Amazon Fine Food Reviews Analysis

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

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

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

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

Attribute Information:

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

Objective:

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

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

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

[1]. Reading Data

[1.1] Loading the data

The dataset is available in two forms

1. .csv file
2. SQLite Database

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

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

In [0]:
```
%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 [0]:

```
# 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 50
0000 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 Sco
re != 3 LIMIT 500000""", con)
# for tsne assignment you can take 5k data points
```

```
filtered_data = pd.read_sql_query(""" SELECT * FROM Reviews WHERE Score
 != 3 LIMIT 5000""", con)

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

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

```
Number of data points in our data (5000, 10)
```

Out[0]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	Helpfulnes
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	Helpfulnes
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1

In [0]:
```
display = pd.read_sql_query("""
SELECT UserId, ProductId, ProfileName, Time, Score, Text, COUNT(*)
FROM Reviews
GROUP BY UserId
HAVING COUNT(*)>1
""", con)
```

In [0]:
```
print(display.shape)
display.head()
```

```
(80668, 7)
```

Out[0]:

	UserId	ProductId	ProfileName	Time	Score	Text	COUI
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

	UserId	ProductId	ProfileName	Time	Score	Text	COUI
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 [0]:
```
display[display['UserId']=='AZY10LLTJ71NX']
```

Out[0]:

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

In [0]:
```
display['COUNT(*)'].sum()
```

Out[0]: 393063

[2] Exploratory Data Analysis

[2.1] Data Cleaning: Deduplication

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

In [0]:
```
display= pd.read_sql_query("""
SELECT *
FROM Reviews
WHERE Score != 3 AND UserId="AR5J8UI46CURR"
ORDER BY ProductID
""", con)
display.head()
```

Out[0]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulN
0	78445	B000HDL1RQ	AR5J8UI46CURR	Geetha Krishnan	2	2
1	138317	B000HDOPYC	AR5J8UI46CURR	Geetha Krishnan	2	2

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulN
2	138277	B000HDOPYM	AR5J8UI46CURR	Geetha Krishnan	2	2
3	73791	B000HDOPZG	AR5J8UI46CURR	Geetha Krishnan	2	2
4	155049	B000PAQ75C	AR5J8UI46CURR	Geetha Krishnan	2	2

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

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

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

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

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

In [0]:
```
#Sorting data according to ProductId in ascending order
sorted_data=filtered_data.sort_values('ProductId', axis=0, ascending=True, inplace=False, kind='quicksort', na_position='last')
```

In [0]:
```
#Deduplication of entries
final=sorted_data.drop_duplicates(subset={"UserId","ProfileName","Time","Text"}, keep='first', inplace=False)
final.shape
```

Out[0]:
```
(4986, 10)
```

In [0]:
```
#Checking to see how much % of data still remains
(final['Id'].size*1.0)/(filtered_data['Id'].size*1.0)*100
```

Out[0]:
```
99.72
```

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 [0]:
```
display= pd.read_sql_query("""
SELECT *
FROM Reviews
WHERE Score != 3 AND Id=44737 OR Id=64422
ORDER BY ProductID
""", con)

display.head()
```

Out[0]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulIn
0	64422	B000MIDROQ	A161DK06JJMCYF	J. E. Stephens "Jeanne"	3	1
1	44737	B001EQ55RW	A2V0I904FH7ABY	Ram	3	2

In [0]:
```
final=final[final.HelpfulnessNumerator<=final.HelpfulnessDenominator]
```

In [0]:
```
#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()
```

```
(4986, 10)
```

Out[0]:
```
1    4178
0     808
Name: Score, dtype: int64
```

[3] Preprocessing

[3.1]. Preprocessing Review Text

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

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

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

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

In [0]:
```
# 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)
```

```
Why is this $[...] when the same product is available for $[...] here?<
br />http://www.amazon.com/VICTOR-FLY-MAGNET-BAIT-REEL/dp/B00004RBDY<
```

```
br /><br />The Victor M380 and M502 traps are unreal, of course -- tota
l fly genocide. Pretty stinky, but only right nearby.
==================================================
I recently tried this flavor/brand and was surprised at how delicious t
hese chips are.  The best thing was that there were a lot of "brown" ch
ips in the bsg (my favorite), so I bought some more through amazon and
shared with family and friends.  I am a little disappointed that there
are not, so far, very many brown chips in these bags, but the flavor is
still very good.  I like them better than the yogurt and green onion fl
avor because they do not seem to be as salty, and the onion flavor is b
etter.  If you haven't eaten Kettle chips before, I recommend that you
try a bag before buying bulk.  They are thicker and crunchier than Lays
but just as fresh out of the bag.
==================================================
Wow.  So far, two two-star reviews.  One obviously had no idea what the
y were ordering; the other wants crispy cookies.  Hey, I'm sorry; but t
hese reviews do nobody any good beyond reminding us to look  before ord
ering.<br /><br />These are chocolate-oatmeal cookies.  If you don't li
ke that combination, don't order this type of cookie.  I find the combo
quite nice, really.  The oatmeal sort of "calms" the rich chocolate fla
vor and gives the cookie sort of a coconut-type consistency.  Now let's
also remember that tastes differ; so, I've given my opinion.<br /><br /
>Then, these are soft, chewy cookies -- as advertised.  They are not "c
rispy" cookies, or the blurb would say "crispy," rather than "chewy."
I happen to like raw cookie dough; however, I don't see where these tas
te like raw cookie dough.  Both are soft, however, so is this the confu
sion?  And, yes, they stick together.  Soft cookies tend to do that.  T
hey aren't individually wrapped, which would add to the cost.  Oh yeah,
chocolate chip cookies tend to be somewhat sweet.<br /><br />So, if you
want something hard and crisp, I suggest Nabiso's Ginger Snaps.  If you
want a cookie that's soft, chewy and tastes like a combination of choco
late and oatmeal, give these a try.  I'm here to place my second order.
==================================================
love to order my coffee on amazon.  easy and shows up quickly.<br />Thi
s k cup is great coffee.  dcaf is very good as well
==================================================
```

In [0]:
```
# remove urls from text python: https://stackoverflow.com/a/40823105/40
84039
```

```
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)
```

```
Why is this $[...] when the same product is available for $[...] here?<
br /> /><br />The Victor M380 and M502 traps are unreal, of course -- t
otal fly genocide. Pretty stinky, but only right nearby.
```

In [0]:

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

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

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

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

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

```
Why is this $[...] when the same product is available for $[...] here?
/>The Victor M380 and M502 traps are unreal, of course -- total fly gen
ocide. Pretty stinky, but only right nearby.
==================================================
I recently tried this flavor/brand and was surprised at how delicious t
hese chips are.  The best thing was that there were a lot of "brown" ch
```

```
ips in the bsg (my favorite), so I bought some more through amazon and
shared with family and friends.  I am a little disappointed that there
are not, so far, very many brown chips in these bags, but the flavor is
still very good.  I like them better than the yogurt and green onion fl
avor because they do not seem to be as salty, and the onion flavor is b
etter.  If you haven't eaten Kettle chips before, I recommend that you
try a bag before buying bulk.  They are thicker and crunchier than Lays
but just as fresh out of the bag.
==================================================
Wow.  So far, two two-star reviews.  One obviously had no idea what the
y were ordering; the other wants crispy cookies.  Hey, I'm sorry; but t
hese reviews do nobody any good beyond reminding us to look  before ord
ering.These are chocolate-oatmeal cookies.  If you don't like that comb
ination, don't order this type of cookie.  I find the combo quite nice,
really.  The oatmeal sort of "calms" the rich chocolate flavor and give
s the cookie sort of a coconut-type consistency.  Now let's also rememb
er that tastes differ; so, I've given my opinion.Then, these are soft,
chewy cookies -- as advertised.  They are not "crispy" cookies, or the
blurb would say "crispy," rather than "chewy."  I happen to like raw co
okie dough; however, I don't see where these taste like raw cookie doug
h.  Both are soft, however, so is this the confusion?  And, yes, they s
tick together.  Soft cookies tend to do that.  They aren't individually
wrapped, which would add to the cost.  Oh yeah, chocolate chip cookies
tend to be somewhat sweet.So, if you want something hard and crisp, I s
uggest Nabiso's Ginger Snaps.  If you want a cookie that's soft, chewy
and tastes like a combination of chocolate and oatmeal, give these a tr
y.  I'm here to place my second order.
==================================================
love to order my coffee on amazon.  easy and shows up quickly.This k cu
p is great coffee.  dcaf is very good as well
```

In [0]:

```
# 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 [0]:
```
sent_1500 = decontracted(sent_1500)
print(sent_1500)
print("="*50)
```

```
Wow.  So far, two two-star reviews.  One obviously had no idea what the
y were ordering; the other wants crispy cookies.  Hey, I am sorry; but
these reviews do nobody any good beyond reminding us to look  before or
dering.<br /><br />These are chocolate-oatmeal cookies.  If you do not
like that combination, do not order this type of cookie.  I find the co
mbo quite nice, really.  The oatmeal sort of "calms" the rich chocolate
flavor and gives the cookie sort of a coconut-type consistency.  Now le
t is also remember that tastes differ; so, I have given my opinion.<br
/><br />Then, these are soft, chewy cookies -- as advertised.  They are
not "crispy" cookies, or the blurb would say "crispy," rather than "che
wy."  I happen to like raw cookie dough; however, I do not see where th
ese taste like raw cookie dough.  Both are soft, however, so is this th
e confusion?  And, yes, they stick together.  Soft cookies tend to do t
hat.  They are not individually wrapped, which would add to the cost.
Oh yeah, chocolate chip cookies tend to be somewhat sweet.<br /><br />S
o, if you want something hard and crisp, I suggest Nabiso is Ginger Sna
ps.  If you want a cookie that is soft, chewy and tastes like a combina
tion of chocolate and oatmeal, give these a try.  I am here to place my
second order.
==================================================
```

In [0]:
```
#remove words with numbers python: https://stackoverflow.com/a/1808237
0/4084039
```

```
sent_0 = re.sub("\S*\d\S*", "", sent_0).strip()
print(sent_0)
```

```
Why is this $[...] when the same product is available for $[...] here?<
br /> /><br />The Victor  and  traps are unreal, of course -- total fly
genocide. Pretty stinky, but only right nearby.
```

In [0]:
```
#remove spacial character: https://stackoverflow.com/a/5843547/4084039
sent_1500 = re.sub('[^A-Za-z0-9]+', ' ', sent_1500)
print(sent_1500)
```

```
Wow So far two two star reviews One obviously had no idea what they wer
e ordering the other wants crispy cookies Hey I am sorry but these revi
ews do nobody any good beyond reminding us to look before ordering br b
r These are chocolate oatmeal cookies If you do not like that combinati
on do not order this type of cookie I find the combo quite nice really
The oatmeal sort of calms the rich chocolate flavor and gives the cooki
e sort of a coconut type consistency Now let is also remember that tast
es differ so I have given my opinion br br Then these are soft chewy co
okies as advertised They are not crispy cookies or the blurb would say
crispy rather than chewy I happen to like raw cookie dough however I do
not see where these taste like raw cookie dough Both are soft however s
o is this the confusion And yes they stick together Soft cookies tend t
o do that They are not individually wrapped which would add to the cost
Oh yeah chocolate chip cookies tend to be somewhat sweet br br So if yo
u want something hard and crisp I suggest Nabiso is Ginger Snaps If you
want a cookie that is soft chewy and tastes like a combination of choco
late and oatmeal give these a try I am here to place my second order
```

In [0]:
```
# https://gist.github.com/sebleier/554280
# we are removing the words from the stop words list: 'no', 'nor', 'no
t'
# <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', 'o
urs', 'ourselves', 'you', "you're", "you've",\
```

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

In [0]:
```
# Combining all the above stundents
from tqdm import tqdm
preprocessed_reviews = []
# tqdm is for printing the status bar
for sentance in tqdm(final['Text'].values):
    sentance = re.sub(r"http\S+", "", sentance)
    sentance = BeautifulSoup(sentance, 'lxml').get_text()
    sentance = decontracted(sentance)
    sentance = re.sub("\S*\d\S*", "", sentance).strip()
    sentance = re.sub('[^A-Za-z]+', ' ', sentance)
    # https://gist.github.com/sebleier/554280
```

```
    sentance = ' '.join(e.lower() for e in sentance.split() if e.lower
() not in stopwords)
    preprocessed_reviews.append(sentance.strip())
```

```
100%|██████████| 4986/4986 [00:01<00:00, 3137.37it/s]
```

In [0]:
```
preprocessed_reviews[1500]
```

Out[0]:
```
'wow far two two star reviews one obviously no idea ordering wants cris
py cookies hey sorry reviews nobody good beyond reminding us look order
ing chocolate oatmeal cookies not like combination not order type cooki
e find combo quite nice really oatmeal sort calms rich chocolate flavor
gives cookie sort coconut type consistency let also remember tastes dif
fer given opinion soft chewy cookies advertised not crispy cookies blur
b would say crispy rather chewy happen like raw cookie dough however no
t see taste like raw cookie dough soft however confusion yes stick toge
ther soft cookies tend not individually wrapped would add cost oh yeah
chocolate chip cookies tend somewhat sweet want something hard crisp su
ggest nabiso ginger snaps want cookie soft chewy tastes like combinatio
n chocolate oatmeal give try place second order'
```

[3.2] Preprocessing Review Summary

In [0]:
```
## Similartly you can do preprocessing for review summary also.
```

[4] Featurization

[4.1] BAG OF WORDS

In [0]:
```
#BoW
count_vect = CountVectorizer() #in scikit-learn
count_vect.fit(preprocessed_reviews)
print("some feature names ", count_vect.get_feature_names()[:10])
```

```
print('='*50)

final_counts = count_vect.transform(preprocessed_reviews)
print("the type of count vectorizer ",type(final_counts))
print("the shape of out text BOW vectorizer ",final_counts.get_shape())
print("the number of unique words ", final_counts.get_shape()[1])
```

```
some feature names  ['aa', 'aahhhs', 'aback', 'abandon', 'abates', 'abb
ott', 'abby', 'abdominal', 'abiding', 'ability']
==================================================
the type of count vectorizer  <class 'scipy.sparse.csr.csr_matrix'>
the shape of out text BOW vectorizer  (4986, 12997)
the number of unique words  12997
```

[4.2] Bi-Grams and n-Grams.

In [0]:

```
#bi-gram, tri-gram and n-gram

#removing stop words like "not" should be avoided before building n-gra
ms
# count_vect = CountVectorizer(ngram_range=(1,2))
# please do read the CountVectorizer documentation http://scikit-learn.
org/stable/modules/generated/sklearn.feature_extraction.text.CountVecto
rizer.html

# you can choose these numebrs min_df=10, max_features=5000, of your ch
oice
count_vect = CountVectorizer(ngram_range=(1,2), min_df=10, max_features
=5000)
final_bigram_counts = count_vect.fit_transform(preprocessed_reviews)
print("the type of count vectorizer ",type(final_bigram_counts))
print("the shape of out text BOW vectorizer ",final_bigram_counts.get_s
hape())
print("the number of unique words including both unigrams and bigrams "
, final_bigram_counts.get_shape()[1])
```

```
the type of count vectorizer  <class 'scipy.sparse.csr.csr_matrix'>
the shape of out text BOW vectorizer  (4986, 3144)
the number of unique words including both unigrams and bigrams  3144
```

the number of unique words including both unigrams and bigrams 3144

[4.3] TF-IDF

In [0]:
```
tf_idf_vect = TfidfVectorizer(ngram_range=(1,2), min_df=10)
tf_idf_vect.fit(preprocessed_reviews)
print("some sample features(unique words in the corpus)",tf_idf_vect.ge
t_feature_names()[0:10])
print('='*50)

final_tf_idf = tf_idf_vect.transform(preprocessed_reviews)
print("the type of count vectorizer ",type(final_tf_idf))
print("the shape of out text TFIDF vectorizer ",final_tf_idf.get_shape
())
print("the number of unique words including both unigrams and bigrams "
, final_tf_idf.get_shape()[1])
```

```
some sample features(unique words in the corpus) ['ability', 'able', 'a
ble find', 'able get', 'absolute', 'absolutely', 'absolutely deliciou
s', 'absolutely love', 'absolutely no', 'according']
==================================================
the type of count vectorizer  <class 'scipy.sparse.csr.csr_matrix'>
the shape of out text TFIDF vectorizer  (4986, 3144)
the number of unique words including both unigrams and bigrams  3144
```

[4.4] Word2Vec

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

In [0]:
```
# Using Google News Word2Vectors

# in this project we are using a pretrained model by google
```

```
# its 3.3G file, once you load this into your memory
# it occupies ~9Gb, so please do this step only if you have >12G of ram
# we will provide a pickle file wich contains a dict ,
# and it contains all our courpus words as keys and  model[word] as val
ues
# To use this code-snippet, download "GoogleNews-vectors-negative300.bi
n"
# from https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edi
t
# it's 1.9GB in size.


