523560184854),
 ('not great', 0.009591566037252981)]
```

In [153]:

```
tfidf_fpr_train, tfidf_tpr_train, _ = roc_curve(Y_train, tfidf_train_proba[:, 1])
tfidf_fpr_test, tfidf_tpr_test, _ = roc_curve(Y_test, tfidf_test_proba[:, 1])
tfidf_test_auc = auc(tfidf_fpr_test, tfidf_tpr_test)
tfidf_train_auc = auc(tfidf_fpr_train, tfidf_tpr_train)
print(tfidf_test_auc)
print(tfidf_train_auc)
```

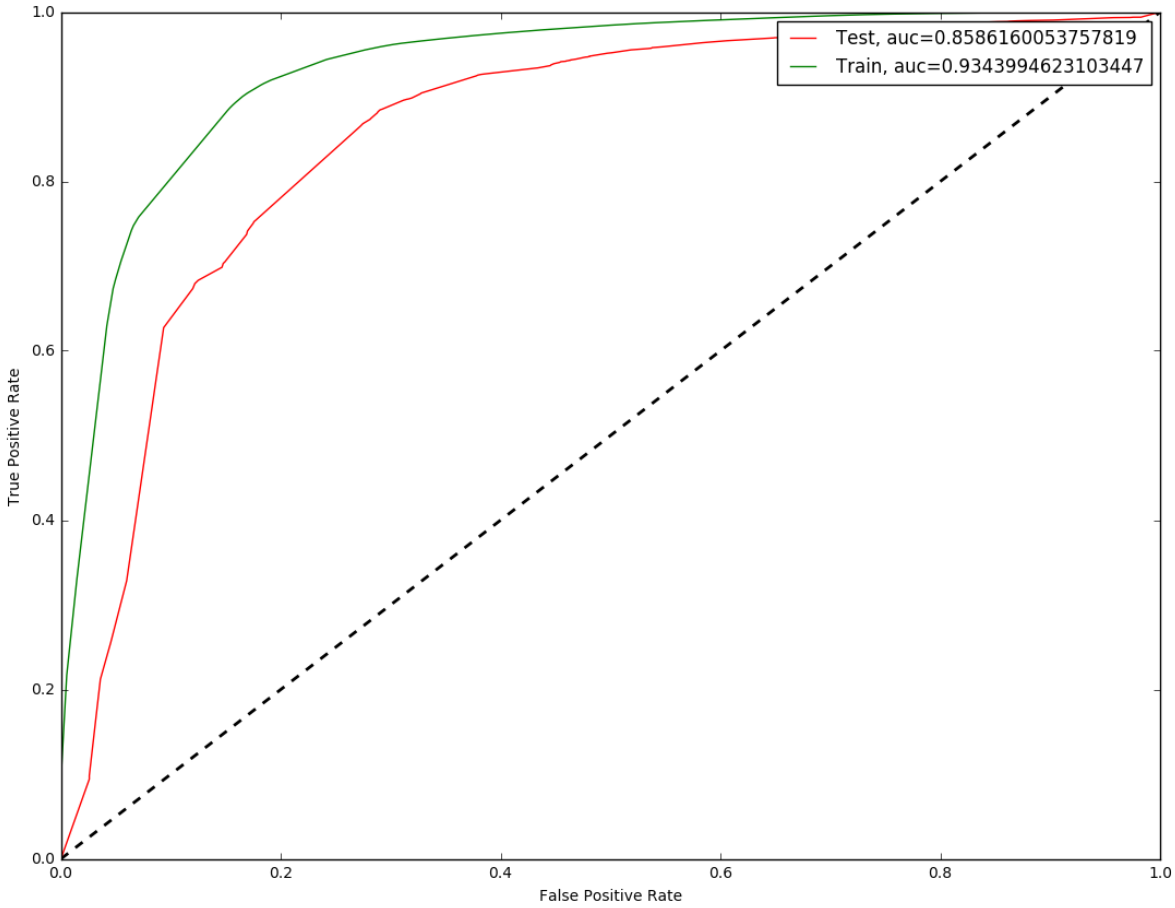
```
0.8586160053757819
0.9343994623103447
```

In [154]:

```
import pylab
plt.figure(figsize=(13, 10))
plt.plot([0,1], [0,1], color='black', lw=2, linestyle='--')
plt.plot(tfidf_fpr_test, tfidf_tpr_test, label="Test, auc="+str(tfidf_test_auc), color = 'r
plt.plot(tfidf_fpr_train, tfidf_tpr_train, label="Train, auc="+str(tfidf_train_auc), color

plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.legend()

plt.show()
```



In [155]:

```
tfidf_test_conf = tfidf_clf.predict(tfidf_dict['test_tf_idf'])
tfidf_train_conf = tfidf_clf.predict(tfidf_dict['train_tf_idf'])
```

In [156]:

```
from sklearn.metrics import classification_report, confusion_matrix
tfidf_train_conf_matrix = confusion_matrix(Y_train, tfidf_train_conf)
tfidf_test_conf_matrix = confusion_matrix(Y_test, tfidf_test_conf)
class_report = classification_report(Y_test, tfidf_test_conf)
print(tfidf_test_conf_matrix)
print(class_report)
```

```
[[ 1335  1341]
 [  833 16491]]
              precision    recall  f1-score   support

           0       0.62      0.50      0.55      2676
           1       0.92      0.95      0.94     17324