# http://kavita-ganesan.com/gensim-word2vec-tutorial-starter-code/#.W17
SRFAzZPY
# you can comment this whole cell
# or change these varible according to your need

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_sentance,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', binary=True)
        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_trai
n_w2v = True, to train your own w2v ")
```

```
[('snack', 0.9951335191726685), ('calorie', 0.9946465492248535), ('wond
```

```
erful', 0.9946032166481018), ('excellent', 0.9944332838058472), ('espec
ially', 0.9941144585609436), ('baked', 0.9940600395202637), ('salted',
0.994047224521637), ('alternative', 0.9937226176261902), ('tasty', 0.99
36816692352295), ('healthy', 0.9936649799346924)]
==================================================
[('varieties', 0.9994194507598877), ('become', 0.9992934465408325), ('p
opcorn', 0.9992750883102417), ('de', 0.9992610216140747), ('miss', 0.99
92451071739197), ('melitta', 0.999218761920929), ('choice', 0.999210238
4567261), ('american', 0.9991837739944458), ('beef', 0.999178051948547
4), ('finish', 0.9991567134857178)]
```

In [0]:
```
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  3817
sample words  ['product', 'available', 'course', 'total', 'pretty', 'st
inky', 'right', 'nearby', 'used', 'ca', 'not', 'beat', 'great', 'receiv
ed', 'shipment', 'could', 'hardly', 'wait', 'try', 'love', 'call', 'ins
tead', 'removed', 'easily', 'daughter', 'designed', 'printed', 'use',
'car', 'windows', 'beautifully', 'shop', 'program', 'going', 'lot', 'fu
n', 'everywhere', 'like', 'tv', 'computer', 'really', 'good', 'idea',
'final', 'outstanding', 'window', 'everybody', 'asks', 'bought', 'mad
e']
```

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

[4.4.1.1] Avg W2v

In [0]:
```
# average Word2Vec
# compute average word2vec for each review.
sent_vectors = []; # the avg-w2v for each sentence/review is stored in
 this list
for sent in tqdm(list_of_sentance): # for each review/sentence
    sent_vec = np.zeros(50) # as word vectors are of zero length 50, yo
```

```
u might need to change this to 300 if you use google's w2v
    cnt_words =0; # num of words with a valid vector in the sentence/re
view
    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]))
```

```
100%|██████████| 4986/4986 [00:03<00:00, 1330.47it/s]
```

```
4986
50
```

[4.4.1.2] TFIDF weighted W2v

In [0]:
```
# S = ["abc def pqr", "def def def abc", "pqr pqr def"]
model = TfidfVectorizer()
tf_idf_matrix = model.fit_transform(preprocessed_reviews)
# we are converting a dictionary with word as a key, and the idf as a v
alue
dictionary = dict(zip(model.get_feature_names(), list(model.idf_)))
```

In [0]:
```
# 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 ce
ll_val = tfidf

tfidf_sent_vectors = []; # the tfidf-w2v for each sentence/review is st
ored in this list
row=0;
for sent in tqdm(list_of_sentance): # for each review/sentence
```

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

```
100%|██████████| 4986/4986 [00:20<00:00, 245.63it/s]
```

[5] Assignment 7: SVM

1. **Apply SVM on these feature sets**

   - SET 1:Review text, preprocessed one converted into vectors using (BOW)
   - SET 2:Review text, preprocessed one converted into vectors using (TFIDF)
   - SET 3:Review text, preprocessed one converted into vectors using (AVG W2v)
   - SET 4:Review text, preprocessed one converted into vectors using (TFIDF W2v)

2. **Procedure**

   - You need to work with 2 versions of SVM
     - Linear kernel
     - RBF kernel

- When you are working with linear kernel, use SGDClassifier' with hinge loss because it is computationally less expensive.
- When you are working with 'SGDClassifier' with hinge loss and trying to find the AUC score, you would have to use CalibratedClassifierCV
- Similarly, like kdtree of knn, when you are working with RBF kernel it's better to reduce the number of dimensions. You can put min_df = 10, max_features = 500 and consider a sample size of 40k points.

3. **Hyper paramter tuning (find best alpha in range [10^-4 to 10^4], and the best penalty among 'l1', 'l2')**

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

4. **Feature importance**

   - When you are working on the linear kernel with BOW or TFIDF please print the top 10 best features for each of the positive and negative classes.

5. **Feature engineering**

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

6. **Representation of results**

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

on both train and test.

- Along with plotting ROC curve, you need to print the confusion matrix with predicted and original labels of test data points. Please visualize your confusion matrices using seaborn heatmaps.

7. **Conclusion**

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

Note: Data Leakage

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

In [3]:
```
# Load the Drive helper and mount
from google.colab import drive
drive.mount('/content/drive')
```

```
Go to this URL in a browser: https://accounts.google.com/o/oauth2/auth?
client_id=947318989803-6bn6qk8qdgf4n4g3pfee6491hc0brc4i.apps.googleuser
content.com&redirect_uri=urn%3Aietf%3Awg%3Aoauth%3A2.0%3Aoob&scope=emai
l%20https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fdocs.test%20https%3A%2F%2
Fwww.googleapis.com%2Fauth%2Fdrive%20https%3A%2F%2Fwww.googleapis.com%2
Fauth%2Fdrive.photos.readonly%20https%3A%2F%2Fwww.googleapis.com%2Faut
h%2Fpeopleapi.readonly&response_type=code

Enter your authorization code:
··········
Mounted at /content/drive
```

In [4]:
```
cd drive/My Drive
```

```
/content/drive/My Drive
```

In [5]:
```
!ls
```

```
'03 Amazon Fine Food Reviews Analysis_KNN.ipynb'
'07 Amazon Fine Food Reviews Analysis_Support Vector Machines.ipynb'
'100 days ml.pdf'
 1553267672997_aravindresume.docx
 1.jpg
 2.jpg
 545.jpg
 5_6339045594849345659.pdf
'ACFrOgDlTE5HY5CK84NZo_AitoojMxCIJpRr_hyu9jql3e2FWbSaNDy-SXhhA5bUSAhiM-
tib_flUk1uTWtuLyITmV2jRdUaAgaVba1XSMuknXX08XLgp8Q3CcBz5Ps=.pdf'
 aravinds1.pdf
 aravinds2.pdf
'Assignment Guide Lines.pdf'
"Basic Interview Q's on ML.pdf"
 birthdate1.jpg
'Boarding Pass | Campus Recruiters.pdf'
'Chat backup_ 1498179864984.txt'
 Classroom
'Colab Notebooks'
'data science interview.pdf'
'deep learning.pdf'
'deep learning.Pdf'
 featureeng.sqlite
 final.sqlite
'Fundamentals-of-Mathematical-Statistics_Gupta-Kapoor- By EasyEngineeri
ng.net.pdf'
'Google Photos'
 hikeCam1491449461817_1491449566795.mp4
 IMG_1492188497114_1492188503978.jpg
 IMG_1492451321815_1492451326618.jpg
 IMG_1497623513343_1497623516687.jpg
 IMG_20170130_053213918.jpg
 IMG_20170130_053503405.jpg
```

```
 IMG_20170130_053538312.jpg
 IMG_20170130_053601196.jpg
 IMG_20170130_053804923.jpg
 IMG_20170130_053908690.jpg
 IMG_20170328_122908153_HDR.jpg
 IMG-20170414-WA0005_1492225019550.jpg
 IMG_20170415_072106832_1492221133024_1492224251474.jpg
 IMG_20170415_072106832.jpg
 Keras_Mnist.ipynb
 Lecture2.pdf
'machine learning book.pdf'
'mastery ml.pdf'
'Ng-MLY01-13 (1).pdf'
'Notes for Live Session on Feb 23 2019: How to code effecively and buil
d a web scraper.pdf'
 Pathos_5bviolin_5d.mp3
'Statistics and Probability for Data Science.pdf'
 thinkstats2.pdf
 TO.docx
'Week 2'
```

In [6]:
```
con = sqlite3.connect('final.sqlite')
final = pd.read_sql_query(""" SELECT * FROM Reviews WHERE Score != 3
 """, con)
final.head()
```

Out[6]:

	index	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator
0	138706	150524	0006641040	ACITT7DI6IDDL	shari zychinski	0

	index	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator
1	138688	150506	0006641040	A2IW4PEEKO2R0U	Tracy	1
2	138689	150507	0006641040	A1S4A3IQ2MU7V4	sally sue "sally sue"	1
3	138690	150508	0006641040	AZGXZ2UUK6X	Catherine Hallberg " (Kate)"	1
4	138691	150509	0006641040	A3CMRKGE0P909G	Teresa	3

In [0]:
```
from sklearn.model_selection import train_test_split
##Sorting data according to Time in ascending order for Time Based Spli
tting
time_sorted_data = final.sort_values('Time', axis=0, ascending=True, in
place=False, kind='quicksort', na_position='last')

time_sorted_data = time_sorted_data[:50000]

x = time_sorted_data['CleanedText'].values
y = time_sorted_data['Score']

# split the data set into train and test
X_train, X_test, Y_train, Y_test = train_test_split(x, y, test_size=0.3
, random_state=0,shuffle=False)
```

In [0]:
```
print(X_train.shape)
print(X_test.shape)
```

```
(35000,)
(15000,)
```

Applying SVM

[5.1] Linear SVM

[5.1.1] Applying Linear SVM on BOW, SET 1

In [0]:
```
# Please write all the code with proper documentation
#BoW
count_vect = CountVectorizer(min_df = 100)
X_train_vec = count_vect.fit_transform(X_train)
X_test_vec = count_vect.transform(X_test)
print("the type of count vectorizer :",type(X_train_vec))
print("the shape of out text BOW vectorizer : ",X_train_vec.get_shape
```

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

import warnings
warnings.filterwarnings('ignore')
# Data-preprocessing: Standardizing the data

from sklearn.preprocessing import StandardScaler
sc = StandardScaler(with_mean=False)
X_train_vec_standardized = sc.fit_transform(X_train_vec)
X_test_vec_standardized = sc.transform(X_test_vec)
```

```
the type of count vectorizer : <class 'scipy.sparse.csr.csr_matrix'>
the shape of out text BOW vectorizer :  (35000, 1457)
the number of unique words : 1457
```

In [0]:

```
# Importing libraries
from sklearn.linear_model import SGDClassifier
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import RandomizedSearchCV
from sklearn.metrics import accuracy_score,confusion_matrix,f1_score,pr
ecision_score,recall_score


param_grid = {'alpha': [0.0001,0.001, 0.01, 0.1, 1, 10,100,1000,10000]}
model = GridSearchCV(SGDClassifier(class_weight = 'balanced'), param_gr
id, scoring = 'roc_auc', cv=3 , n_jobs = -1,pre_dispatch=2)
model.fit(X_train_vec_standardized, Y_train)
print("Model with best parameters :\n",model.best_estimator_)
print("Accuracy of the model : ",model.score(X_test_vec_standardized, Y
_test))

optimal_alpha = model.best_estimator_.alpha
print("The optimal value of alpha(1/C) is : ",optimal_alpha)

print("\nBest Score:", model.best_score_)
```

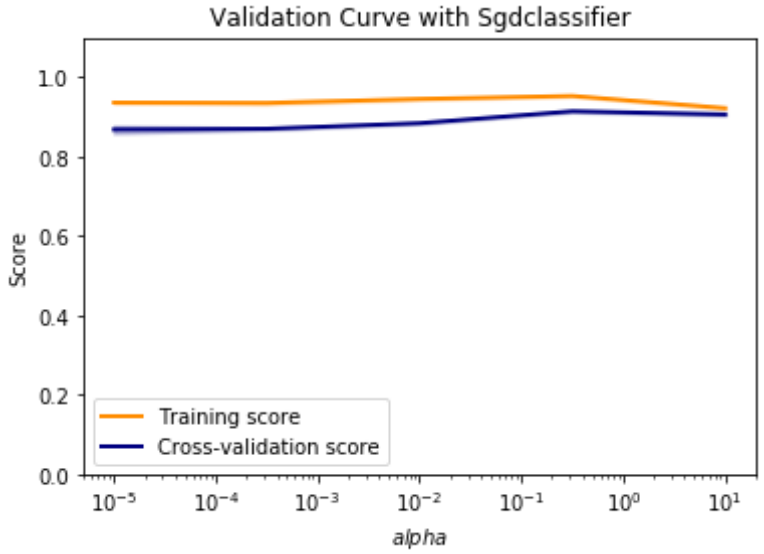
```
Model with best parameters :
 SGDClassifier(alpha=1, average=False, class_weight='balanced', epsilon

 0.1
```

```
=0.1,
       eta0=0.0, fit_intercept=True, l1_ratio=0.15,
       learning_rate='optimal', loss='hinge', max_iter=None, n_iter=Non
e,
       n_jobs=1, penalty='l2', power_t=0.5, random_state=None,
       shuffle=True, tol=None, verbose=0, warm_start=False)
Accuracy of the model :  0.9092037945197431
The optimal value of alpha(1/C) is :  1

Best Score: 0.9170572407378853
```

In [0]:

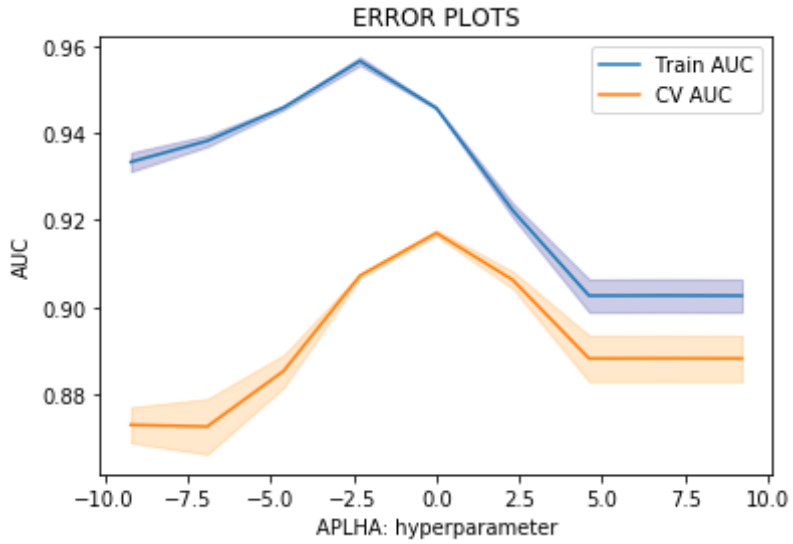
```
import matplotlib.pyplot as plt
import numpy as np

from sklearn.model_selection import validation_curve


param_range = np.logspace(-5, 1, 5)
train_scores, test_scores = validation_curve(
    SGDClassifier(class_weight = 'balanced'),X_train_vec_standardized,
Y_train, param_name="alpha", param_range=param_range,
    cv=3, scoring="roc_auc", n_jobs=1)
train_scores_mean = np.mean(train_scores, axis=1)
train_scores_std = np.std(train_scores, axis=1)
test_scores_mean = np.mean(test_scores, axis=1)
test_scores_std = np.std(test_scores, axis=1)

plt.title("Validation Curve with Sgdclassifier")
plt.xlabel("$alpha$")
plt.ylabel("Score")
plt.ylim(0.0, 1.1)
lw = 2
plt.semilogx(param_range, train_scores_mean, label="Training score",
             color="darkorange", lw=lw)
plt.fill_between(param_range, train_scores_mean - train_scores_std,
                 train_scores_mean + train_scores_std, alpha=0.2,
                 color="darkorange", lw=lw)
plt.semilogx(param_range, test_scores_mean, label="Cross-validation sco
re",
```

```
                 color="navy", lw=lw)
plt.fill_between(param_range, test_scores_mean - test_scores_std,
                 test_scores_mean + test_scores_std, alpha=0.2,
                 color="navy", lw=lw)
plt.legend(loc="best")
plt.show()
```



In [0]:
```
alpha = [10**-4, 10**-3, 10**-2,10**-1, 10**0, 10**1, 10**2, 10**3, 10**4]
train_auc= model.cv_results_['mean_train_score']
train_auc_std= model.cv_results_['std_train_score']
cv_auc = model.cv_results_['mean_test_score']
cv_auc_std= model.cv_results_['std_test_score']

plt.plot(np.log(alpha), train_auc, label='Train AUC')
# this code is copied from here: https://stackoverflow.com/a/48803361/4084039
plt.gca().fill_between(np.log(alpha),train_auc - train_auc_std,train_auc + train_auc_std,alpha=0.2,color='darkblue')

plt.plot(np.log(alpha), cv_auc, label='CV AUC')
```

```
# this code is copied from here: https://stackoverflow.com/a/48803361/4
084039
plt.gca().fill_between(np.log(alpha),cv_auc - cv_auc_std,cv_auc + cv_au
c_std,alpha=0.2,color='darkorange')
plt.legend()
plt.xlabel("APLHA: hyperparameter")
plt.ylabel("AUC")
plt.title("ERROR PLOTS")
plt.show()
```



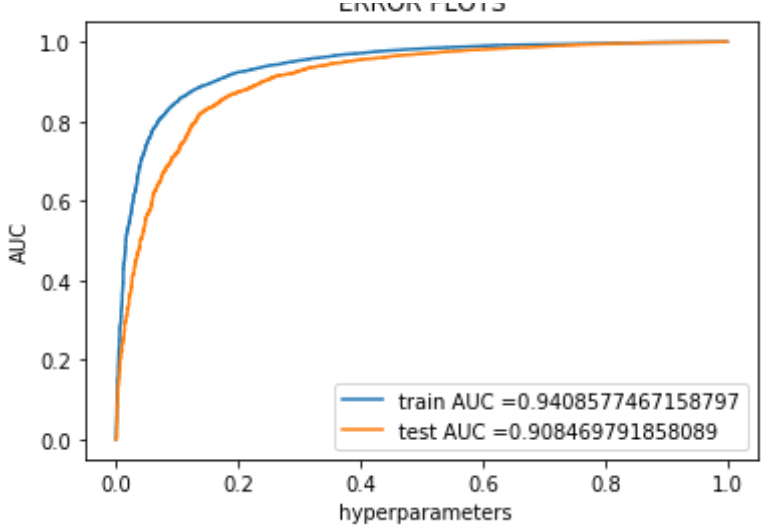
In [0]:
```
# SGDClassifier with Optimal value of alpha i.e.(1/C)
sgd = SGDClassifier(alpha=optimal_alpha, n_jobs=-1,class_weight = 'bala
nced')
sgd.fit(X_train_vec_standardized,Y_train)
predictions = sgd.predict(X_test_vec_standardized)
predictions1 = sgd.predict(X_train_vec_standardized)

# Variables that will be used for  making table in Conclusion part of t
his assignment
bow_grid_alpha = optimal_alpha
bow_grid_train_acc = model.score(X_test_vec_standardized, Y_test)*100
bow_grid_test_acc = accuracy_score(Y_test, predictions) * 100
```

In [0]:
```
# linear svc with sigmoid calibration
from sklearn.calibration import CalibratedClassifierCV
calib = CalibratedClassifierCV(sgd, method = "sigmoid", cv = "prefit")
calib.fit(X_train_vec_standardized,Y_train)
```

Out[0]:
```
CalibratedClassifierCV(base_estimator=SGDClassifier(alpha=1, average=Fa
lse, class_weight='balanced', epsilon=0.1,
       eta0=0.0, fit_intercept=True, l1_ratio=0.15,
       learning_rate='optimal', loss='hinge', max_iter=None, n_iter=Non
e,
       n_jobs=-1, penalty='l2', power_t=0.5, random_state=None,
       shuffle=True, tol=None, verbose=0, warm_start=False),
            cv='prefit', method='sigmoid')
```

In [0]:
```
train_fpr, train_tpr, thresholds = roc_curve(Y_train, calib.predict_pro
ba(X_train_vec_standardized)[:,1])
test_fpr, test_tpr, thresholds = roc_curve(Y_test, calib.predict_proba(
X_test_vec_standardized)[:,1])

plt.plot(train_fpr, train_tpr, label="train AUC ="+str(auc(train_fpr, t
rain_tpr)))
plt.plot(test_fpr, test_tpr, label="test AUC ="+str(auc(test_fpr, test_
tpr)))
plt.legend()
plt.xlabel(" hyperparameters")
plt.ylabel("AUC")
plt.title("ERROR PLOTS")
plt.show()
```

ERROR PLOTS



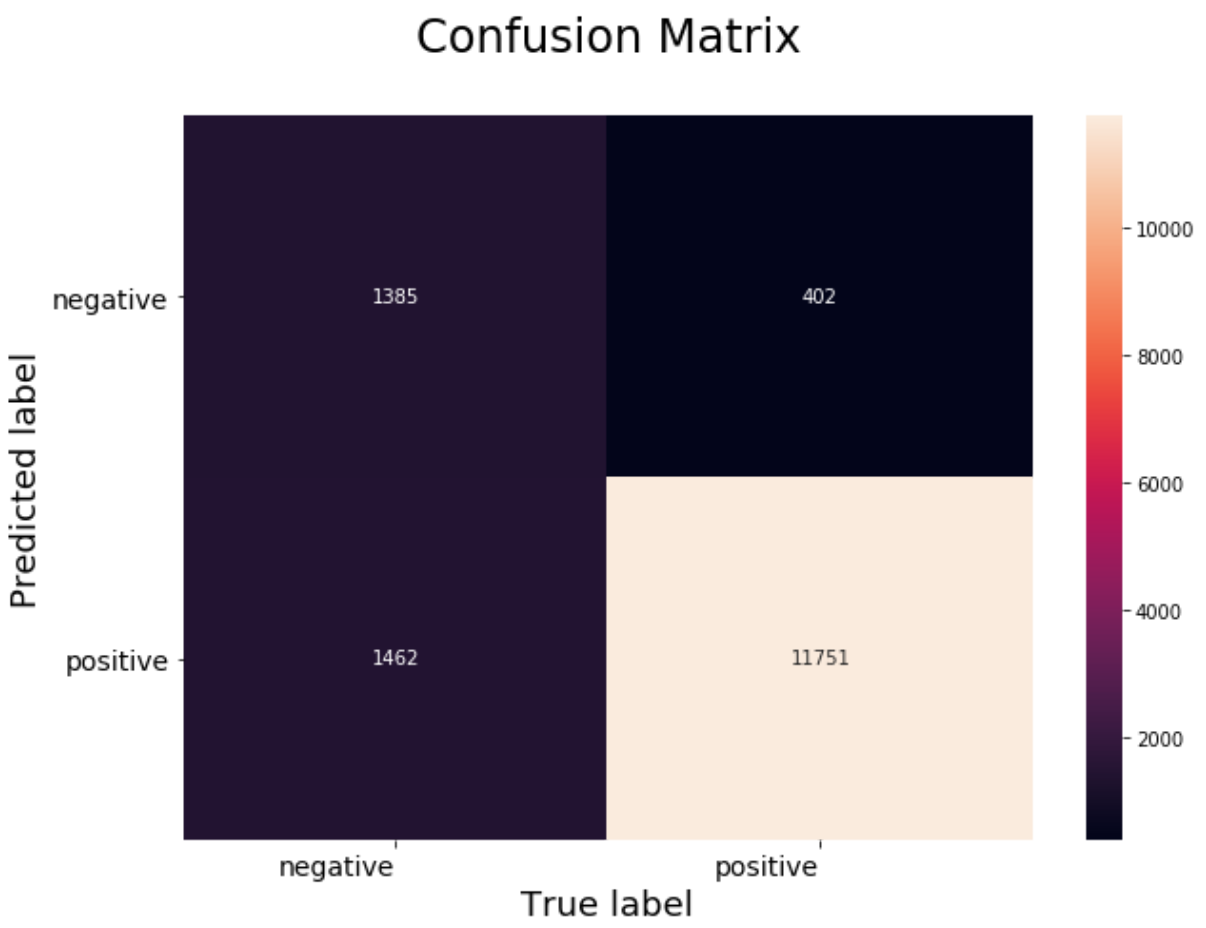
In [0]:
```
# evaluate accuracy on test data
acc = accuracy_score(Y_test, predictions) * 100
print('\nThe Test Accuracy of the SGDClassifier for alpha = %.3f is %f%
%' % (optimal_alpha, acc))

# evaluate precision
acc = precision_score(Y_test, predictions, pos_label = 1)
print('\nThe Test Precision of the SGDClassifier for alpha = %.3f is %f
' % (optimal_alpha, acc))

# evaluate recall
acc = recall_score(Y_test, predictions, pos_label = 1)
print('\nThe Test Recall of the SGDClassifier for alpha = %.3f is %f' %
 (optimal_alpha, acc))

# evaluate f1-score
acc = f1_score(Y_test, predictions, pos_label = 1)
print('\nThe Test F1-Score of the SGDClassifier for alpha = %.3f is %f'
 % (optimal_alpha, acc))
```

```
The Test Accuracy of the SGDClassifier for alpha = 1.000 is 87.573333%


The Test Precision of the SGDClassifier for alpha = 1.000 is 0.966922
```

```
The Test Precision of the SGDClassifier for alpha = 1.000 is 0.906922

The Test Recall of the SGDClassifier for alpha = 1.000 is 0.889351

The Test F1-Score of the SGDClassifier for alpha = 1.000 is 0.926516
```

In [0]:

```
# Code for drawing seaborn heatmaps on test data
class_names = ['negative','positive']
df_heatmap = pd.DataFrame(confusion_matrix(Y_test, predictions), index=
class_names, columns=class_names )
fig = plt.figure(figsize=(10,7))
heatmap = sns.heatmap(df_heatmap, annot=True, fmt="d")

# Setting tick labels for heatmap
heatmap.yaxis.set_ticklabels(heatmap.yaxis.get_ticklabels(), rotation=0
, ha='right', fontsize=14)
heatmap.xaxis.set_ticklabels(heatmap.xaxis.get_ticklabels(), rotation=0
, ha='right', fontsize=14)
plt.ylabel('Predicted label',size=18)
plt.xlabel('True label',size=18)
plt.title("Confusion Matrix\n",size=24)
plt.show()
```



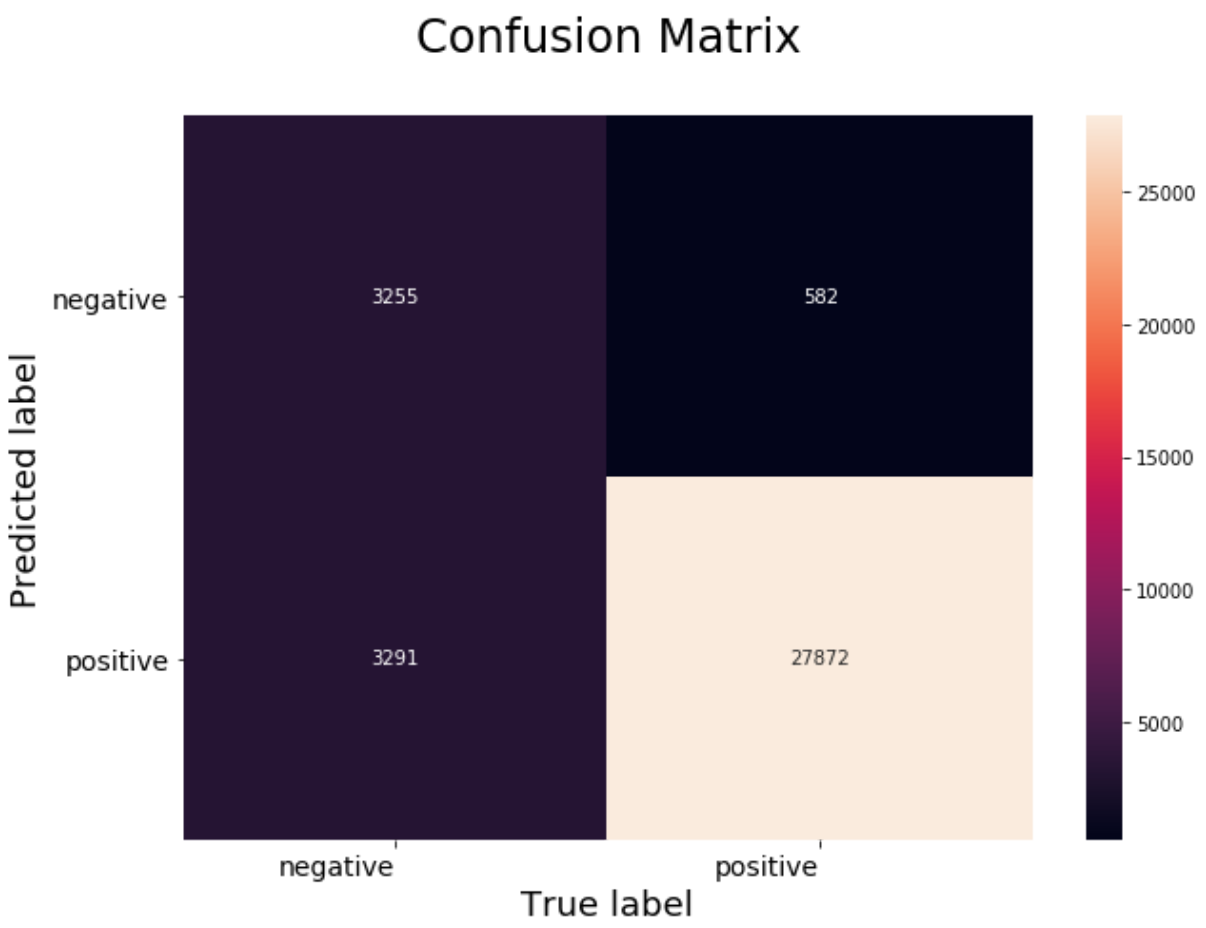
In [0]:
```
# evaluate accuracy on train data
acc = accuracy_score(Y_train, predictions1) * 100
print('\nThe Train Accuracy of the SGDClassifier for alpha = %.3f is %
f%%' % (optimal_alpha, acc))

# evaluate precision
acc = precision_score(Y_train, predictions1, pos_label = 1)
print('\nThe Train Precision of the SGDClassifier for alpha = %.3f is %
f' % (optimal_alpha, acc))

# evaluate recall
```

```
acc = recall_score(Y_train, predictions1, pos_label = 1)
print('\nThe Train Recall of the SGDClassifier for alpha = %.3f is %f'
% (optimal_alpha, acc))

# evaluate f1-score
acc = f1_score(Y_train, predictions1, pos_label = 1)
print('\nThe Train F1-Score of the SGDClassifier for alpha = %.3f is %f
' % (optimal_alpha, acc))
```

```
The Train Accuracy of the SGDClassifier for alpha = 1.000 is 88.934286%

The Train Precision of the SGDClassifier for alpha = 1.000 is 0.979546

The Train Recall of the SGDClassifier for alpha = 1.000 is 0.894394

The Train F1-Score of the SGDClassifier for alpha = 1.000 is 0.935035
```

In [0]:

```
# Code for drawing seaborn heatmaps on train data
class_names = ['negative','positive']
df_heatmap = pd.DataFrame(confusion_matrix(Y_train, predictions1), inde
x=class_names, columns=class_names )
fig = plt.figure(figsize=(10,7))
heatmap = sns.heatmap(df_heatmap, annot=True, fmt="d")

# Setting tick labels for heatmap
heatmap.yaxis.set_ticklabels(heatmap.yaxis.get_ticklabels(), rotation=0
, ha='right', fontsize=14)
heatmap.xaxis.set_ticklabels(heatmap.xaxis.get_ticklabels(), rotation=0
, ha='right', fontsize=14)
plt.ylabel('Predicted label',size=18)
plt.xlabel('True label',size=18)
plt.title("Confusion Matrix\n",size=24)
plt.show()
```



[5.1.2] Applying Linear SVM on TFIDF, SET 2

In [0]:
```
# Please write all the code with proper documentation
tf_idf_vect = TfidfVectorizer(min_df=100)
X_train_vec = tf_idf_vect.fit_transform(X_train)
X_test_vec = tf_idf_vect.transform(X_test)
print("the type of count vectorizer :",type(X_train_vec))
print("the shape of out text TFIDF vectorizer : ",X_train_vec.get_shape
())
```

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

# Data-preprocessing: Standardizing the data
sc = StandardScaler(with_mean=False)
X_train_vec_standardized = sc.fit_transform(X_train_vec)
X_test_vec_standardized = sc.transform(X_test_vec)
```

```
the type of count vectorizer : <class 'scipy.sparse.csr.csr_matrix'>
the shape of out text TFIDF vectorizer :  (35000, 1457)
the number of unique words : 1457
```

In [0]:

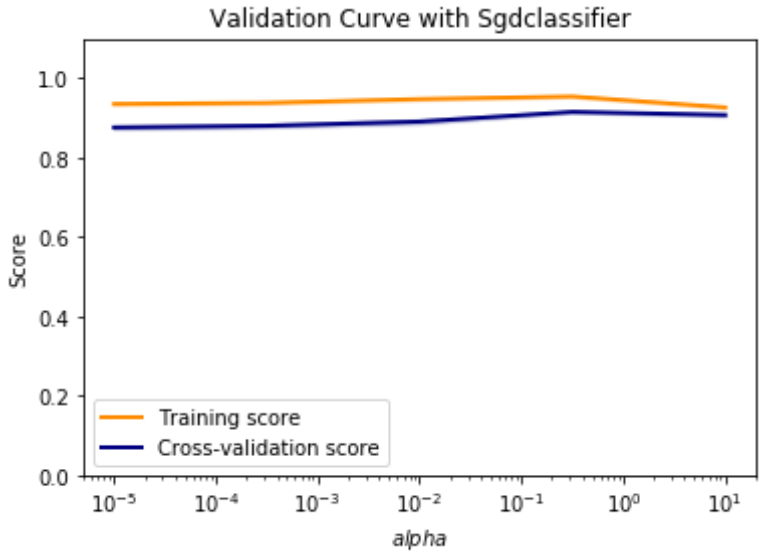
```
# Importing libraries
from sklearn.linear_model import SGDClassifier
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import RandomizedSearchCV
from sklearn.metrics import accuracy_score,confusion_matrix,f1_score,pr
ecision_score,recall_score


param_grid = {'alpha': [0.0001,0.001, 0.01, 0.1, 1, 10,100,1000,10000]}
model = GridSearchCV(SGDClassifier(class_weight = 'balanced'), param_gr
id, scoring = 'roc_auc', cv=3 , n_jobs = -1,pre_dispatch=2)
model.fit(X_train_vec_standardized, Y_train)
print("Model with best parameters :\n",model.best_estimator_)
print("Accuracy of the model : ",model.score(X_test_vec_standardized, Y
_test))

optimal_alpha = model.best_estimator_.alpha
print("The optimal value of alpha(1/C) is : ",optimal_alpha)

print("\nBest Score:", model.best_score_)
```

```
Model with best parameters :
 SGDClassifier(alpha=1, average=False, class_weight='balanced', epsilon
=0.1,
       eta0=0.0, fit_intercept=True, l1_ratio=0.15,
       learning_rate='optimal', loss='hinge', max_iter=None, n_iter=Non
e,
       n_jobs=1, penalty='l2', power_t=0.5, random_state=None,
       shuffle=True, tol=None, verbose=0, warm_start=False)
```

```
Accuracy of the model :  0.91050400118484
The optimal value of alpha(1/C) is :  1

Best Score: 0.9176820191112078
```

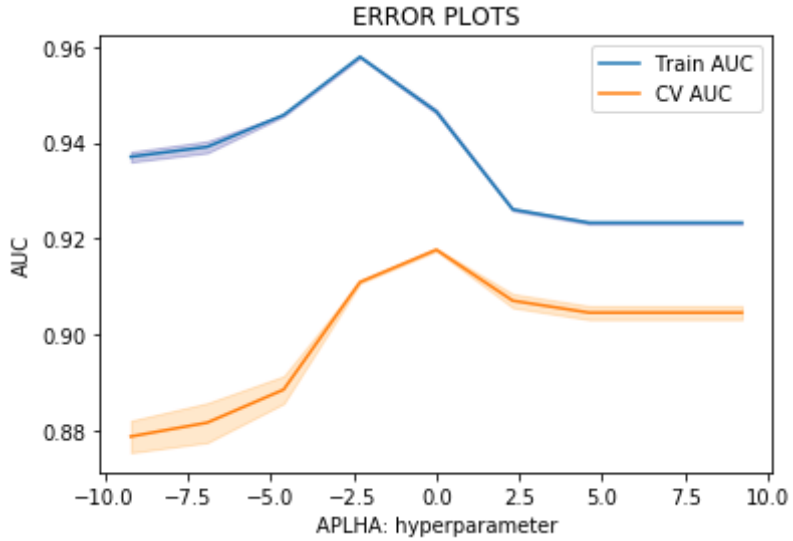
In [0]:

```
import matplotlib.pyplot as plt
import numpy as np

from sklearn.model_selection import validation_curve


param_range = np.logspace(-5, 1, 5)
train_scores, test_scores = validation_curve(
    SGDClassifier(class_weight = 'balanced'),X_train_vec_standardized,
Y_train, param_name="alpha", param_range=param_range,
    cv=3, scoring="roc_auc", n_jobs=1)
train_scores_mean = np.mean(train_scores, axis=1)
train_scores_std = np.std(train_scores, axis=1)
test_scores_mean = np.mean(test_scores, axis=1)
test_scores_std = np.std(test_scores, axis=1)

plt.title("Validation Curve with Sgdclassifier")
plt.xlabel("$alpha$")
plt.ylabel("Score")
plt.ylim(0.0, 1.1)
lw = 2
plt.semilogx(param_range, train_scores_mean, label="Training score",
             color="darkorange", lw=lw)
plt.fill_between(param_range, train_scores_mean - train_scores_std,
                 train_scores_mean + train_scores_std, alpha=0.2,
                 color="darkorange", lw=lw)
plt.semilogx(param_range, test_scores_mean, label="Cross-validation sco
re",
             color="navy", lw=lw)
plt.fill_between(param_range, test_scores_mean - test_scores_std,
                 test_scores_mean + test_scores_std, alpha=0.2,
                 color="navy", lw=lw)
plt.legend(loc="best")
plt.show()
```



In [0]:
```
alpha = [10**-4, 10**-3, 10**-2,10**-1, 10**0, 10**1, 10**2, 10**3, 10**4]
train_auc= model.cv_results_['mean_train_score']
train_auc_std= model.cv_results_['std_train_score']
cv_auc = model.cv_results_['mean_test_score']
cv_auc_std= model.cv_results_['std_test_score']

plt.plot(np.log(alpha), train_auc, label='Train AUC')
# this code is copied from here: https://stackoverflow.com/a/48803361/4084039
plt.gca().fill_between(np.log(alpha),train_auc - train_auc_std,train_auc + train_auc_std,alpha=0.2,color='darkblue')

plt.plot(np.log(alpha), cv_auc, label='CV AUC')
# this code is copied from here: https://stackoverflow.com/a/48803361/4084039
plt.gca().fill_between(np.log(alpha),cv_auc - cv_auc_std,cv_auc + cv_auc_std,alpha=0.2,color='darkorange')
plt.legend()
plt.xlabel("APLHA: hyperparameter")
plt.ylabel("AUC")
```

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



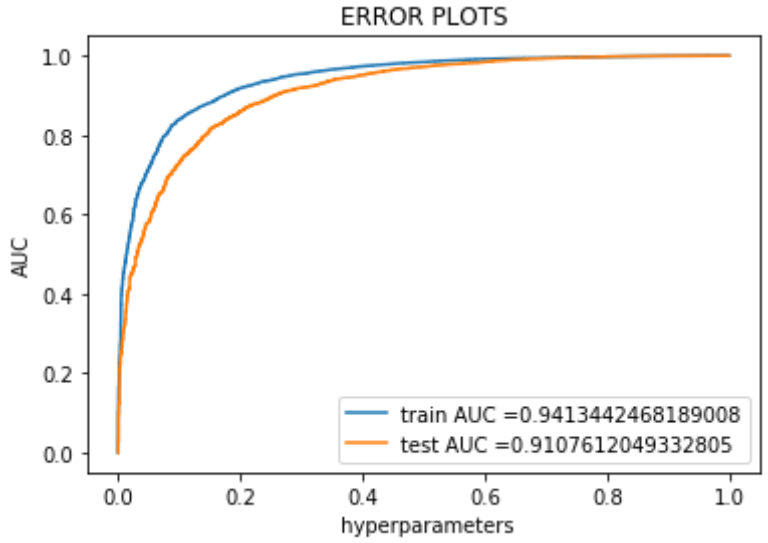
In [0]:
```
# SGDClassifier with Optimal value of alpha i.e.(1/C)
sgd = SGDClassifier(alpha=optimal_alpha, n_jobs=-1,class_weight = 'bala
nced')
sgd.fit(X_train_vec_standardized,Y_train)
predictions = sgd.predict(X_test_vec_standardized)
predictions1 = sgd.predict(X_train_vec_standardized)

# Variables that will be used for  making table in Conclusion part of t
his assignment
tfidf_grid_alpha = optimal_alpha
tfidf_grid_train_acc = model.score(X_test_vec_standardized, Y_test)*100
tfidf_grid_test_acc = accuracy_score(Y_test, predictions) * 100
```

In [0]:
```
calib = CalibratedClassifierCV(sgd, method = "sigmoid", cv = "prefit")
calib.fit(X_train_vec_standardized,Y_train)
```

Out[0]:
```
CalibratedClassifierCV(base_estimator=SGDClassifier(alpha=1, average=Fa
lse, class_weight='balanced', epsilon=0.1,
       eta0=0.0, fit_intercept=True, l1_ratio=0.15,
```

```
        learning_rate='optimal', loss='hinge', max_iter=None, n_iter=Non
e,
        n_jobs=-1, penalty='l2', power_t=0.5, random_state=None,
        shuffle=True, tol=None, verbose=0, warm_start=False),
             cv='prefit', method='sigmoid')
```

In [0]:
```
train_fpr, train_tpr, thresholds = roc_curve(Y_train, calib.predict_pro
ba(X_train_vec_standardized)[:,1])
test_fpr, test_tpr, thresholds = roc_curve(Y_test, calib.predict_proba(
X_test_vec_standardized)[:,1])

plt.plot(train_fpr, train_tpr, label="train AUC ="+str(auc(train_fpr, t
rain_tpr)))
plt.plot(test_fpr, test_tpr, label="test AUC ="+str(auc(test_fpr, test_
tpr)))
plt.legend()
plt.xlabel(" hyperparameters")
plt.ylabel("AUC")
plt.title("ERROR PLOTS")
plt.show()
```



In [0]:
```
# evaluate accuracy on test data
```

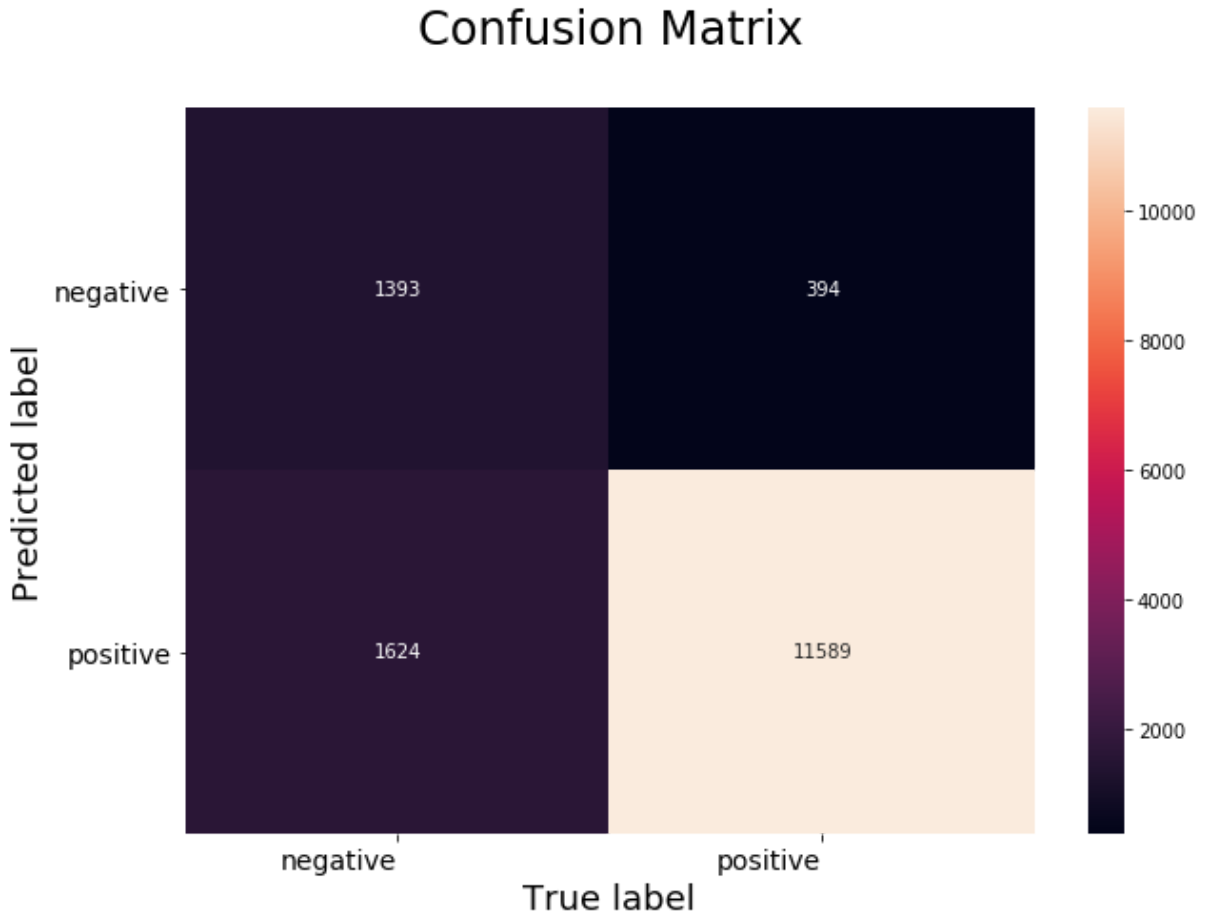
```
acc = accuracy_score(Y_test, predictions) * 100
print('\nThe Test Accuracy of the SGDClassifier for alpha = %.3f is %f%
%' % (optimal_alpha, acc))

# evaluate precision
acc = precision_score(Y_test, predictions, pos_label = 1)
print('\nThe Test Precision of the SGDClassifier for alpha = %.3f is %f
' % (optimal_alpha, acc))

# evaluate recall
acc = recall_score(Y_test, predictions, pos_label = 1)
print('\nThe Test Recall of the SGDClassifier for alpha = %.3f is %f' %
 (optimal_alpha, acc))

# evaluate f1-score
acc = f1_score(Y_test, predictions, pos_label = 1)
print('\nThe Test F1-Score of the SGDClassifier for alpha = %.3f is %f'
 % (optimal_alpha, acc))
```

```
The Test Accuracy of the SGDClassifier for alpha = 1.000 is 86.546667%

The Test Precision of the SGDClassifier for alpha = 1.000 is 0.967120

The Test Recall of the SGDClassifier for alpha = 1.000 is 0.877091

The Test F1-Score of the SGDClassifier for alpha = 1.000 is 0.919908
```

In [0]:

```
# Code for drawing seaborn heatmaps on test data
class_names = ['negative','positive']
df_heatmap = pd.DataFrame(confusion_matrix(Y_test, predictions), index=
class_names, columns=class_names )
fig = plt.figure(figsize=(10,7))
heatmap = sns.heatmap(df_heatmap, annot=True, fmt="d")

# Setting tick labels for heatmap
heatmap.yaxis.set_ticklabels(heatmap.yaxis.get_ticklabels(), rotation=0
, ha='right', fontsize=14)
heatmap.xaxis.set_ticklabels(heatmap.xaxis.get_ticklabels(), rotation=0
, ha='right', fontsize=14)
```

```
plt.ylabel('Predicted label',size=18)
plt.xlabel('True label',size=18)
plt.title("Confusion Matrix\n",size=24)
plt.show()
```



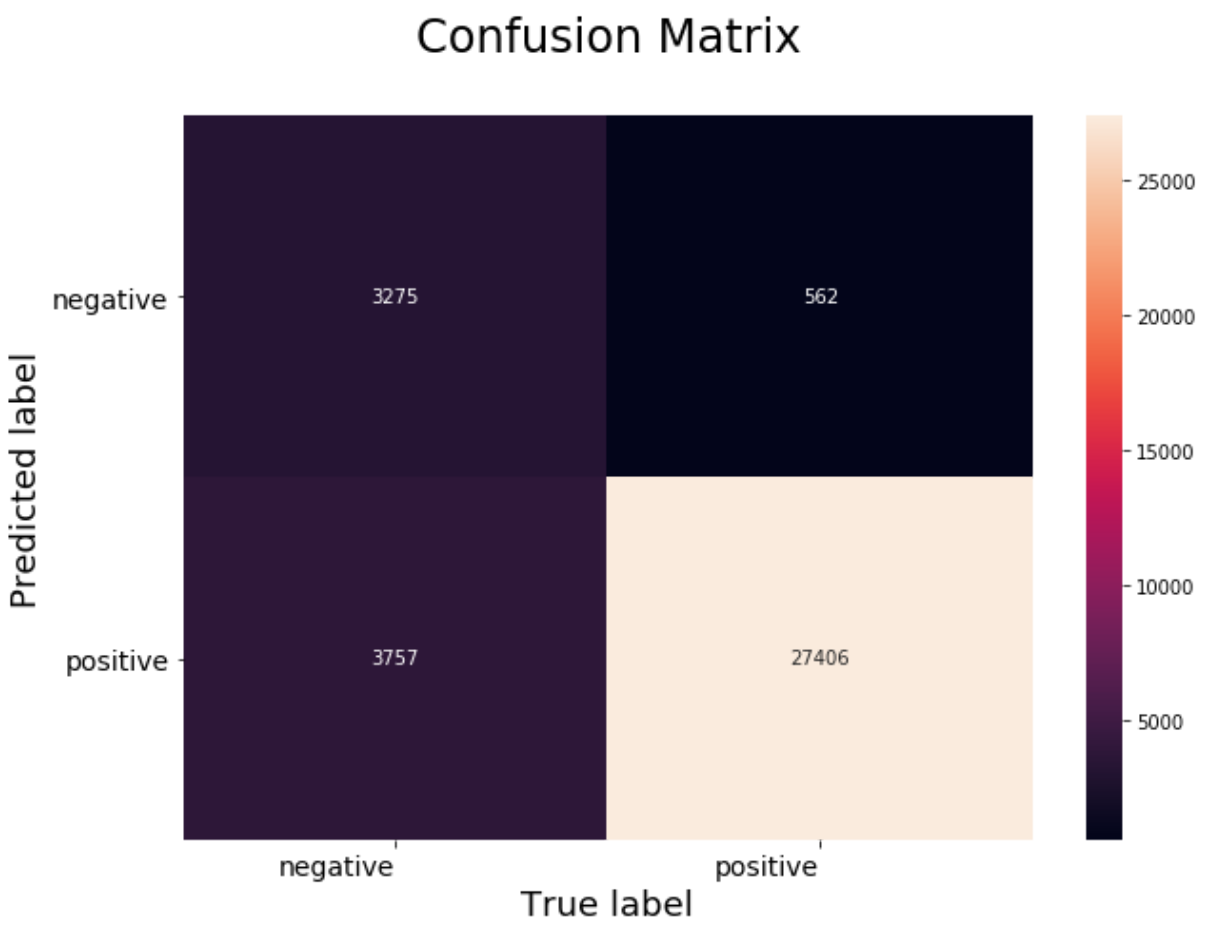
In [0]:
```
# evaluate accuracy on train data
acc = accuracy_score(Y_train, predictions1) * 100
print('\nThe Train Accuracy of the SGDClassifier for alpha = %.3f is %
f%%' % (optimal_alpha, acc))

# evaluate precision
```

```
acc = precision_score(Y_train, predictions1, pos_label = 1)
print('\nThe Train Precision of the SGDClassifier for alpha = %.3f is %
f' % (optimal_alpha, acc))

# evaluate recall
acc = recall_score(Y_train, predictions1, pos_label = 1)
print('\nThe Train Recall of the SGDClassifier for alpha = %.3f is %f'
% (optimal_alpha, acc))

# evaluate f1-score
acc = f1_score(Y_train, predictions1, pos_label = 1)
print('\nThe Train F1-Score of the SGDClassifier for alpha = %.3f is %f
' % (optimal_alpha, acc))
```

```
The Train Accuracy of the SGDClassifier for alpha = 1.000 is 87.660000%

The Train Precision of the SGDClassifier for alpha = 1.000 is 0.979906

The Train Recall of the SGDClassifier for alpha = 1.000 is 0.879440

The Train F1-Score of the SGDClassifier for alpha = 1.000 is 0.926959
```

In [0]:

```
# Code for drawing seaborn heatmaps on train data
class_names = ['negative','positive']
df_heatmap = pd.DataFrame(confusion_matrix(Y_train, predictions1), inde
x=class_names, columns=class_names )
fig = plt.figure(figsize=(10,7))
heatmap = sns.heatmap(df_heatmap, annot=True, fmt="d")

# Setting tick labels for heatmap
heatmap.yaxis.set_ticklabels(heatmap.yaxis.get_ticklabels(), rotation=0
, ha='right', fontsize=14)
heatmap.xaxis.set_ticklabels(heatmap.xaxis.get_ticklabels(), rotation=0
, ha='right', fontsize=14)
plt.ylabel('Predicted label',size=18)
plt.xlabel('True label',size=18)
plt.title("Confusion Matrix\n",size=24)
plt.show()
```



[5.1.3] Applying Linear SVM on AVG W2V, SET 3

In [0]:
```
# Please write all the code with proper documentation
# List of sentence in X_train text
sent_of_train=[]
for sent in X_train:
    sent_of_train.append(sent.split())

# List of sentence in X_est text
```

```
sent_of_test=[]
for sent in X_test:
    sent_of_test.append(sent.split())

# Train your own Word2Vec model using your own train text corpus
# min_count = 5 considers only words that occured atleast 5 times
w2v_model=Word2Vec(sent_of_train,min_count=5,size=50, workers=4)

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

```
number of words that occured minimum 5 times  7936
```

In [0]:

```
# compute average word2vec for each review for X_train .
train_vectors = [];
for sent in sent_of_train:
    sent_vec = np.zeros(50)
    cnt_words =0;
    for word in sent: #
        if word in w2v_words:
            vec = w2v_model.wv[word]
            sent_vec += vec
            cnt_words += 1
    if cnt_words != 0:
        sent_vec /= cnt_words
    train_vectors.append(sent_vec)

# compute average word2vec for each review for X_test .
test_vectors = [];
for sent in sent_of_test:
    sent_vec = np.zeros(50)
    cnt_words =0;
    for word in sent: #
        if word in w2v_words:
            vec = w2v_model.wv[word]
            sent_vec += vec
            cnt_words += 1
    if cnt_words != 0:
        sent_vec /= cnt_words
```

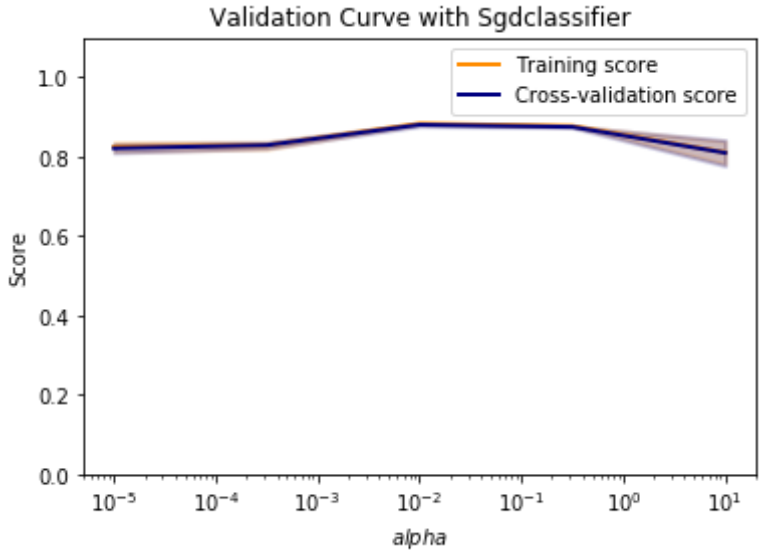
```
    test_vectors.append(sent_vec)

# Data-preprocessing: Standardizing the data
sc = StandardScaler()
X_train_vec_standardized = sc.fit_transform(train_vectors)
X_test_vec_standardized = sc.transform(test_vectors)
```

In [0]:
```
param_grid = {'alpha': [0.0001,0.001, 0.01, 0.1, 1, 10,100,1000,10000]}
model = GridSearchCV(SGDClassifier(class_weight = 'balanced'), param_gr
id, scoring = 'roc_auc', cv=3 , n_jobs = -1,pre_dispatch=2)
model.fit(X_train_vec_standardized, Y_train)
print("Model with best parameters :\n",model.best_estimator_)
print("Accuracy of the model : ",model.score(X_test_vec_standardized, Y
_test))

optimal_alpha = model.best_estimator_.alpha
print("The optimal value of alpha(1/C) is : ",optimal_alpha)

print("\nBest Score:", model.best_score_)
```

```
Model with best parameters :
 SGDClassifier(alpha=0.01, average=False, class_weight='balanced', epsi
lon=0.1,
       eta0=0.0, fit_intercept=True, l1_ratio=0.15,
       learning_rate='optimal', loss='hinge', max_iter=None, n_iter=Non
e,
       n_jobs=1, penalty='l2', power_t=0.5, random_state=None,
       shuffle=True, tol=None, verbose=0, warm_start=False)
Accuracy of the model :  0.8759619570541316
The optimal value of alpha(1/C) is :  0.01

Best Score: 0.8807630971721992
```

In [0]:
```
import matplotlib.pyplot as plt
import numpy as np

from sklearn.model_selection import validation_curve
```

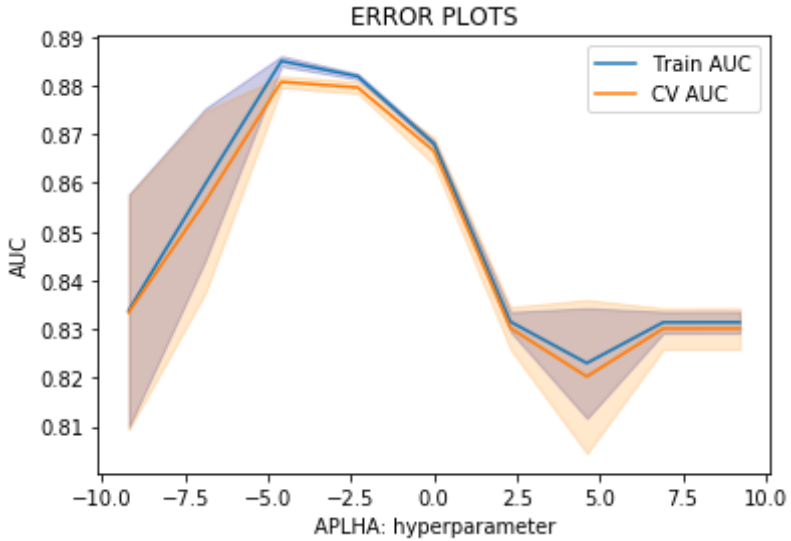
```
param_range = np.logspace(-5, 1, 5)
train_scores, test_scores = validation_curve(
    SGDClassifier(class_weight = 'balanced'),X_train_vec_standardized,
Y_train, param_name="alpha", param_range=param_range,
    cv=3, scoring="roc_auc", n_jobs=1)
train_scores_mean = np.mean(train_scores, axis=1)
train_scores_std = np.std(train_scores, axis=1)
test_scores_mean = np.mean(test_scores, axis=1)
test_scores_std = np.std(test_scores, axis=1)

plt.title("Validation Curve with Sgdclassifier")
plt.xlabel("$alpha$")
plt.ylabel("Score")
plt.ylim(0.0, 1.1)
lw = 2
plt.semilogx(param_range, train_scores_mean, label="Training score",
             color="darkorange", lw=lw)
plt.fill_between(param_range, train_scores_mean - train_scores_std,
                 train_scores_mean + train_scores_std, alpha=0.2,
                 color="darkorange", lw=lw)
plt.semilogx(param_range, test_scores_mean, label="Cross-validation sco
re",
             color="navy", lw=lw)
plt.fill_between(param_range, test_scores_mean - test_scores_std,
                 test_scores_mean + test_scores_std, alpha=0.2,
                 color="navy", lw=lw)
plt.legend(loc="best")
plt.show()
```



In [0]:
```
# SGDClassifier with Optimal value of alpha i.e.(1/C)
sgd = SGDClassifier(alpha=optimal_alpha, n_jobs=-1,class_weight = 'balanced')
sgd.fit(X_train_vec_standardized,Y_train)
predictions = sgd.predict(X_test_vec_standardized)
predictions1 = sgd.predict(X_train_vec_standardized)

# Variables that will be used for  making table in Conclusion part of this assignment
avg_w2v_grid_alpha = optimal_alpha
avg_w2v_grid_train_acc = model.score(X_test_vec_standardized, Y_test)*100
avg_w2v_grid_test_acc = accuracy_score(Y_test, predictions) * 100
```

In [0]:
```
alpha = [10**-4, 10**-3, 10**-2,10**-1, 10**0, 10**1, 10**2, 10**3, 10**4]
train_auc= model.cv_results_['mean_train_score']
train_auc_std= model.cv_results_['std_train_score']
cv_auc = model.cv_results_['mean_test_score']
cv_auc_std= model.cv_results_['std_test_score']
```

```
plt.plot(np.log(alpha), train_auc, label='Train AUC')
# this code is copied from here: https://stackoverflow.com/a/48803361/4
084039
plt.gca().fill_between(np.log(alpha),train_auc - train_auc_std,train_au
c + train_auc_std,alpha=0.2,color='darkblue')

plt.plot(np.log(alpha), cv_auc, label='CV AUC')
# this code is copied from here: https://stackoverflow.com/a/48803361/4
084039
plt.gca().fill_between(np.log(alpha),cv_auc - cv_auc_std,cv_auc + cv_au
c_std,alpha=0.2,color='darkorange')
plt.legend()
plt.xlabel("APLHA: hyperparameter")
plt.ylabel("AUC")
plt.title("ERROR PLOTS")
plt.show()
```

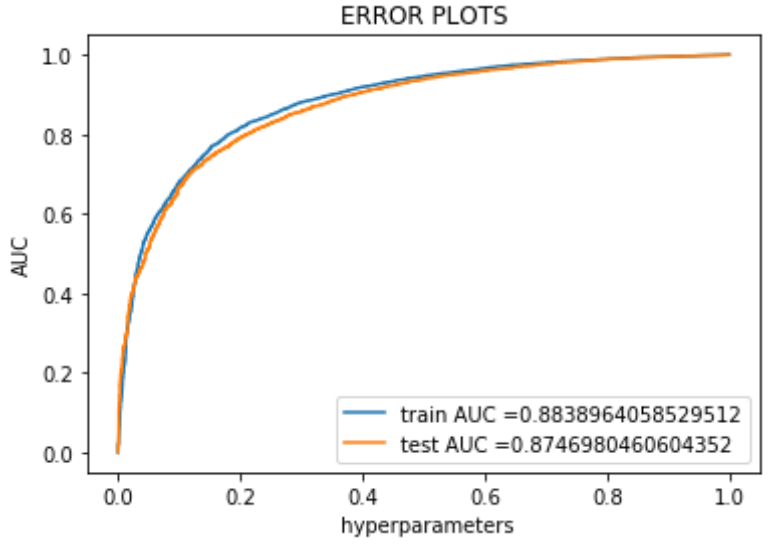


In [0]:
```
calib = CalibratedClassifierCV(sgd, method = "sigmoid", cv = "prefit")
calib.fit(X_train_vec_standardized,Y_train)
```

Out[0]:
```
CalibratedClassifierCV(base_estimator=SGDClassifier(alpha=0.01, average
=False, class_weight='balanced', epsilon=0.1,
```

```
        eta0=0.0, fit_intercept=True, l1_ratio=0.15,
        learning_rate='optimal', loss='hinge', max_iter=None, n_iter=Non
e,
        n_jobs=-1, penalty='l2', power_t=0.5, random_state=None,
        shuffle=True, tol=None, verbose=0, warm_start=False),
             cv='prefit', method='sigmoid')
```

In [0]:
```
train_fpr, train_tpr, thresholds = roc_curve(Y_train, calib.predict_pro
ba(X_train_vec_standardized)[:,1])
test_fpr, test_tpr, thresholds = roc_curve(Y_test, calib.predict_proba(
X_test_vec_standardized)[:,1])

plt.plot(train_fpr, train_tpr, label="train AUC ="+str(auc(train_fpr, t
rain_tpr)))
plt.plot(test_fpr, test_tpr, label="test AUC ="+str(auc(test_fpr, test_
tpr)))
plt.legend()
plt.xlabel(" hyperparameters")
plt.ylabel("AUC")
plt.title("ERROR PLOTS")
plt.show()
```

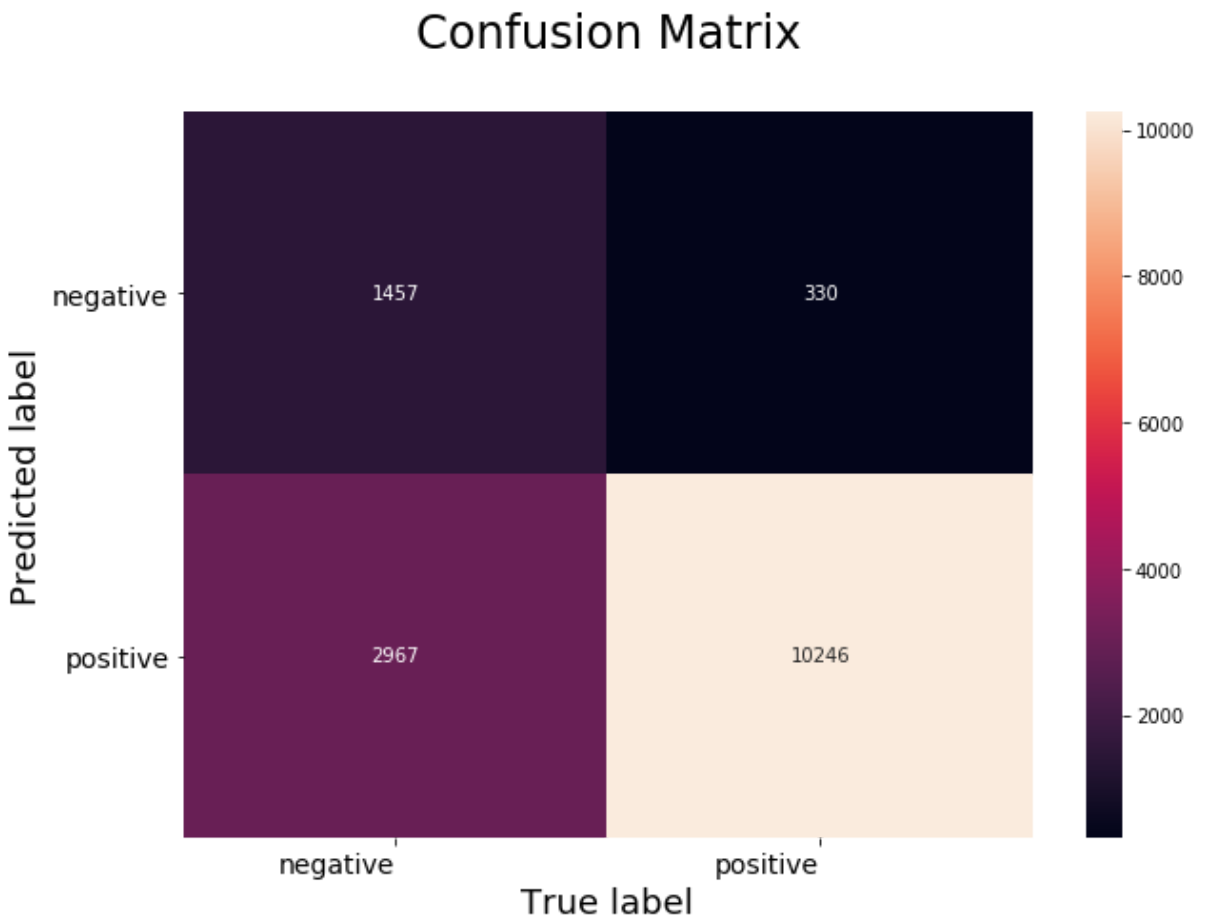


In [0]:
```
# evaluate accuracy on test data
acc = accuracy_score(Y_test, predictions) * 100
print('\nThe Test Accuracy of the SGDClassifier for alpha = %.3f is %f%
%' % (optimal_alpha, acc))

# evaluate precision
acc = precision_score(Y_test, predictions, pos_label = 1)
print('\nThe Test Precision of the SGDClassifier for alpha = %.3f is %f
' % (optimal_alpha, acc))

# evaluate recall
acc = recall_score(Y_test, predictions, pos_label = 1)
print('\nThe Test Recall of the SGDClassifier for alpha = %.3f is %f' %
 (optimal_alpha, acc))

# evaluate f1-score
acc = f1_score(Y_test, predictions, pos_label = 1)
print('\nThe Test F1-Score of the SGDClassifier for alpha = %.3f is %f'
 % (optimal_alpha, acc))
```

```
The Test Accuracy of the SGDClassifier for alpha = 0.010 is 78.020000%

The Test Precision of the SGDClassifier for alpha = 0.010 is 0.968797

The Test Recall of the SGDClassifier for alpha = 0.010 is 0.775448

The Test F1-Score of the SGDClassifier for alpha = 0.010 is 0.861407
```

In [0]:
```
# Code for drawing seaborn heatmaps on test data
class_names = ['negative','positive']
df_heatmap = pd.DataFrame(confusion_matrix(Y_test, predictions), index=
class_names, columns=class_names )
fig = plt.figure(figsize=(10,7))
heatmap = sns.heatmap(df_heatmap, annot=True, fmt="d")

# Setting tick labels for heatmap
heatmap.yaxis.set_ticklabels(heatmap.yaxis.get_ticklabels(), rotation=0
, ha='right', fontsize=14)
heatmap.xaxis.set_ticklabels(heatmap.xaxis.get_ticklabels(), rotation=0
```

```
, ha='right', fontsize=14)
plt.ylabel('Predicted label',size=18)
plt.xlabel('True label',size=18)
plt.title("Confusion Matrix\n",size=24)
plt.show()
```



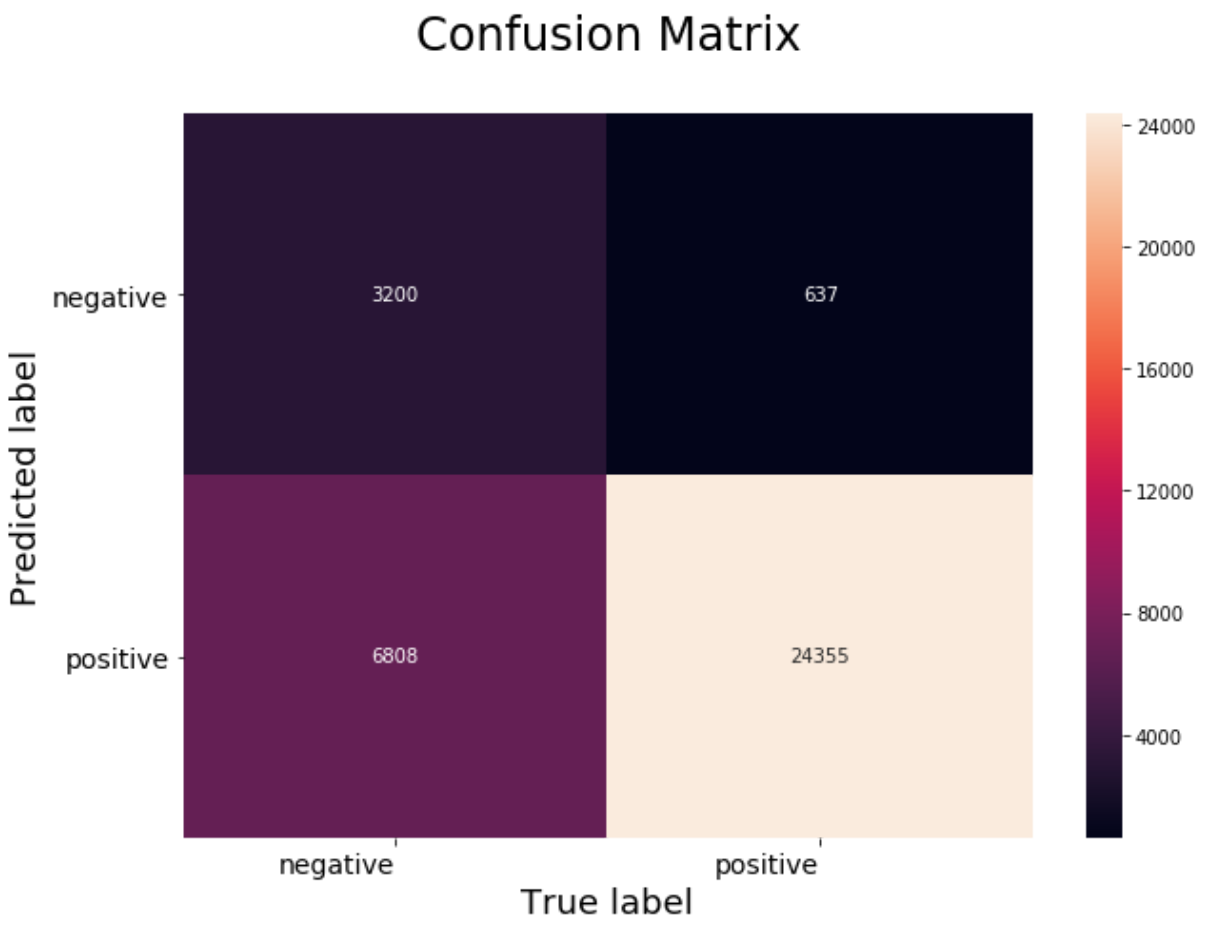
In [0]:
```
# evaluate accuracy on train data
acc = accuracy_score(Y_train, predictions1) * 100
print('\nThe Train Accuracy of the SGDClassifier for alpha = %.3f is %
f%%' % (optimal_alpha, acc))
```

```
# evaluate precision
acc = precision_score(Y_train, predictions1, pos_label = 1)
print('\nThe Train Precision of the SGDClassifier for alpha = %.3f is %
f' % (optimal_alpha, acc))

# evaluate recall
acc = recall_score(Y_train, predictions1, pos_label = 1)
print('\nThe Train Recall of the SGDClassifier for alpha = %.3f is %f'
% (optimal_alpha, acc))

# evaluate f1-score
acc = f1_score(Y_train, predictions1, pos_label = 1)
print('\nThe Train F1-Score of the SGDClassifier for alpha = %.3f is %f
' % (optimal_alpha, acc))
```

```
The Train Accuracy of the SGDClassifier for alpha = 0.010 is 78.728571%

The Train Precision of the SGDClassifier for alpha = 0.010 is 0.974512

The Train Recall of the SGDClassifier for alpha = 0.010 is 0.781536

The Train F1-Score of the SGDClassifier for alpha = 0.010 is 0.867421
```

In [0]:

```
# Code for drawing seaborn heatmaps on train data
class_names = ['negative','positive']
df_heatmap = pd.DataFrame(confusion_matrix(Y_train, predictions1), inde
x=class_names, columns=class_names )
fig = plt.figure(figsize=(10,7))
heatmap = sns.heatmap(df_heatmap, annot=True, fmt="d")

# Setting tick labels for heatmap
heatmap.yaxis.set_ticklabels(heatmap.yaxis.get_ticklabels(), rotation=0
, ha='right', fontsize=14)
heatmap.xaxis.set_ticklabels(heatmap.xaxis.get_ticklabels(), rotation=0
, ha='right', fontsize=14)
plt.ylabel('Predicted label',size=18)
plt.xlabel('True label',size=18)
plt.title("Confusion Matrix\n",size=24)
plt.show()
```



[5.1.4] Applying Linear SVM on TFIDF W2V, SET 4

In [0]:
```
# Please write all the code with proper documentation
# TF-IDF weighted Word2Vec
tf_idf_vect = TfidfVectorizer()

# final_tf_idf1 is the sparse matrix with row= sentence, col=word and c
ell_val = tfidf
final_tf_idf1 = tf_idf_vect.fit_transform(X_train)
```

```
# tfidf words/col-names
tfidf_feat = tf_idf_vect.get_feature_names()

# compute TFIDF Weighted Word2Vec for each review for X_test .
tfidf_test_vectors = [];
row=0;
for sent in sent_of_test:
    sent_vec = np.zeros(50)
    weight_sum =0;
    for word in sent:
        if word in w2v_words:
            vec = w2v_model.wv[word]
            # obtain the tf_idfidf of a word in a sentence/review
            tf_idf = final_tf_idf1[row, tfidf_feat.index(word)]
            sent_vec += (vec * tf_idf)
            weight_sum += tf_idf
    if weight_sum != 0:
        sent_vec /= weight_sum
    tfidf_test_vectors.append(sent_vec)
    row += 1
```

In [0]:
```
# compute TFIDF Weighted Word2Vec for each review for X_train .
tfidf_train_vectors = [];
row=0;
for sent in sent_of_train:
    sent_vec = np.zeros(50)
    weight_sum =0;
    for word in sent:
        if word in w2v_words:
            vec = w2v_model.wv[word]
            # obtain the tf_idfidf of a word in a sentence/review
            tf_idf = final_tf_idf1[row, tfidf_feat.index(word)]
            sent_vec += (vec * tf_idf)
            weight_sum += tf_idf
    if weight_sum != 0:
        sent_vec /= weight_sum
    tfidf_train_vectors.append(sent_vec)
    row += 1
```

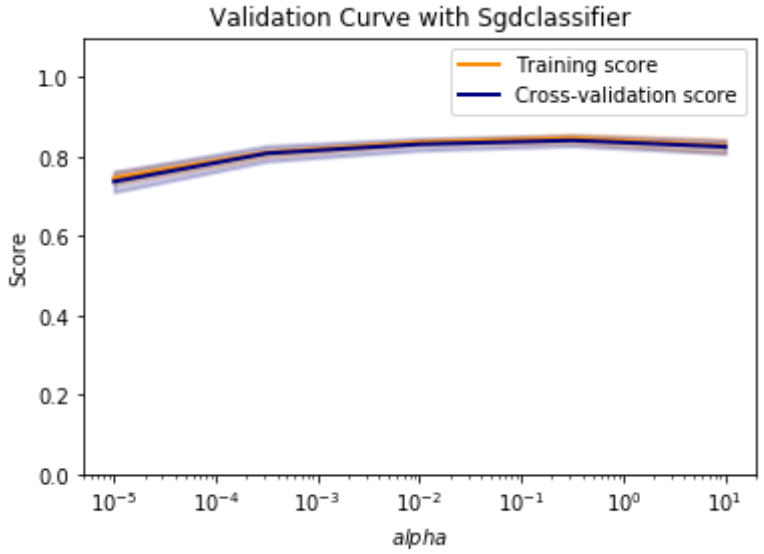
```
# Data-preprocessing: Standardizing the data
sc = StandardScaler()
X_train_vec_standardized = sc.fit_transform(tfidf_train_vectors)
X_test_vec_standardized = sc.transform(tfidf_test_vectors)
```

In [0]:
```
param_grid = {'alpha': [0.0001,0.001, 0.01, 0.1, 1, 10,100,1000,10000]}
model = GridSearchCV(SGDClassifier(), param_grid, scoring = 'roc_auc',
cv=10 , n_jobs = -1,pre_dispatch=2)
model.fit(X_train_vec_standardized, Y_train)
print("Model with best parameters :\n",model.best_estimator_)
print("Accuracy of the model : ",model.score(X_test_vec_standardized, Y
_test))

optimal_alpha = model.best_estimator_.alpha
print("The optimal value of alpha(1/C) is : ",optimal_alpha)

print("\nBest Score:", model.best_score_)
```

```
Model with best parameters :
 SGDClassifier(alpha=0.1, average=False, class_weight=None, epsilon=0.
1,
       eta0=0.0, fit_intercept=True, l1_ratio=0.15,
       learning_rate='optimal', loss='hinge', max_iter=None, n_iter=Non
e,
       n_jobs=1, penalty='l2', power_t=0.5, random_state=None,
       shuffle=True, tol=None, verbose=0, warm_start=False)
Accuracy of the model :  0.6063530935241195
The optimal value of alpha(1/C) is :  0.1

Best Score: 0.8512616825443738
```

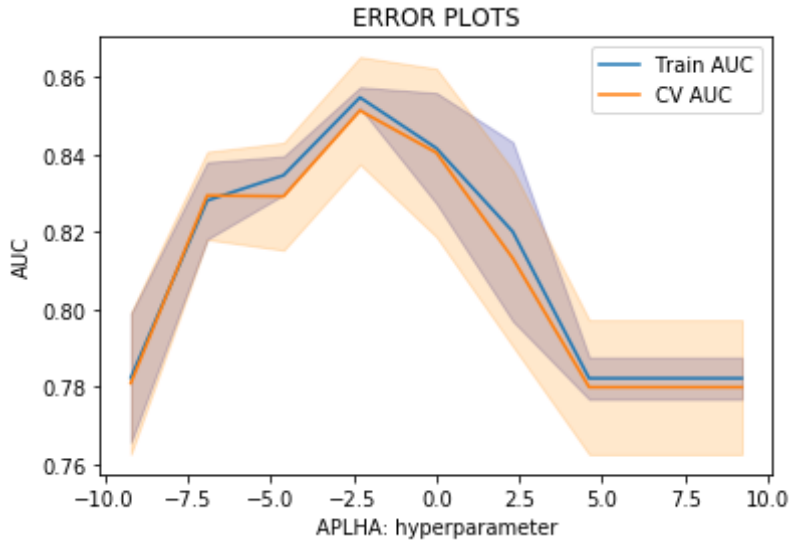
In [0]:
```
import matplotlib.pyplot as plt
import numpy as np

from sklearn.model_selection import validation_curve


param_range = np.logspace(-5, 1, 5)
```

```
train_scores, test_scores = validation_curve(
    SGDClassifier(),X_train_vec_standardized, Y_train, param_name="alph
a", param_range=param_range,
    cv=10, scoring="roc_auc", n_jobs=1)
train_scores_mean = np.mean(train_scores, axis=1)
train_scores_std = np.std(train_scores, axis=1)
test_scores_mean = np.mean(test_scores, axis=1)
test_scores_std = np.std(test_scores, axis=1)

plt.title("Validation Curve with Sgdclassifier")
plt.xlabel("$alpha$")
plt.ylabel("Score")
plt.ylim(0.0, 1.1)
lw = 2
plt.semilogx(param_range, train_scores_mean, label="Training score",
             color="darkorange", lw=lw)
plt.fill_between(param_range, train_scores_mean - train_scores_std,
                 train_scores_mean + train_scores_std, alpha=0.2,
                 color="darkorange", lw=lw)
plt.semilogx(param_range, test_scores_mean, label="Cross-validation sco
re",
             color="navy", lw=lw)
plt.fill_between(param_range, test_scores_mean - test_scores_std,
                 test_scores_mean + test_scores_std, alpha=0.2,
                 color="navy", lw=lw)
plt.legend(loc="best")
plt.show()
```



In [0]:
```
alpha = [10**-4, 10**-3, 10**-2,10**-1, 10**0, 10**1, 10**2, 10**3, 10**4]
train_auc= model.cv_results_['mean_train_score']
train_auc_std= model.cv_results_['std_train_score']
cv_auc = model.cv_results_['mean_test_score']
cv_auc_std= model.cv_results_['std_test_score']

plt.plot(np.log(alpha), train_auc, label='Train AUC')
# this code is copied from here: https://stackoverflow.com/a/48803361/4084039
plt.gca().fill_between(np.log(alpha),train_auc - train_auc_std,train_auc + train_auc_std,alpha=0.2,color='darkblue')

plt.plot(np.log(alpha), cv_auc, label='CV AUC')
# this code is copied from here: https://stackoverflow.com/a/48803361/4084039
plt.gca().fill_between(np.log(alpha),cv_auc - cv_auc_std,cv_auc + cv_auc_std,alpha=0.2,color='darkorange')
plt.legend()
plt.xlabel("APLHA: hyperparameter")
plt.ylabel("AUC")
```

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



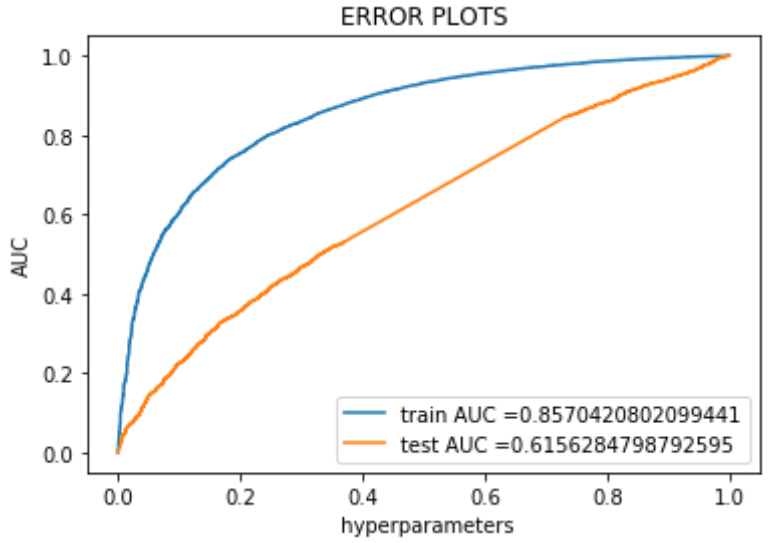
In [0]:
```
# SGDClassifier with Optimal value of alpha i.e.(1/C)
sgd = SGDClassifier(alpha=optimal_alpha, n_jobs=-1)
sgd.fit(X_train_vec_standardized,Y_train)
predictions = sgd.predict(X_test_vec_standardized)
predictions1 = sgd.predict(X_train_vec_standardized)

# Variables that will be used for  making table in Conclusion part of t
his assignment
tfidf_avg_w2v_grid_alpha = optimal_alpha
tfidf_avg_w2v_grid_train_acc = model.score(X_test_vec_standardized, Y_t
est)*100
tfidf_avg_w2v_grid_test_acc = accuracy_score(Y_test, predictions) * 100
```

In [0]:
```
calib = CalibratedClassifierCV(sgd, method = "sigmoid", cv = "prefit")
calib.fit(X_train_vec_standardized,Y_train)
```

Out[0]:
```
CalibratedClassifierCV(base_estimator=SGDClassifier(alpha=0.1, average=
False, class_weight=None, epsilon=0.1,
       eta0=0.0, fit_intercept=True, l1_ratio=0.15,
```

```
        learning_rate='optimal', loss='hinge', max_iter=None, n_iter=Non
e,
        n_jobs=-1, penalty='l2', power_t=0.5, random_state=None,
        shuffle=True, tol=None, verbose=0, warm_start=False),
             cv='prefit', method='sigmoid')
```

In [0]:
```
train_fpr, train_tpr, thresholds = roc_curve(Y_train, calib.predict_pro
ba(X_train_vec_standardized)[:,1])
test_fpr, test_tpr, thresholds = roc_curve(Y_test, calib.predict_proba(
X_test_vec_standardized)[:,1])

plt.plot(train_fpr, train_tpr, label="train AUC ="+str(auc(train_fpr, t
rain_tpr)))
plt.plot(test_fpr, test_tpr, label="test AUC ="+str(auc(test_fpr, test_
tpr)))
plt.legend()
plt.xlabel(" hyperparameters")
plt.ylabel("AUC")
plt.title("ERROR PLOTS")
plt.show()
```



In [0]:
```
# evaluate accuracy on test data
```

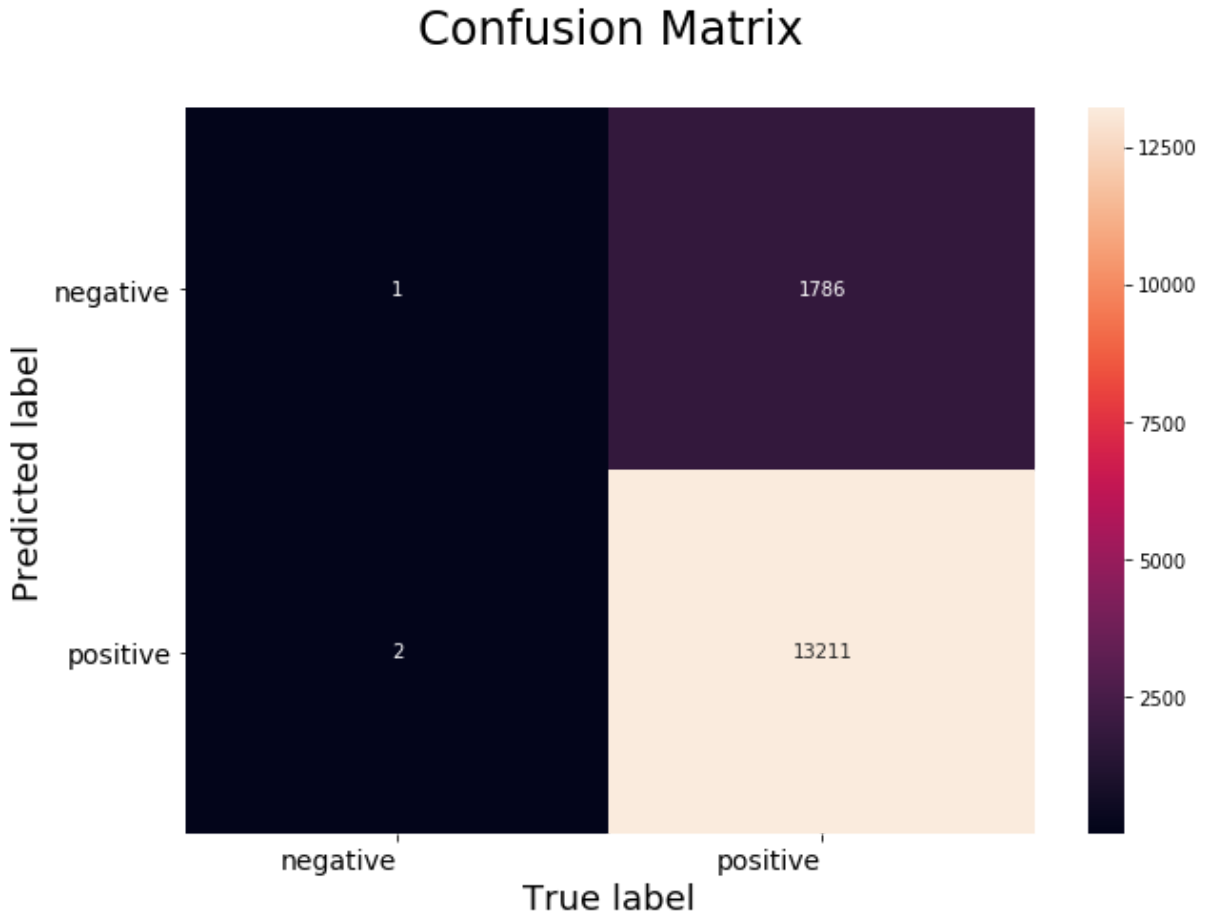
```
acc = accuracy_score(Y_test, predictions) * 100
print('\nThe Test Accuracy of the SGDClassifier for alpha = %.3f is %f%
%' % (optimal_alpha, acc))

# evaluate precision
acc = precision_score(Y_test, predictions, pos_label = 1)
print('\nThe Test Precision of the SGDClassifier for alpha = %.3f is %f
' % (optimal_alpha, acc))

# evaluate recall
acc = recall_score(Y_test, predictions, pos_label = 1)
print('\nThe Test Recall of the SGDClassifier for alpha = %.3f is %f' %
 (optimal_alpha, acc))

# evaluate f1-score
acc = f1_score(Y_test, predictions, pos_label = 1)
print('\nThe Test F1-Score of the SGDClassifier for alpha = %.3f is %f'
 % (optimal_alpha, acc))
```

```
The Test Accuracy of the SGDClassifier for alpha = 0.100 is 88.080000%

The Test Precision of the SGDClassifier for alpha = 0.100 is 0.880910

The Test Recall of the SGDClassifier for alpha = 0.100 is 0.999849

The Test F1-Score of the SGDClassifier for alpha = 0.100 is 0.936618
```

In [0]:

```
# Code for drawing seaborn heatmaps on test data
class_names = ['negative','positive']
df_heatmap = pd.DataFrame(confusion_matrix(Y_test, predictions), index=
class_names, columns=class_names )
fig = plt.figure(figsize=(10,7))
heatmap = sns.heatmap(df_heatmap, annot=True, fmt="d")

# Setting tick labels for heatmap
heatmap.yaxis.set_ticklabels(heatmap.yaxis.get_ticklabels(), rotation=0
, ha='right', fontsize=14)
heatmap.xaxis.set_ticklabels(heatmap.xaxis.get_ticklabels(), rotation=0
, ha='right', fontsize=14)
```

```
plt.ylabel('Predicted label',size=18)
plt.xlabel('True label',size=18)
plt.title("Confusion Matrix\n",size=24)
plt.show()
```



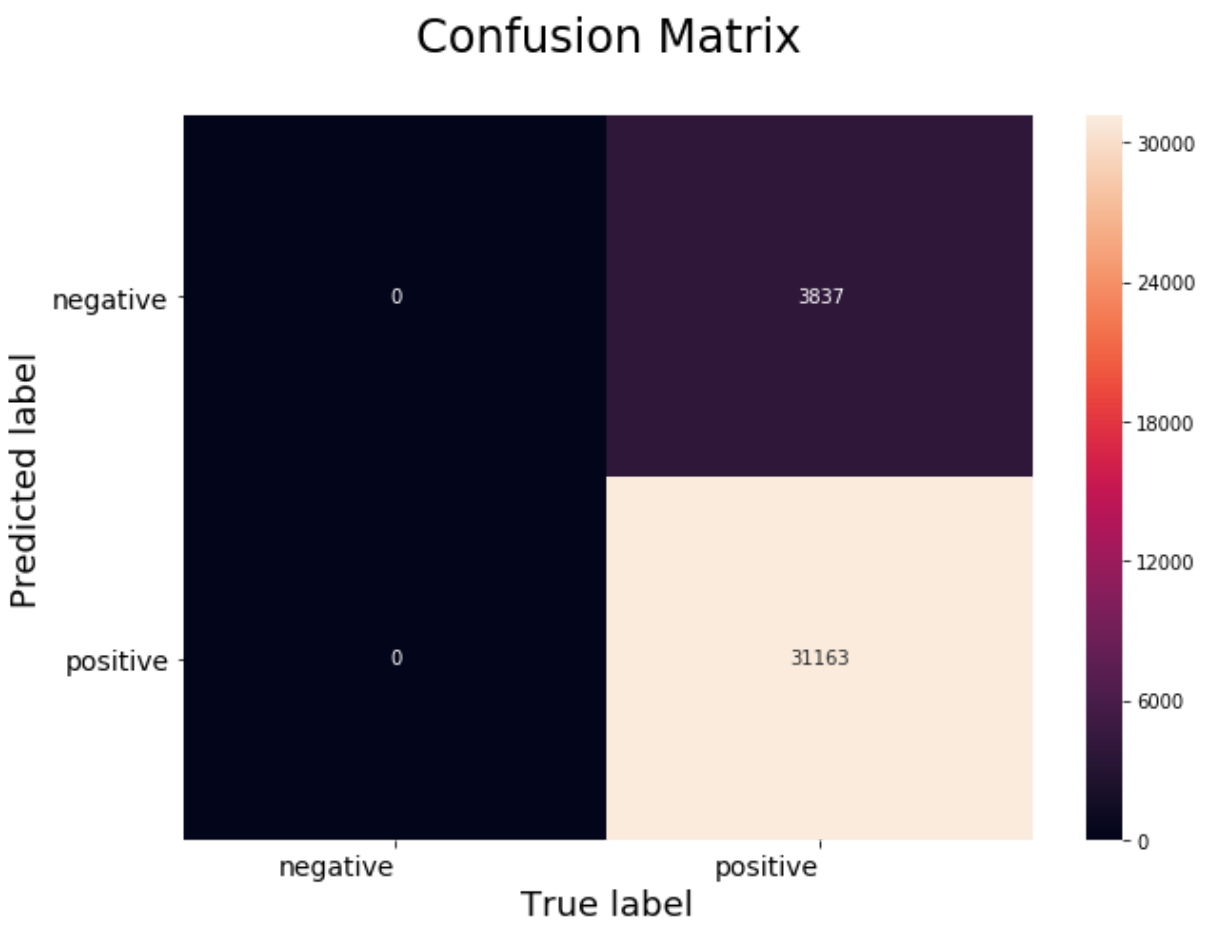
In [0]:
```
# evaluate accuracy on train data
acc = accuracy_score(Y_train, predictions1) * 100
print('\nThe Train Accuracy of the SGDClassifier for alpha = %.3f is %
f%%' % (optimal_alpha, acc))

# evaluate precision
```

```
acc = precision_score(Y_train, predictions1, pos_label = 1)
print('\nThe train Precision of the SGDClassifier for alpha = %.3f is %
f' % (optimal_alpha, acc))

# evaluate recall
acc = recall_score(Y_train, predictions1, pos_label = 1)
print('\nThe Train Recall of the SGDClassifier for alpha = %.3f is %f'
% (optimal_alpha, acc))

# evaluate f1-score
acc = f1_score(Y_train, predictions1, pos_label = 1)
print('\nThe Train F1-Score of the SGDClassifier for alpha = %.3f is %f
' % (optimal_alpha, acc))
```

```
The Train Accuracy of the SGDClassifier for alpha = 0.100 is 89.037143%

The train Precision of the SGDClassifier for alpha = 0.100 is 0.890371

The Train Recall of the SGDClassifier for alpha = 0.100 is 1.000000

The Train F1-Score of the SGDClassifier for alpha = 0.100 is 0.942007
```

In [0]:

```
# Code for drawing seaborn heatmaps on train data
class_names = ['negative','positive']
df_heatmap = pd.DataFrame(confusion_matrix(Y_train, predictions1), inde
x=class_names, columns=class_names )
fig = plt.figure(figsize=(10,7))
heatmap = sns.heatmap(df_heatmap, annot=True, fmt="d")

# Setting tick labels for heatmap
heatmap.yaxis.set_ticklabels(heatmap.yaxis.get_ticklabels(), rotation=0
, ha='right', fontsize=14)
heatmap.xaxis.set_ticklabels(heatmap.xaxis.get_ticklabels(), rotation=0
, ha='right', fontsize=14)
plt.ylabel('Predicted label',size=18)
plt.xlabel('True label',size=18)
plt.title("Confusion Matrix\n",size=24)
plt.show()
```



[5.2] RBF SVM

In [0]:
```
from sklearn.model_selection import train_test_split
##Sorting data according to Time in ascending order for Time Based Spli
tting
time_sorted_data = final.sort_values('Time', axis=0, ascending=True, in
place=False, kind='quicksort', na_position='last')

time_sorted_data = time_sorted_data[:20000]
```

```
x = time_sorted_data['CleanedText'].values
y = time_sorted_data['Score']

# split the data set into train and test
X_train, X_test, Y_train, Y_test = train_test_split(x, y, test_size=0.3
, random_state=0,shuffle=False)
```

In [8]:
```
X_train.shape
```

Out[8]:
```
(14000,)
```

In [9]:
```
# Please write all the code with proper documentation
#BoW
count_vect = CountVectorizer(min_df = 10,max_features = 5000)
X_train_vec = count_vect.fit_transform(X_train)
X_test_vec = count_vect.transform(X_test)
print("the type of count vectorizer :",type(X_train_vec))
print("the shape of out text BOW vectorizer : ",X_train_vec.get_shape
())
print("the number of unique words :", X_train_vec.get_shape()[1])

import warnings
warnings.filterwarnings('ignore')
# Data-preprocessing: Standardizing the data

from sklearn.preprocessing import StandardScaler
sc = StandardScaler(with_mean=False)
X_train_vec_standardized = sc.fit_transform(X_train_vec)
X_test_vec_standardized = sc.transform(X_test_vec)
```

```
the type of count vectorizer : <class 'scipy.sparse.csr.csr_matrix'>
the shape of out text BOW vectorizer :  (14000, 3478)
the number of unique words : 3478
```

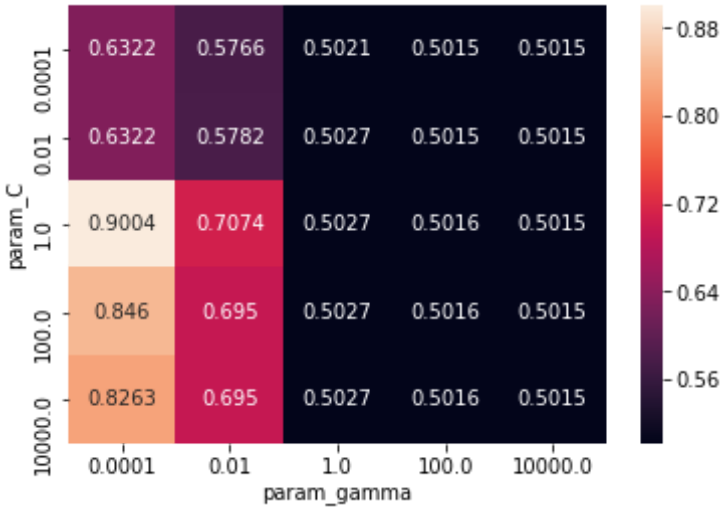
[5.2.1] Applying RBF SVM on BOW, SET 1

In [11]:
```
# Please write all the code with proper documentation
from sklearn.svm import SVC
from sklearn.linear_model import SGDClassifier
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import RandomizedSearchCV
from sklearn.metrics import accuracy_score,confusion_matrix,f1_score,pr
ecision_score,recall_score


param_grid = {'C':[10**-4, 10**-2, 10**0, 10**2, 10**4],'gamma':[10**-4
, 10**-2, 10**0, 10**2, 10**4]}
model = GridSearchCV(SVC(kernel = "rbf", class_weight = "balanced"), pa
ram_grid, scoring = 'roc_auc', cv=3 , n_jobs = -1,pre_dispatch=2)
model.fit(X_train_vec_standardized, Y_train)
print("Model with best parameters :\n",model.best_estimator_)
print("Accuracy of the model : ",model.score(X_test_vec_standardized, Y
_test))

print("\nOptimal C:", model.best_estimator_.C)
print("\nOptimal Gamma:", model.best_estimator_.gamma)
print("\nBest kernel:", model.best_estimator_.kernel)
print("\nBest Score:", model.best_score_)
```

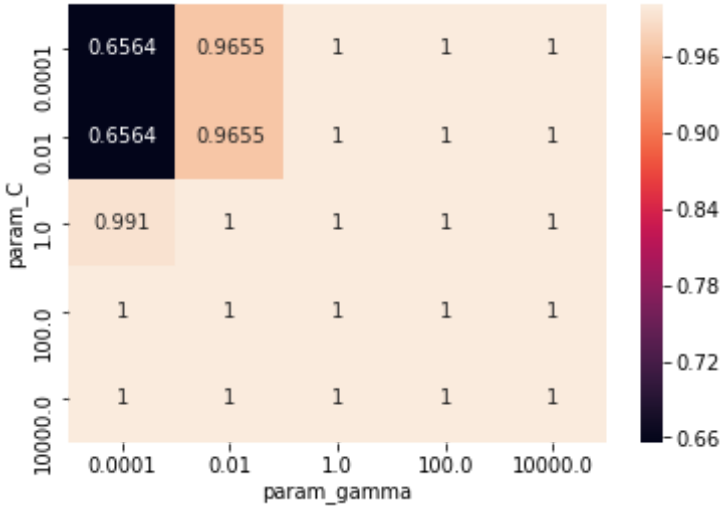
```
Model with best parameters :
 SVC(C=1, cache_size=200, class_weight='balanced', coef0=0.0,
  decision_function_shape='ovr', degree=3, gamma=0.0001, kernel='rbf',
  max_iter=-1, probability=False, random_state=None, shrinking=True,
  tol=0.001, verbose=False)
Accuracy of the model :  0.8922956883011604

Optimal C: 1

Optimal Gamma: 0.0001

Best kernel: rbf

Best Score: 0.9004093244022555
```

In [12]:
```
df_gridsearch = pd.DataFrame(model.cv_results_)
```

```
max_scores = df_gridsearch.groupby(['param_C','param_gamma']).max()
max_scores = max_scores.unstack()[['mean_test_score', 'mean_train_scor
e']]
sns.heatmap(max_scores.mean_test_score, annot=True, fmt='.4g')
plt.show()
```



In [29]:
```
df_gridsearch = pd.DataFrame(model.cv_results_)
max_scores = df_gridsearch.groupby(['param_C','param_gamma']).max()
max_scores = max_scores.unstack()[['mean_test_score', 'mean_train_scor
e']]
sns.heatmap(max_scores.mean_train_score, annot=True, fmt='.4g')
plt.show()
```



In [30]:
```
optimal_C = model.best_estimator_.C
print("The optimal value of C is : ",optimal_C)

optimal_gamma = model.best_estimator_.gamma
print("The optimal value of C is : ",optimal_gamma)
```

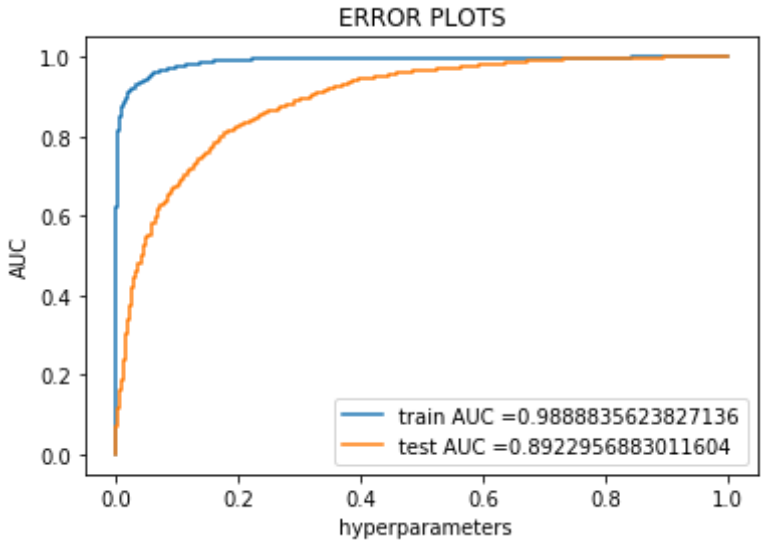
```
The optimal value of C is :  1
The optimal value of C is :  0.0001
```

In [0]:
```
# SVC with RBF kernel with Optimal value of C
svc = SVC(C=optimal_C,gamma=optimal_gamma,kernel = "rbf", class_weight
= "balanced", probability = True)
svc.fit(X_train_vec_standardized,Y_train)
predictions = svc.predict(X_test_vec_standardized)
predictions1 = svc.predict(X_train_vec_standardized)

bow_gridrbf_C = optimal_C
bow_gridrbf_gamma = optimal_gamma
bow_gridrbf_train_acc = model.score(X_test_vec_standardized, Y_test)*10
0
bow_gridrbf_test_acc = accuracy_score(Y_test, predictions) * 100
```

In [0]:
```
calib = CalibratedClassifierCV(svc, method = "sigmoid", cv = "prefit")
calib.fit(X_train_vec_standardized, Y_train)
```

Out[0]:
```
CalibratedClassifierCV(base_estimator=SVC(C=1, cache_size=200, class_we
ight='balanced', coef0=0.0,
  decision_function_shape='ovr', degree=3, gamma=0.0001, kernel='rbf',
  max_iter=-1, probability=True, random_state=None, shrinking=True,
  tol=0.001, verbose=False),
            cv='prefit', method='sigmoid')
```

In [0]:
```
train_fpr, train_tpr, thresholds = roc_curve(Y_train, calib.predict_pro
ba(X_train_vec_standardized)[:,1])
test_fpr, test_tpr, thresholds = roc_curve(Y_test, calib.predict_proba(
X_test_vec_standardized)[:,1])

plt.plot(train_fpr, train_tpr, label="train AUC ="+str(auc(train_fpr, t
rain_tpr)))
plt.plot(test_fpr, test_tpr, label="test AUC ="+str(auc(test_fpr, test_
tpr)))
plt.legend()
plt.xlabel(" hyperparameters")
plt.ylabel("AUC")
plt.title("ERROR PLOTS")
plt.show()
```



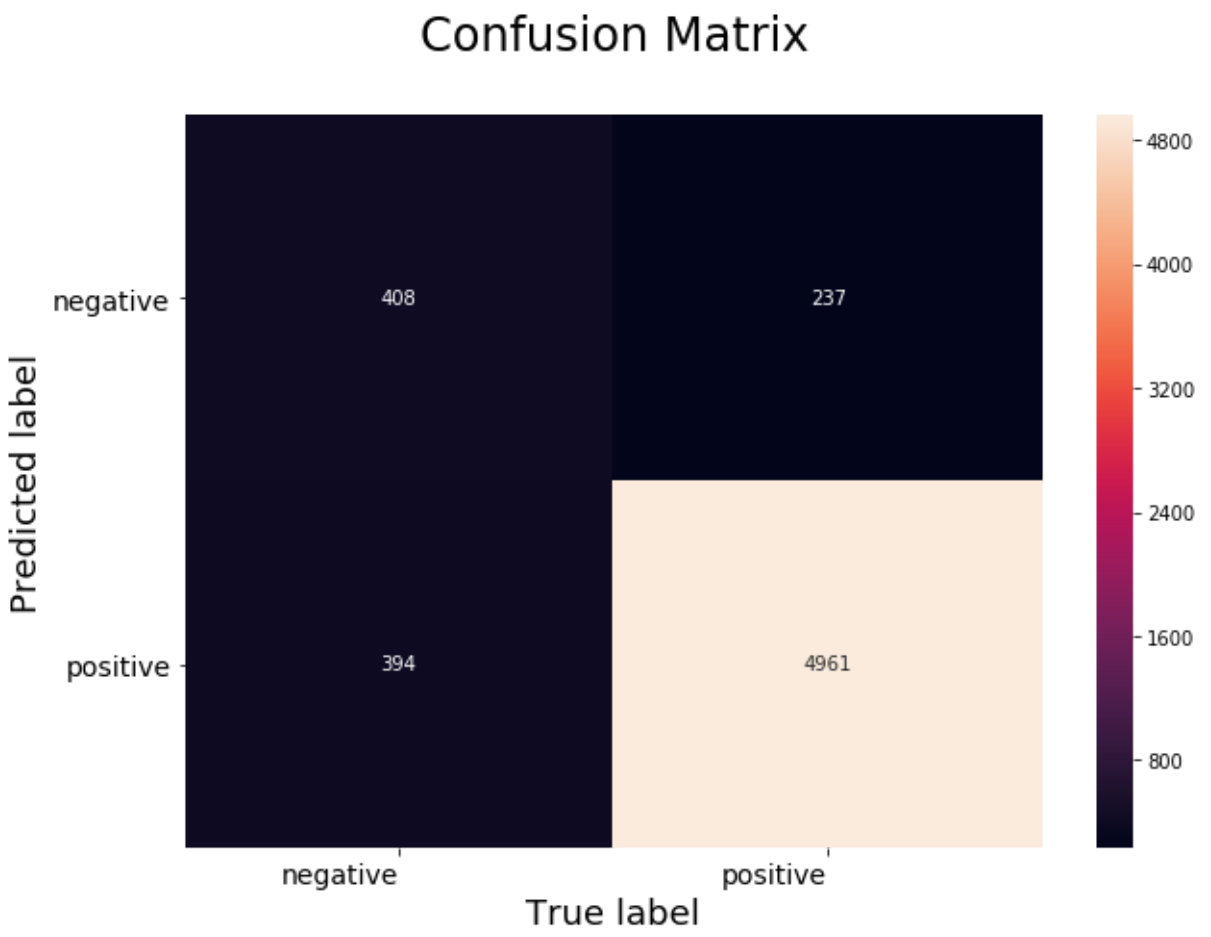
In [0]:
```
# evaluate accuracy on test data
acc = accuracy_score(Y_test, predictions) * 100
print('\nThe Test Accuracy of SVC with RBF kernel for C = %.3f is %f%%'
 % (optimal_C, acc))
print('\nThe Test Accuracy of SVC with RBF kernel for gamma = %.4f is %
f%%' % (optimal_gamma, acc))

# evaluate precision
acc = precision_score(Y_test, predictions, pos_label = 1)
print('\nThe Test Precision of SVC with RBF kernel for C = %.3f is %f'
% (optimal_C, acc))
print('\nThe Test Precision of SVC with RBF kernel for gammma = %.4f is
 %f' % (optimal_gamma, acc))

# evaluate recall
acc = recall_score(Y_test, predictions, pos_label = 1)
print('\nThe Test Recall of SVC with RBF kernel for C = %.3f is %f' % (
optimal_C, acc))
print('\nThe Test Recall of SVC with RBF kernel for gamma = %.4f is %f'
 % (optimal_gamma, acc))

# evaluate f1-score
```

```
acc = f1_score(Y_test, predictions, pos_label = 1)
print('\nThe Test F1-Score of SVC with RBF kernel for C = %.3f is %f' %
 (optimal_C, acc))
print('\nThe Test F1-Score of SVC with RBF kernel for gamma = %.4f is %
f' % (optimal_gamma, acc))
```

```
The Test Accuracy of SVC with RBF kernel for C = 1.000 is 89.483333%

The Test Accuracy of SVC with RBF kernel for gamma = 0.0001 is 89.48333
3%

The Test Precision of SVC with RBF kernel for C = 1.000 is 0.954406

The Test Precision of SVC with RBF kernel for gammma = 0.0001 is 0.9544
06

The Test Recall of SVC with RBF kernel for C = 1.000 is 0.926424

The Test Recall of SVC with RBF kernel for gamma = 0.0001 is 0.926424

The Test F1-Score of SVC with RBF kernel for C = 1.000 is 0.940207

The Test F1-Score of SVC with RBF kernel for gamma = 0.0001 is 0.940207
```

In [0]:
```
# Code for drawing seaborn heatmaps on test data
class_names = ['negative','positive']
df_heatmap = pd.DataFrame(confusion_matrix(Y_test, predictions), index=
class_names, columns=class_names )
fig = plt.figure(figsize=(10,7))
heatmap = sns.heatmap(df_heatmap, annot=True, fmt="d")

# Setting tick labels for heatmap
heatmap.yaxis.set_ticklabels(heatmap.yaxis.get_ticklabels(), rotation=0
, ha='right', fontsize=14)
heatmap.xaxis.set_ticklabels(heatmap.xaxis.get_ticklabels(), rotation=0
, ha='right', fontsize=14)
plt.ylabel('Predicted label',size=18)
plt.xlabel('True label',size=18)
```

```
plt.title("Confusion Matrix\n",size=24)
plt.show()
```



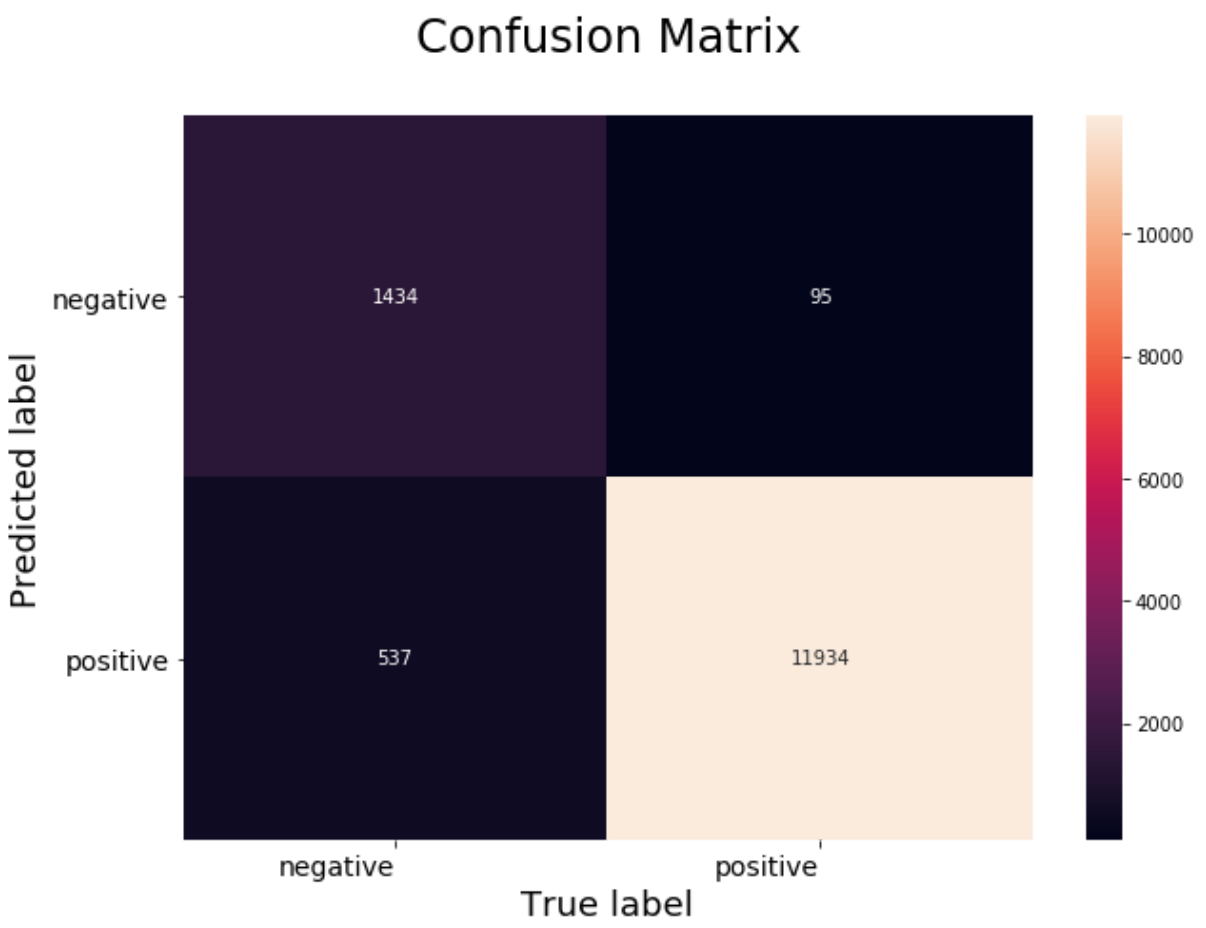
In [0]:
```
# evaluate accuracy on train data
acc = accuracy_score(Y_train, predictions1) * 100
print('\nThe Train accuracy of SVC with RBF kernel for C = %.3f is %f%%
' % (optimal_C, acc))
print('\nThe Train accuracy of SVC with RBF kernel for gamma = %.4f is
%f%%' % (optimal_gamma, acc))

# evaluate precision
```

```
acc = precision_score(Y_train, predictions1, pos_label = 1)
print('\nThe Train precision of SVC with RBF kernel for C = %.3f is %f'
 % (optimal_C, acc))
print('\nThe Train precision of SVC with RBF kernel for gamma = %.4f is
 %f' % (optimal_gamma, acc))

# evaluate recall
acc = recall_score(Y_train, predictions1, pos_label = 1)
print('\nThe Train Recall of SVC with RBF kernel for C = %.3f is %f' %
(optimal_C, acc))
print('\nThe Train Recall of SVC with RBF kernel for gamma = %.4f is %f
' % (optimal_gamma, acc))

# evaluate f1-score
acc = f1_score(Y_train, predictions1, pos_label = 1)
print('\nThe Train f1score of SVC with RBF kernel for C = %.3f is %f' %
 (optimal_C, acc))
print('\nThe Train f1score of SVC with RBF kernel for gamma = %.4f is %
f' % (optimal_gamma, acc))
```

```
The Train accuracy of SVC with RBF kernel for C = 1.000 is 95.485714%

The Train accuracy of SVC with RBF kernel for gamma = 0.0001 is 95.4857
14%

The Train precision of SVC with RBF kernel for C = 1.000 is 0.992102

The Train precision of SVC with RBF kernel for gamma = 0.0001 is 0.9921
02

The Train Recall of SVC with RBF kernel for C = 1.000 is 0.956940

The Train Recall of SVC with RBF kernel for gamma = 0.0001 is 0.956940

The Train f1score of SVC with RBF kernel for C = 1.000 is 0.974204

The Train f1score of SVC with RBF kernel for gamma = 0.0001 is 0.974204
```

In [0]:
```
# Code for drawing seaborn heatmaps on train data
```

```
class_names = ['negative','positive']
df_heatmap = pd.DataFrame(confusion_matrix(Y_train, predictions1), inde
x=class_names, columns=class_names )
fig = plt.figure(figsize=(10,7))
heatmap = sns.heatmap(df_heatmap, annot=True, fmt="d")

# Setting tick labels for heatmap
heatmap.yaxis.set_ticklabels(heatmap.yaxis.get_ticklabels(), rotation=0
, ha='right', fontsize=14)
heatmap.xaxis.set_ticklabels(heatmap.xaxis.get_ticklabels(), rotation=0
, ha='right', fontsize=14)
plt.ylabel('Predicted label',size=18)
plt.xlabel('True label',size=18)
plt.title("Confusion Matrix\n",size=24)
plt.show()
```



[5.2.2] Applying RBF SVM on TFIDF, SET 2

In [32]:
```
# Please write all the code with proper documentation
tf_idf_vect = TfidfVectorizer(min_df=10,max_features=5000)
X_train_vec = tf_idf_vect.fit_transform(X_train)
X_test_vec = tf_idf_vect.transform(X_test)
print("the type of count vectorizer :",type(X_train_vec))
print("the shape of out text TFIDF vectorizer : ",X_train_vec.get_shape
())
```

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

# Data-preprocessing: Standardizing the data
sc = StandardScaler(with_mean=False)
X_train_vec_standardized = sc.fit_transform(X_train_vec)
X_test_vec_standardized = sc.transform(X_test_vec)
```

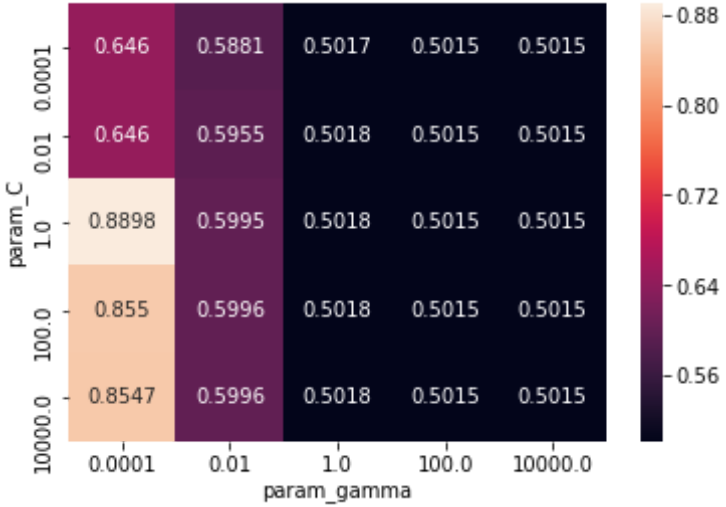
```
the type of count vectorizer : <class 'scipy.sparse.csr.csr_matrix'>
the shape of out text TFIDF vectorizer :  (14000, 3478)
the number of unique words : 3478
```

In [33]:

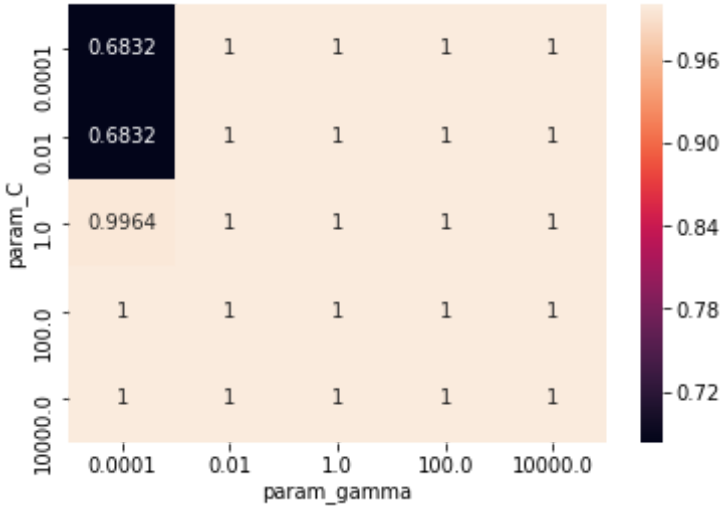
```
# Please write all the code with proper documentation
param_grid = {'C':[10**-4, 10**-2, 10**0, 10**2, 10**4],'gamma':[10**-4
, 10**-2, 10**0, 10**2, 10**4]}
model = GridSearchCV(SVC(kernel = "rbf", class_weight = "balanced"), pa
ram_grid, scoring = 'roc_auc', cv=3 , n_jobs = -1,pre_dispatch=2)
model.fit(X_train_vec_standardized, Y_train)
print("Model with best parameters :\n",model.best_estimator_)
print("Accuracy of the model : ",model.score(X_test_vec_standardized, Y
_test))

print("\nOptimal C:", model.best_estimator_.C)
print("\nOptimal Gamma:", model.best_estimator_.gamma)
print("\nBest kernel:", model.best_estimator_.kernel)
print("\nBest Score:", model.best_score_)
df_gridsearch = pd.DataFrame(model.cv_results_)
max_scores = df_gridsearch.groupby(['param_C','param_gamma']).max()
max_scores = max_scores.unstack()[['mean_test_score', 'mean_train_scor
e']]
sns.heatmap(max_scores.mean_test_score, annot=True, fmt='.4g')
plt.show()
```

```
Model with best parameters :
 SVC(C=1, cache_size=200, class_weight='balanced', coef0=0.0,
  decision_function_shape='ovr', degree=3, gamma=0.0001, kernel='rbf',
  max_iter=-1, probability=False, random_state=None, shrinking=True,
  tol=0.001, verbose=False)
Accuracy of the model :  0.8865886406242083

Optimal C: 1
```

```
Optimal Gamma: 0.0001

Best kernel: rbf

Best Score: 0.8898060877722092
```



In [34]:
```
df_gridsearch = pd.DataFrame(model.cv_results_)
max_scores = df_gridsearch.groupby(['param_C','param_gamma']).max()
max_scores = max_scores.unstack()[['mean_test_score', 'mean_train_scor
e']]
sns.heatmap(max_scores.mean_train_score, annot=True, fmt='.4g')
plt.show()
```



In [0]:
```
optimal_C = model.best_estimator_.C
print("The optimal value of C is : ",optimal_C)

optimal_gamma = model.best_estimator_.gamma
print("The optimal value of C is : ",optimal_gamma)
```

```
The optimal value of C is :  1
The optimal value of C is :  0.0001
```

In [0]:
```
svc = SVC(C=optimal_C,gamma=optimal_gamma,kernel = "rbf", class_weight
= "balanced", probability = True)
svc.fit(X_train_vec_standardized,Y_train)
predictions = svc.predict(X_test_vec_standardized)
predictions1 = svc.predict(X_train_vec_standardized)

tfidf_gridrbf_C = optimal_C
tfidf_gridrbf_gamma = optimal_gamma
tfidf_gridrbf_train_acc = model.score(X_test_vec_standardized, Y_test)*
100
tfidf_gridrbf_test_acc = accuracy_score(Y_test, predictions) * 100
```

In [0]:
```
calib = CalibratedClassifierCV(svc, method = "sigmoid", cv = "prefit")
```

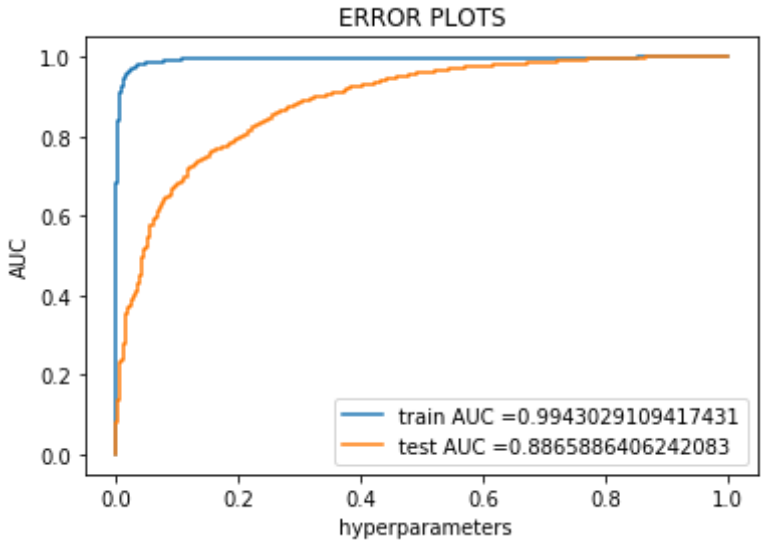
```
calib.fit(X_train_vec_standardized, Y_train)
```

Out[0]:

```
CalibratedClassifierCV(base_estimator=SVC(C=1, cache_size=200, class_we
ight='balanced', coef0=0.0,
  decision_function_shape='ovr', degree=3, gamma=0.0001, kernel='rbf',
  max_iter=-1, probability=True, random_state=None, shrinking=True,
  tol=0.001, verbose=False),
            cv='prefit', method='sigmoid')
```

In [0]:

```
train_fpr, train_tpr, thresholds = roc_curve(Y_train, calib.predict_pro
ba(X_train_vec_standardized)[:,1])
test_fpr, test_tpr, thresholds = roc_curve(Y_test, calib.predict_proba(
X_test_vec_standardized)[:,1])

plt.plot(train_fpr, train_tpr, label="train AUC ="+str(auc(train_fpr, t
rain_tpr)))
plt.plot(test_fpr, test_tpr, label="test AUC ="+str(auc(test_fpr, test_
tpr)))
plt.legend()
plt.xlabel(" hyperparameters")
plt.ylabel("AUC")
plt.title("ERROR PLOTS")
plt.show()
```



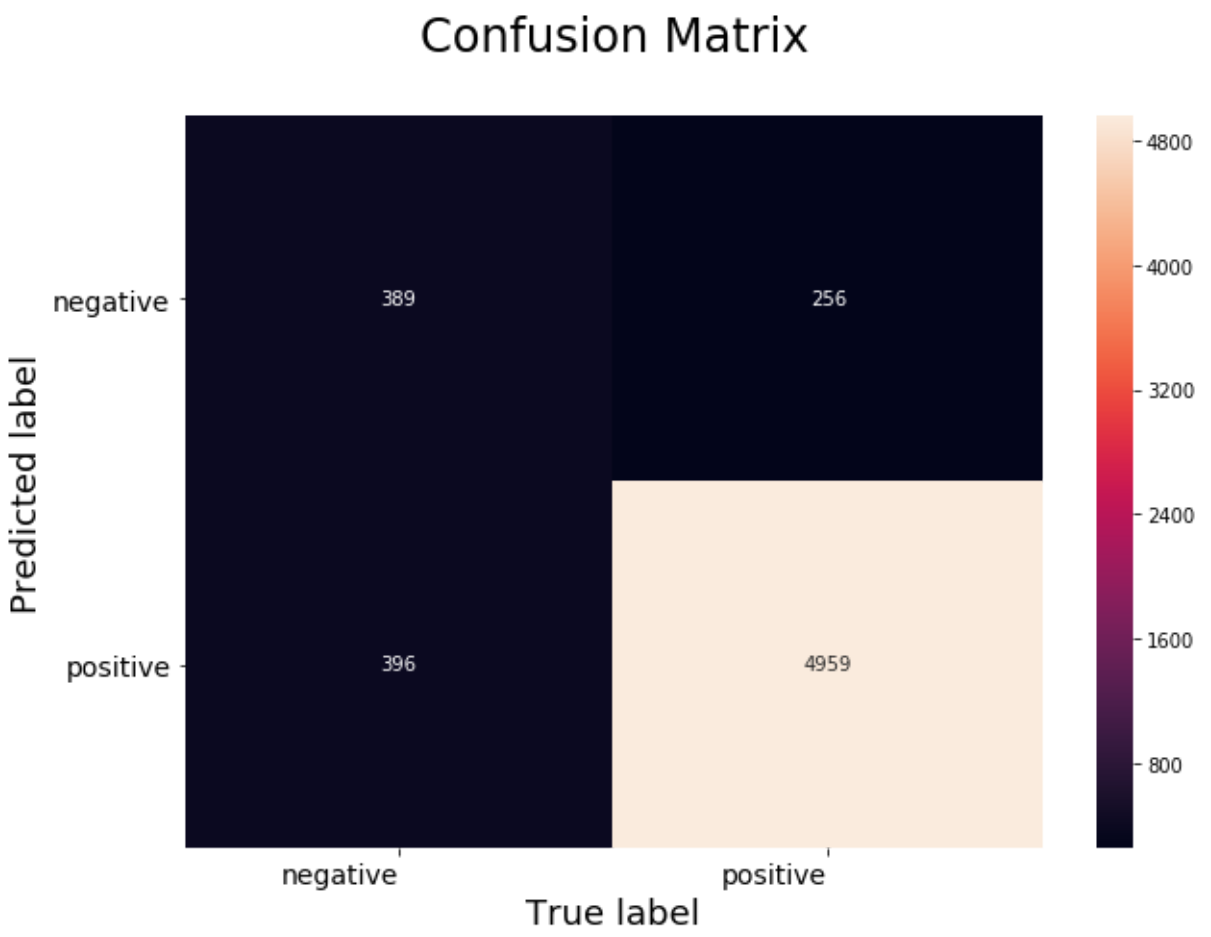
In [0]:
```
# evaluate accuracy on test data
acc = accuracy_score(Y_test, predictions) * 100
print('\nThe Test Accuracy of SVC with RBF kernel for C = %.3f is %f%%'
 % (optimal_C, acc))
print('\nThe Test Accuracy of SVC with RBF kernel for gamma = %.4f is %
f%%' % (optimal_gamma, acc))

# evaluate precision
acc = precision_score(Y_test, predictions, pos_label = 1)
print('\nThe Test Precision of SVC with RBF kernel for C = %.3f is %f'
% (optimal_C, acc))
print('\nThe Test Precision of SVC with RBF kernel for gammma = %.4f is
 %f' % (optimal_gamma, acc))

# evaluate recall
acc = recall_score(Y_test, predictions, pos_label = 1)
print('\nThe Test Recall of SVC with RBF kernel for C = %.3f is %f' % (
optimal_C, acc))
print('\nThe Test Recall of SVC with RBF kernel for gamma = %.4f is %f'
 % (optimal_gamma, acc))

# evaluate f1-score
```

```
acc = f1_score(Y_test, predictions, pos_label = 1)
print('\nThe Test F1-Score of SVC with RBF kernel for C = %.3f is %f' %
 (optimal_C, acc))
print('\nThe Test F1-Score of SVC with RBF kernel for gamma = %.4f is %
f' % (optimal_gamma, acc))
```

```
The Test Accuracy of SVC with RBF kernel for C = 1.000 is 89.133333%

The Test Accuracy of SVC with RBF kernel for gamma = 0.0001 is 89.13333
3%

The Test Precision of SVC with RBF kernel for C = 1.000 is 0.950911

The Test Precision of SVC with RBF kernel for gammma = 0.0001 is 0.9509
11

The Test Recall of SVC with RBF kernel for C = 1.000 is 0.926050

The Test Recall of SVC with RBF kernel for gamma = 0.0001 is 0.926050

The Test F1-Score of SVC with RBF kernel for C = 1.000 is 0.938316

The Test F1-Score of SVC with RBF kernel for gamma = 0.0001 is 0.938316
```

In [0]:
```
# Code for drawing seaborn heatmaps on test data
class_names = ['negative','positive']
df_heatmap = pd.DataFrame(confusion_matrix(Y_test, predictions), index=
class_names, columns=class_names )
fig = plt.figure(figsize=(10,7))
heatmap = sns.heatmap(df_heatmap, annot=True, fmt="d")

# Setting tick labels for heatmap
heatmap.yaxis.set_ticklabels(heatmap.yaxis.get_ticklabels(), rotation=0
, ha='right', fontsize=14)
heatmap.xaxis.set_ticklabels(heatmap.xaxis.get_ticklabels(), rotation=0
, ha='right', fontsize=14)
plt.ylabel('Predicted label',size=18)
plt.xlabel('True label',size=18)
```

```
plt.title("Confusion Matrix\n",size=24)
plt.show()
```