   micro avg       0.89      0.89      0.89     20000
   macro avg       0.77      0.73      0.74     20000
weighted avg       0.88      0.89      0.89     20000
```

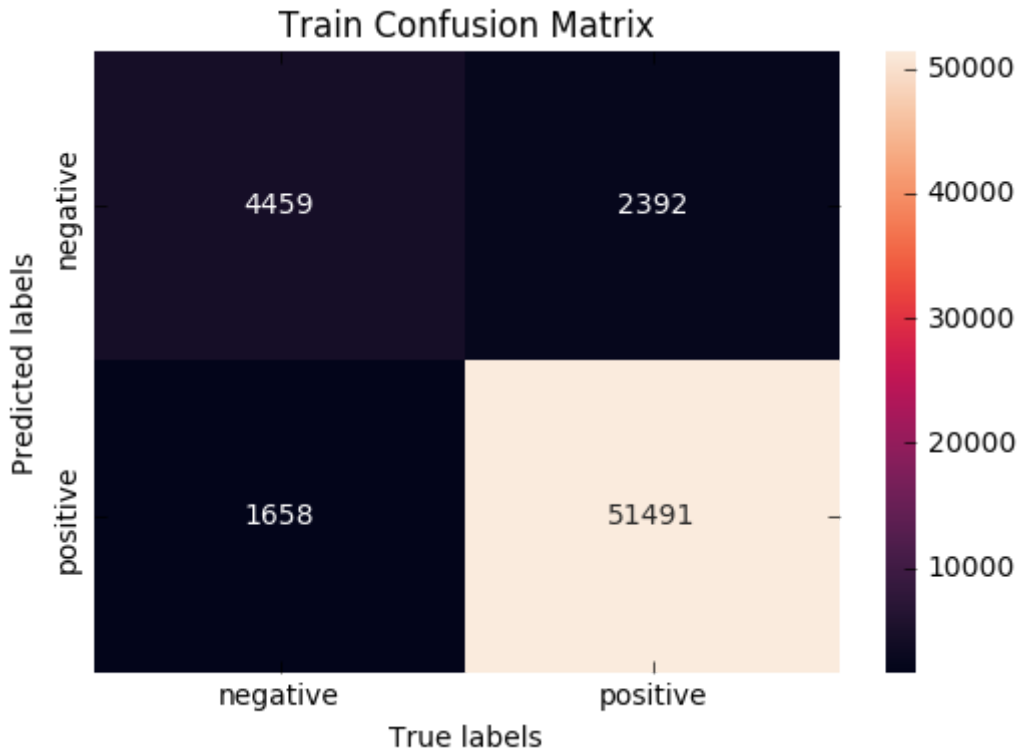
In [157]:

```
ax= plt.subplot()
sns.heatmap(tfidf_train_conf_matrix, annot=True, ax = ax, fmt='g')

ax.set_ylabel('Predicted labels')
ax.set_xlabel('True labels')
ax.set_title('Train Confusion Matrix')
ax.xaxis.set_ticklabels(['negative', 'positive'])
ax.yaxis.set_ticklabels(['negative', 'positive'])
```

Out[157]:

```
[<matplotlib.text.Text at 0x1e9812b0>, <matplotlib.text.Text at 0x1aae50b8>]
```



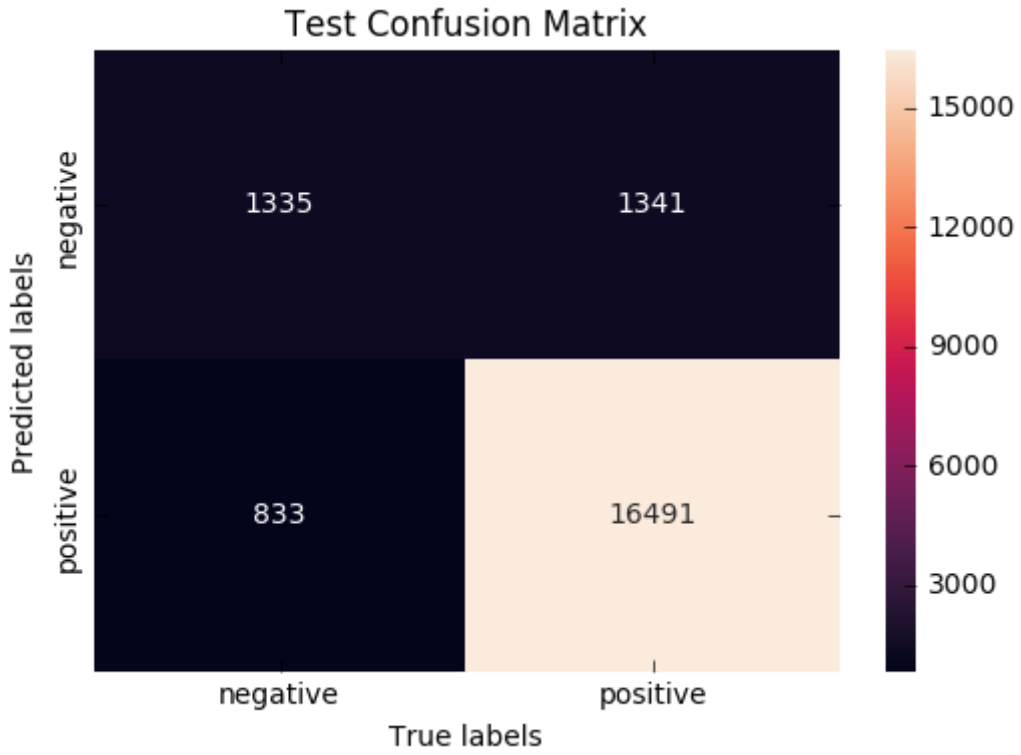
In [158]:

```
ax= plt.subplot()
sns.heatmap(tfidf_test_conf_matrix, annot=True, ax = ax, fmt='g')

ax.set_ylabel('Predicted labels')
ax.set_xlabel('True labels')
ax.set_title('Test Confusion Matrix')
ax.xaxis.set_ticklabels(['negative', 'positive'])
ax.yaxis.set_ticklabels(['negative', 'positive'])
```

Out[158]:

```
[<matplotlib.text.Text at 0x1e6fab00>, <matplotlib.text.Text at 0x2c4b0d30>]
```



Exporting Decision Tree

In [115]:

```
tclassifier = tree.DecisionTreeClassifier(max_depth = 3)
tclassifier.fit(tfidf_dict['train_tf_idf'],Y_train)
```

Out[115]:

```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=3,
            max_features=None, max_leaf_nodes=None, min_samples_leaf=1,
            min_samples_split=2, min_weight_fraction_leaf=0.0,
            presort=False, random_state=None, splitter='best')
```

In [ ]:

```
from graphviz import Source
import os
os.environ["PATH"] += os.pathsep + 'C:/Program Files (x86)/Graphviz2.38/bin/'
graph = Source(tree.export_graphviz(tclassifier, out_file='tftree.dot', feature_names=count
png_bytes = graph.pipe(format='png')
with open('tfidf.png','wb') as f:
    f.write(png_bytes)
```

Decision Tree on Avg_w2v

In [159]:

```
import pickle
with open(r"avg_w2v.pkl", "rb") as input_file:
    avg_tfidf_dict = pickle.load(input_file)
```

In [160]:

```
from scipy.sparse import vstack
X_train_val_avg = vstack((avg_tfidf_dict['X_train_avgw2v'], avg_tfidf_dict['X_val_avgw2v'])
```

In [161]:

```
print(X_train_val_avg.shape)
```

```
(80000, 50)
```

In [162]:

```
fpr_val = dict()
tpr_val = dict()
roc_auc_val = dict()
fpr_train = dict()
tpr_train = dict()
roc_auc_train = dict()
from sklearn.model_selection import GridSearchCV
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import roc_auc_score

param_grid = {'max_depth': [1, 5, 10, 50, 100, 500, 1000], 'min_samples_split' : [5, 10, 10

grid_search_cv_avg = GridSearchCV(DecisionTreeClassifier(), param_grid, scoring = 'roc_auc'
grid_search_cv_avg.fit(X_train_val_avg,Y_train_val)
```

Out[162]:

```
GridSearchCV(cv='warn', error_score='raise-deprecating',
       estimator=DecisionTreeClassifier(class_weight=None, 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='warn', n_jobs=None,
       param_grid={'max_depth': [1, 5, 10, 50, 100, 500, 1000], 'min_samples
_split': [5, 10, 100, 500]},
       pre_dispatch='2*n_jobs', refit=True, return_train_score='warn',
       scoring='roc_auc', verbose=0)
```

In [163]:

```
avg_results = grid_search_cv_avg.cv_results_
avg_results
```

```
        0.03799995, 0.04066666, 0.04299998, 0.04933333, 0.03966657,
        0.03866665, 0.03733333, 0.03866665, 0.04966672, 0.04533331,
        0.04200006, 0.04166667, 0.04666662, 0.04800002, 0.04500008,
        0.04200006, 0.04466669, 0.045     ]),
 'mean_test_score': array([0.67052526, 0.67052526, 0.67052526, 0.67052526,
0.81666947,
        0.81666947, 0.81666947, 0.8166891 , 0.79981036, 0.80398502,
        0.8305592 , 0.83421497, 0.67677021, 0.69470864, 0.79727571,
        0.8301696 , 0.67750268, 0.69242636, 0.79725363, 0.83010636,
        0.67691919, 0.69364174, 0.79676406, 0.83018103, 0.67756365,
        0.69290973, 0.79693989, 0.83033067]),
 'mean_train_score': array([0.67200073, 0.67200073, 0.67200073, 0.6720007
3, 0.83554298,
        0.83554298, 0.83554298, 0.83528583, 0.92314333, 0.92162778,
        0.90664611, 0.87967986, 0.99960135, 0.99731843, 0.95228873,
        0.89636143, 0.99959714, 0.99730292, 0.95234019, 0.89635257,
        0.99959933, 0.99731721, 0.95231958, 0.89636651, 0.99959485,
        0.99734129, 0.95230246, 0.89640726]),
 'param_max_depth': masked_array(data=[1, 1, 1, 1, 5, 5, 5, 5, 10, 10, 10,
```

In [164]:

```
avg_x_list = []
avg_y_list = []
for c1 in grid_search_cv_avg.cv_results_['params']:
    avg_x_list.append(c1['max_depth'])
for c2 in grid_search_cv_avg.cv_results_['params']:
    avg_y_list.append(c2['min_samples_split'])
print(avg_x_list, avg_y_list)
```

```
[1, 1, 1, 1, 5, 5, 5, 5, 10, 10, 10, 10, 50, 50, 50, 50, 100, 100, 100, 100,
500, 500, 500, 500, 1000, 1000, 1000, 1000] [5, 10, 100, 500, 5, 10, 100, 50
0, 5, 10, 100, 500, 5, 10, 100, 500, 5, 10, 100, 500, 5, 10, 100, 500, 5, 1
0, 100, 500]
```

In [165]:

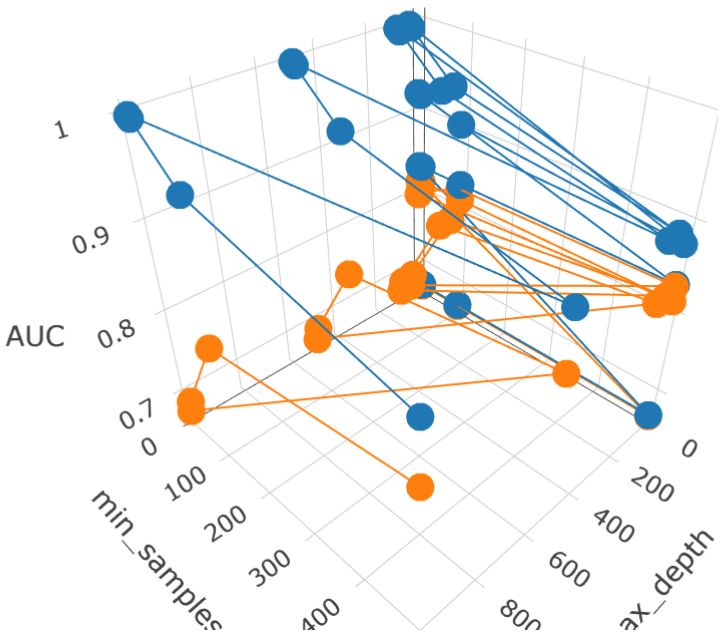
```
x1 = avg_x_list
y1 = avg_y_list
z1 = grid_search_cv_avg.cv_results_['mean_train_score'].tolist()
x2 = avg_x_list
y2 = avg_y_list
z2 = grid_search_cv_avg.cv_results_['mean_test_score'].tolist()
```

In [166]:

```
# https://plot.ly/python/3d-axes/
trace1 = go.Scatter3d(x=x1,y=y1,z=z1, name = 'train')
trace2 = go.Scatter3d(x=x2,y=y2,z=z2, name = 'Cross validation')
data = [trace1, trace2]

layout = go.Layout(scene = dict(
        xaxis = dict(title='max_depth'),
        yaxis = dict(title='min_samples_split'),
        zaxis = dict(title='AUC'),))

fig = go.Figure(data=data, layout=layout)
offline.iplot(fig, filename='3d-scatter-colorscale')
```



In [167]:

```
grid_search_cv_avg.best_params_
```

Out[167]:

```
{'max_depth': 10, 'min_samples_split': 500}
```

In [168]:

```
avg_best_max_depth = grid_search_cv_avg.best_params_['max_depth']
avg_best_min_samples_split = grid_search_cv_avg.best_params_['min_samples_split']
```

In [171]:

```
avg_clf = tree.DecisionTreeClassifier(max_depth = avg_best_max_depth, min_samples_split = a
avg_clf.fit(avg_tfidf_dict['X_train_avgw2v'],Y_train)
avg_train_proba = avg_clf.predict_proba(avg_tfidf_dict['X_train_avgw2v'])
avg_test_proba = avg_clf.predict_proba(avg_tfidf_dict['X_test_avgw2v'])
avg_test_proba
```

Out[171]:

```
array([[0.1025641 , 0.8974359 ],
       [0.90909091, 0.09090909],
       [0.02313702, 0.97686298],
       ...,
       [0.04298643, 0.95701357],
       [0.00585294, 0.99414706],
       [0.546875  , 0.453125  ]])
```

In [172]:

```
avg_fpr_train, avg_tpr_train, _ = roc_curve(Y_train, avg_train_proba[:, 1])
avg_fpr_test, avg_tpr_test, _ = roc_curve(Y_test, avg_test_proba[:, 1])
avg_test_auc = auc(avg_fpr_test, avg_tpr_test)
avg_train_auc = auc(avg_fpr_train, avg_tpr_train)
print(avg_test_auc)
print(avg_train_auc)
```

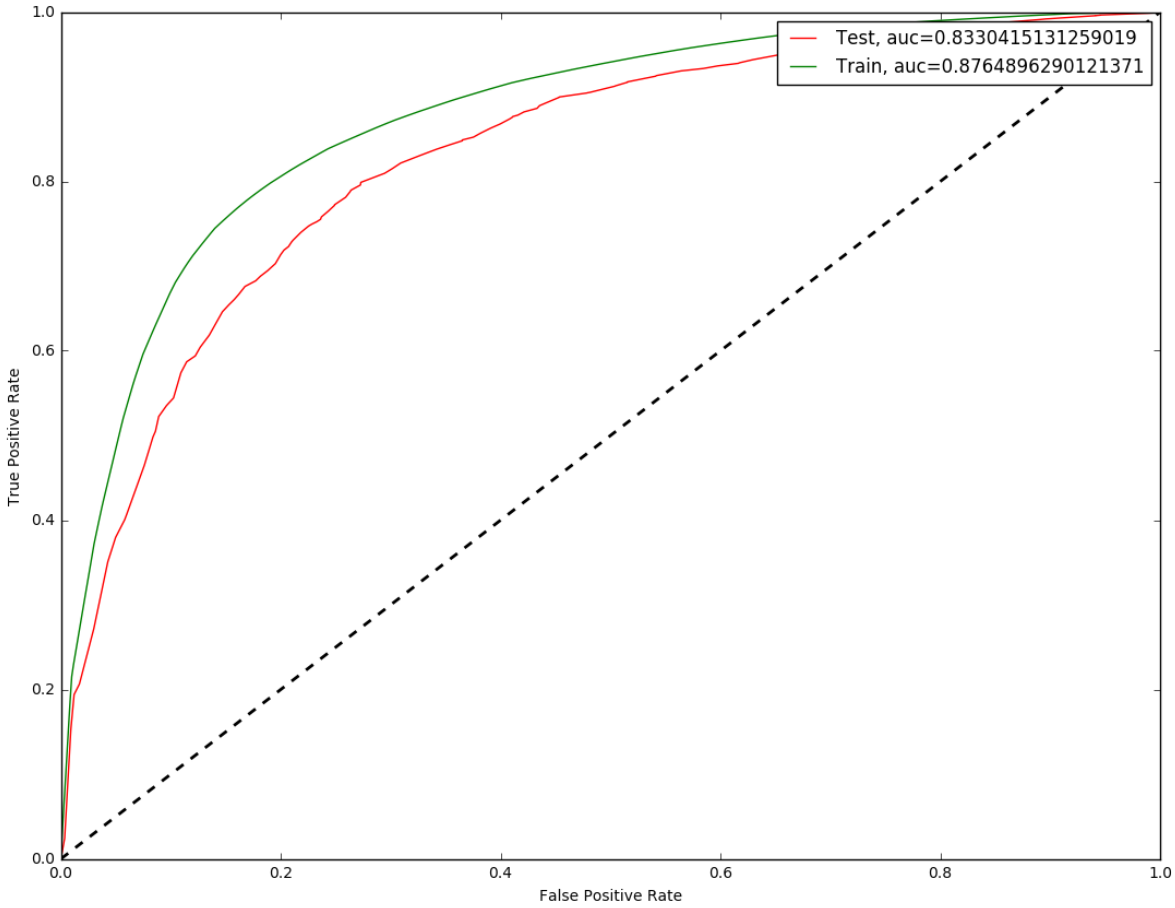
```
0.8330415131259019
0.8764896290121371
```

In [173]:

```
import pylab
plt.figure(figsize=(13, 10))
plt.plot([0,1], [0,1], color='black', lw=2, linestyle='--')
plt.plot(avg_fpr_test, avg_tpr_test, label="Test, auc="+str(avg_test_auc), color = 'red')
plt.plot(avg_fpr_train, avg_tpr_train, label="Train, auc="+str(avg_train_auc), color = 'gre

plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.legend()

plt.show()
```



In [174]:

```
avg_test_conf = avg_clf.predict(avg_tfidf_dict['X_test_avgw2v'])
avg_train_conf = avg_clf.predict(avg_tfidf_dict['X_train_avgw2v'])
```

In [175]:

```
from sklearn.metrics import classification_report, confusion_matrix
avg_train_conf_matrix = confusion_matrix(Y_train, avg_train_conf)
avg_test_conf_matrix = confusion_matrix(Y_test, avg_test_conf)
class_report = classification_report(Y_test, avg_test_conf)
print(avg_test_conf_matrix)
print(class_report)
```

```
[[  706  1970]
 [  474 16850]]
              precision    recall  f1-score   support

           0       0.60      0.26      0.37      2676
           1       0.90      0.97      0.93     17324

   micro avg       0.88      0.88      0.88     20000
   macro avg       0.75      0.62      0.65     20000
weighted avg       0.86      0.88      0.86     20000
```

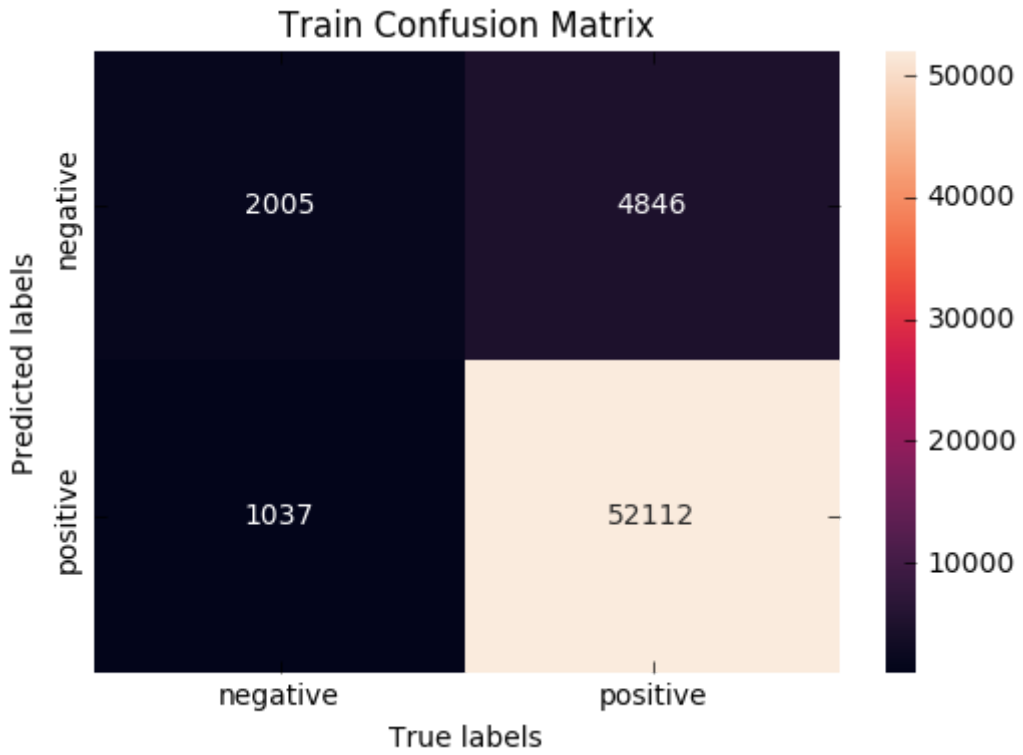
In [176]:

```
ax= plt.subplot()
sns.heatmap(avg_train_conf_matrix, annot=True, ax = ax, fmt='g')

ax.set_ylabel('Predicted labels')
ax.set_xlabel('True labels')
ax.set_title('Train Confusion Matrix')
ax.xaxis.set_ticklabels(['negative', 'positive'])
ax.yaxis.set_ticklabels(['negative', 'positive'])
```

Out[176]:

```
[<matplotlib.text.Text at 0x1b42eeb8>, <matplotlib.text.Text at 0x28339ac8>]
```



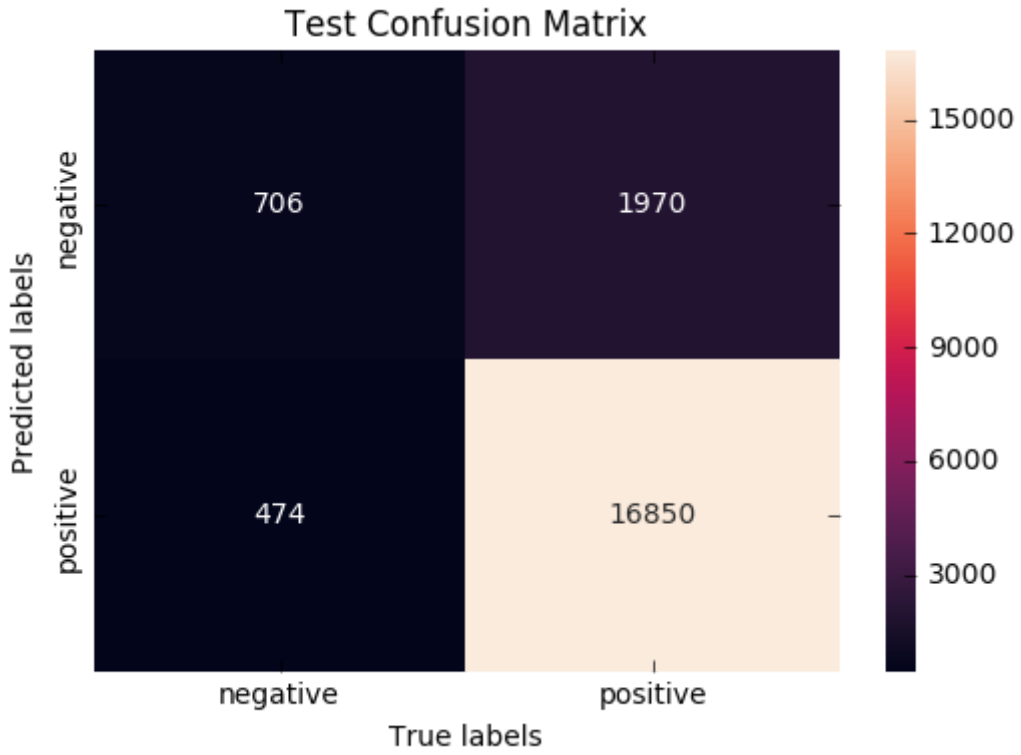
In [177]:

```
ax= plt.subplot()
sns.heatmap(avg_test_conf_matrix, annot=True, ax = ax, fmt='g')

ax.set_ylabel('Predicted labels')
ax.set_xlabel('True labels')
ax.set_title('Test Confusion Matrix')
ax.xaxis.set_ticklabels(['negative', 'positive'])
ax.yaxis.set_ticklabels(['negative', 'positive'])
```

Out[177]:

```
[<matplotlib.text.Text at 0x2fb11c18>, <matplotlib.text.Text at 0x17a97358>]
```



In [139]:

```
tclassifier = tree.DecisionTreeClassifier(max_depth = 3)
tclassifier.fit(avg_tfidf_dict['X_train_avgw2v'],Y_train)
```

Out[139]:

```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=3,
            max_features=None, max_leaf_nodes=None, min_samples_leaf=1,
            min_samples_split=2, min_weight_fraction_leaf=0.0,
            presort=False, random_state=None, splitter='best')
```

In [ ]:

```
from graphviz import Source
import os
os.environ["PATH"] += os.pathsep + 'C:/Program Files (x86)/Graphviz2.38/bin/'
graph = Source(tree.export_graphviz(tclassifier, out_file='avg.dot', feature_names=count_ve
png_bytes = graph.pipe(format='png')
with open('tfidf.png','wb') as f:
    f.write(png_bytes)
```

Decision Tree on TFIDF-w2v

In [178]:

```
import pickle
with open(r"tfidf_w2v.pkl", "rb") as input_file:
    tfidfw2v_dict = pickle.load(input_file)
```

In [180]:

```
from scipy.sparse import vstack
X_train_val_tfw2v = vstack((tfidfw2v_dict['X_train_tfidfw2v'], tfidfw2v_dict['X_val_tfidfw2
```

In [181]:

```
print(X_train_val_tfw2v.shape)
```

```
(80000, 50)
```

In [182]:

```
fpr_val = dict()
tpr_val = dict()
roc_auc_val = dict()
fpr_train = dict()
tpr_train = dict()
roc_auc_train = dict()
from sklearn.model_selection import GridSearchCV
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import roc_auc_score

param_grid = {'max_depth': [1, 5, 10, 50, 100, 500, 1000], 'min_samples_split' : [5, 10, 10

grid_search_cv_tfw2v = GridSearchCV(DecisionTreeClassifier(), param_grid, scoring = 'roc_au
grid_search_cv_tfw2v.fit(X_train_val_tfw2v,Y_train_val)
```

Out[182]:

```
GridSearchCV(cv='warn', error_score='raise-deprecating',
       estimator=DecisionTreeClassifier(class_weight=None, 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='warn', n_jobs=None,
       param_grid={'max_depth': [1, 5, 10, 50, 100, 500, 1000], 'min_samples
_split': [5, 10, 100, 500]},
       pre_dispatch='2*n_jobs', refit=True, return_train_score='warn',
       scoring='roc_auc', verbose=0)
```

In [183]:

```
tfw2v_results = grid_search_cv_tfw2v.cv_results_
tfw2v_results
```

Out[183]:

```
{'mean_fit_time': array([ 0.46433338,  0.488     ,  0.46566677,  0.4603332
7,  2.13033334,
         2.18599987,  2.17600004,  2.284     ,  6.79233329,  6.13433329,
         5.9216667 ,  4.8956666 , 15.05966671, 14.46066666, 13.10500002,
         8.33300002, 14.31166665, 14.73033333, 13.05933332,  8.42466664,
        13.94766672, 13.75700013, 12.81066656,  8.53700002, 14.60699995,
        14.15933339, 13.00566665,  8.66666667]),
 'mean_score_time': array([0.03500001, 0.04100005, 0.03666671, 0.03633332,
0.03766664,
        0.03666679, 0.03599993, 0.046     , 0.03866673, 0.03900003,
        0.03933326, 0.03833334, 0.04633331, 0.04233336, 0.03933342,
        0.04199998, 0.03866673, 0.03933334, 0.04199998, 0.04566669,
        0.03933318, 0.03799995, 0.04133336, 0.04100005, 0.04600008,
        0.04166659, 0.04200006, 0.04500008]),
 'mean_test_score': array([0.64312817, 0.64312817, 0.64312817, 0.64312817,
0.78511677,
        0.78513053, 0.78511677, 0.78494906, 0.77329466, 0.77559351,
        0.79806523, 0.80227924, 0.65070033, 0.66743603, 0.77156221,
```

In [184]:

```
tfw2v_x_list = []
tfw2v_y_list = []
for c1 in grid_search_cv_tfw2v.cv_results_['params']:
    tfw2v_x_list.append(c1['max_depth'])
for c2 in grid_search_cv_tfw2v.cv_results_['params']:
    tfw2v_y_list.append(c2['min_samples_split'])
print(tfw2v_x_list, tfw2v_y_list)
```

```
[1, 1, 1, 1, 5, 5, 5, 5, 10, 10, 10, 10, 50, 50, 50, 50, 100, 100, 100, 100,
500, 500, 500, 500, 1000, 1000, 1000, 1000] [5, 10, 100, 500, 5, 10, 100, 50
0, 5, 10, 100, 500, 5, 10, 100, 500, 5, 10, 100, 500, 5, 10, 100, 500, 5, 1
0, 100, 500]
```

In [185]:

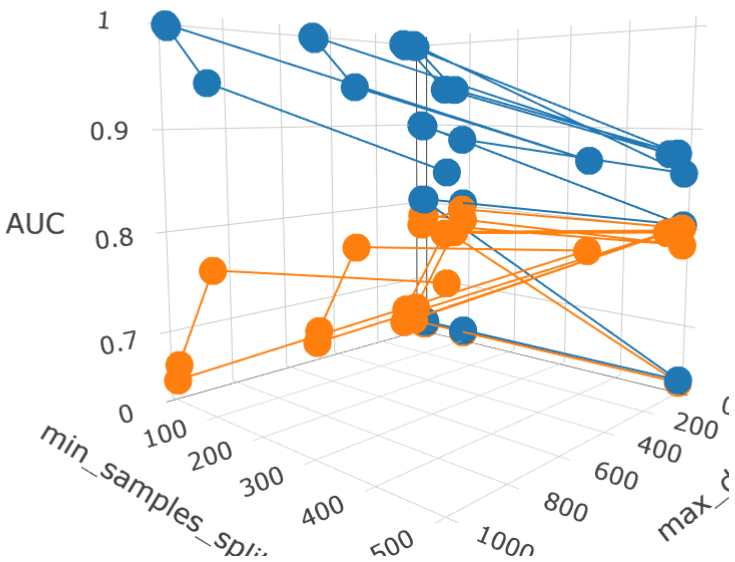
```
x1 = tfw2v_x_list
y1 = tfw2v_y_list
z1 = grid_search_cv_tfw2v.cv_results_['mean_train_score'].tolist()
x2 = tfw2v_x_list
y2 = tfw2v_y_list
z2 = grid_search_cv_tfw2v.cv_results_['mean_test_score'].tolist()
```

In [186]:

```
# https://plot.ly/python/3d-axes/
trace1 = go.Scatter3d(x=x1,y=y1,z=z1, name = 'train')
trace2 = go.Scatter3d(x=x2,y=y2,z=z2, name = 'Cross validation')
data = [trace1, trace2]

layout = go.Layout(scene = dict(
        xaxis = dict(title='max_depth'),
        yaxis = dict(title='min_samples_split'),
        zaxis = dict(title='AUC'),))

fig = go.Figure(data=data, layout=layout)
offline.iplot(fig, filename='3d-scatter-colorscale')
```



In [187]:

```
grid_search_cv_tfw2v.best_params_
```

Out[187]:

```
{'max_depth': 10, 'min_samples_split': 500}
```

In [188]:

```
tfw2v_best_max_depth = grid_search_cv_tfw2v.best_params_['max_depth']
tfw2v_best_min_samples_split = grid_search_cv_tfw2v.best_params_['min_samples_split']
```

In [190]:

```
tfw2v_clf = tree.DecisionTreeClassifier(max_depth = tfw2v_best_max_depth, min_samples_split
tfw2v_clf.fit(tfidfw2v_dict['X_train_tfidfw2v'],Y_train)
tfw2v_train_proba = tfw2v_clf.predict_proba(tfidfw2v_dict['X_train_tfidfw2v'])
tfw2v_test_proba = tfw2v_clf.predict_proba(tfidfw2v_dict['X_test_tfidfw2v'])
tfw2v_test_proba
```

Out[190]:

```
array([[0.12802768, 0.87197232],
       [0.30407524, 0.69592476],
       [0.05617978, 0.94382022],
       ...,
       [0.06041987, 0.93958013],
       [0.02999595, 0.97000405],
       [0.52434457, 0.47565543]])
```

In [191]:

```
tfw2v_fpr_train, tfw2v_tpr_train, _ = roc_curve(Y_train, tfw2v_train_proba[:, 1])
tfw2v_fpr_test, tfw2v_tpr_test, _ = roc_curve(Y_test, tfw2v_test_proba[:, 1])
tfw2v_test_auc = auc(tfw2v_fpr_test, tfw2v_tpr_test)
tfw2v_train_auc = auc(tfw2v_fpr_train, tfw2v_tpr_train)
print(tfw2v_test_auc)
print(tfw2v_train_auc)
```

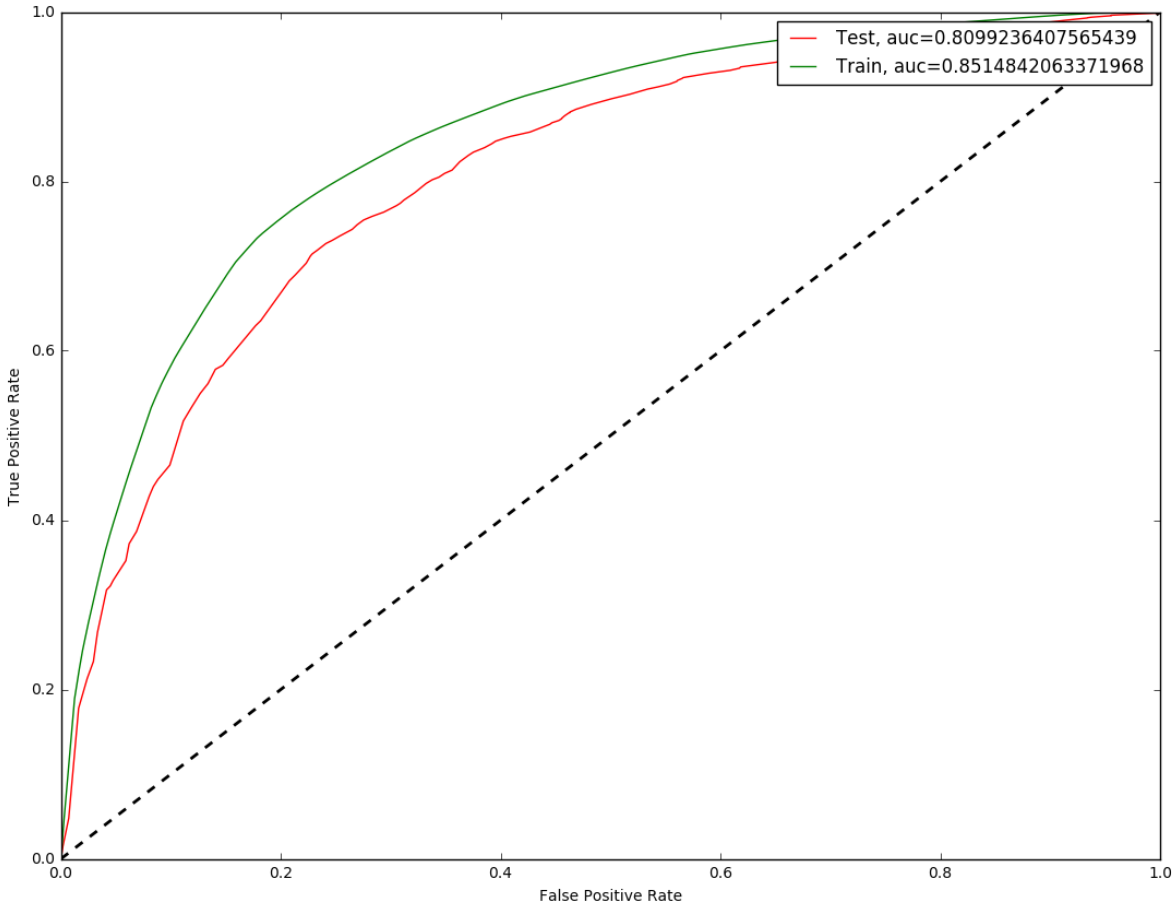
```
0.8099236407565439
0.8514842063371968
```

In [192]:

```
import pylab
plt.figure(figsize=(13, 10))
plt.plot([0,1], [0,1], color='black', lw=2, linestyle='--')
plt.plot(tfw2v_fpr_test, tfw2v_tpr_test, label="Test, auc="+str(tfw2v_test_auc), color = 'r
plt.plot(tfw2v_fpr_train, tfw2v_tpr_train, label="Train, auc="+str(tfw2v_train_auc), color

plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.legend()

plt.show()
```



In [194]:

```
tfw2v_test_conf = tfw2v_clf.predict(tfidfw2v_dict['X_test_tfidfw2v'])
tfw2v_train_conf = tfw2v_clf.predict(tfidfw2v_dict['X_train_tfidfw2v'])
```

In [195]:

```
from sklearn.metrics import classification_report, confusion_matrix
tfw2v_train_conf_matrix = confusion_matrix(Y_train, tfw2v_train_conf)
tfw2v_test_conf_matrix = confusion_matrix(Y_test, tfw2v_test_conf)
class_report = classification_report(Y_test, tfw2v_test_conf)
print(tfw2v_test_conf_matrix)
print(class_report)
```

```
[[  739  1937]
 [  665 16659]]
              precision    recall  f1-score   support

           0       0.53      0.28      0.36      2676
           1       0.90      0.96      0.93     17324

   micro avg       0.87      0.87      0.87     20000
   macro avg       0.71      0.62      0.64     20000
weighted avg       0.85      0.87      0.85     20000
```

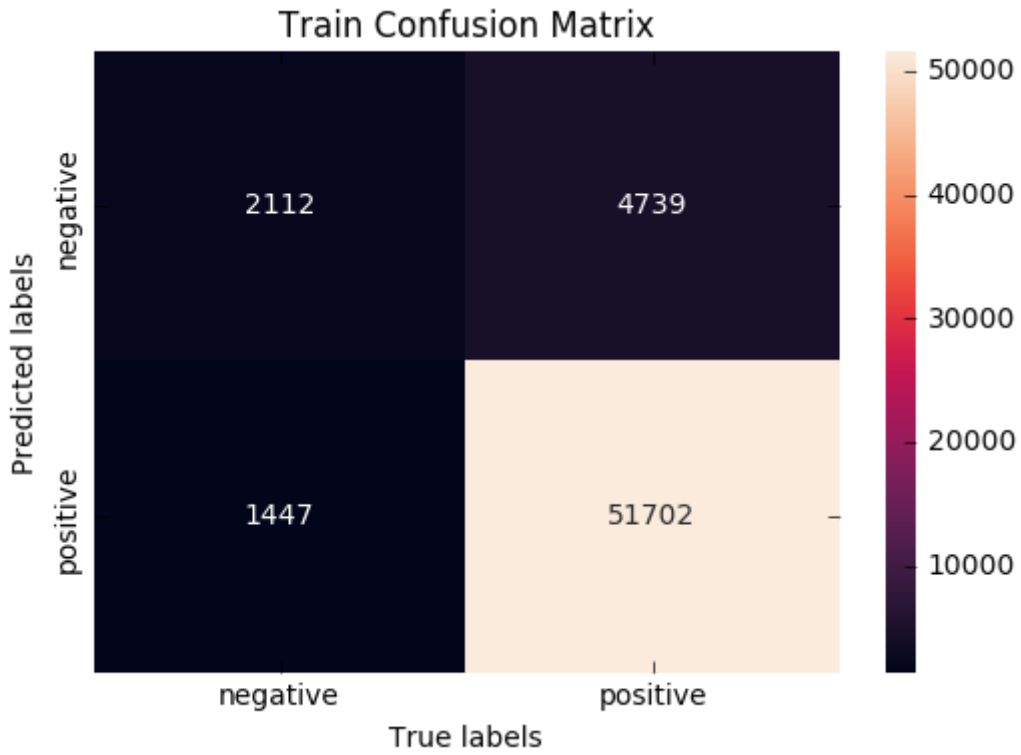
In [197]:

```
ax= plt.subplot()
sns.heatmap(tfw2v_train_conf_matrix, annot=True, ax = ax, fmt='g')

ax.set_ylabel('Predicted labels')
ax.set_xlabel('True labels')
ax.set_title('Train Confusion Matrix')
ax.xaxis.set_ticklabels(['negative', 'positive'])
ax.yaxis.set_ticklabels(['negative', 'positive'])
```

Out[197]:

```
[<matplotlib.text.Text at 0x322f56d8>, <matplotlib.text.Text at 0x1f89fe80>]
```



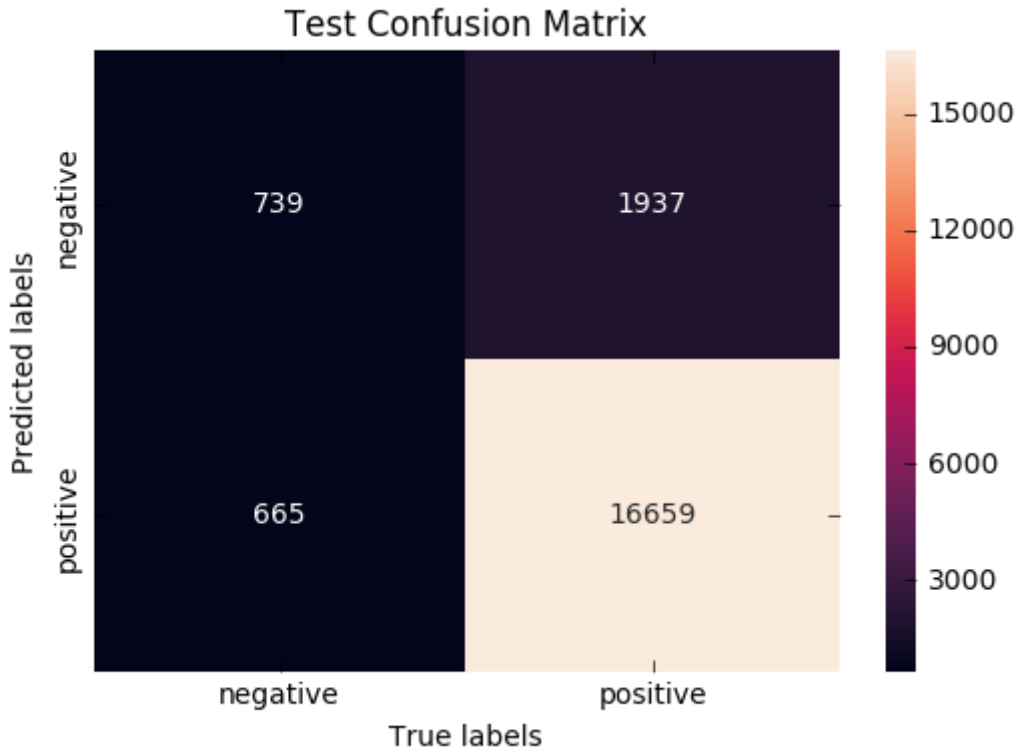
In [198]:

```
ax= plt.subplot()
sns.heatmap(tfw2v_test_conf_matrix, annot=True, ax = ax, fmt='g')

ax.set_ylabel('Predicted labels')
ax.set_xlabel('True labels')
ax.set_title('Test Confusion Matrix')
ax.xaxis.set_ticklabels(['negative', 'positive'])
ax.yaxis.set_ticklabels(['negative', 'positive'])
```

Out[198]:

```
[<matplotlib.text.Text at 0x4ce9e8d0>, <matplotlib.text.Text at 0x4ceae2e8>]
```



In [164]:

```
tclassifier = tree.DecisionTreeClassifier(max_depth = 3)
tclassifier.fit(tfidfw2v_dict['X_train_tfidfw2v'],Y_train)
```

Out[164]:

```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=3,
            max_features=None, max_leaf_nodes=None, min_samples_leaf=1,
            min_samples_split=2, min_weight_fraction_leaf=0.0,
            presort=False, random_state=None, splitter='best')
```

In [ ]:

```
from graphviz import Source
import os
os.environ["PATH"] += os.pathsep + 'C:/Program Files (x86)/Graphviz2.38/bin/'
graph = Source(tree.export_graphviz(tclassifier, out_file='avg.dot', feature_names=count_ve
png_bytes = graph.pipe(format='png')
with open('tfidf.png','wb') as f:
    f.write(png_bytes)
```

In [199]:

```
from prettytable import PrettyTable

x = PrettyTable()
x.field_names = ["Max_depth", "Min_samples_split",  "Vectorizer", "Train", "Test"]

x.add_row([50, 500, "BoW", 0.920, 0.863])
x.add_row([50, 500, "Tf-idf", 0.934, 0.858])
x.add_row([10, 500, "Avg-w2v", 0.876, 0.833])
x.add_row([10, 500, "tfidf_w2v", 0.851, 0.809])
print(x)
```

```
+-----------+-------------------+------------+-------+-------+
| Max_depth | Min_samples_split | Vectorizer | Train |  Test |
+-----------+-------------------+------------+-------+-------+
|     50    |        500        |    BoW     |  0.92 | 0.863 |
|     50    |        500        |   Tf-idf   | 0.934 | 0.858 |
|     10    |        500        |  Avg-w2v   | 0.876 | 0.833 |
|     10    |        500        | tfidf_w2v  | 0.851 | 0.809 |
+-----------+-------------------+------------+-------+-------+
```

Steps taken to increase performance:
i. Summary and Text columns are appended in single column
ii. length of words is taken from appended column and stacked with sparse matrix

Observations:
i. Accuracy increased around 2% for each vectorizer.

In [ ]: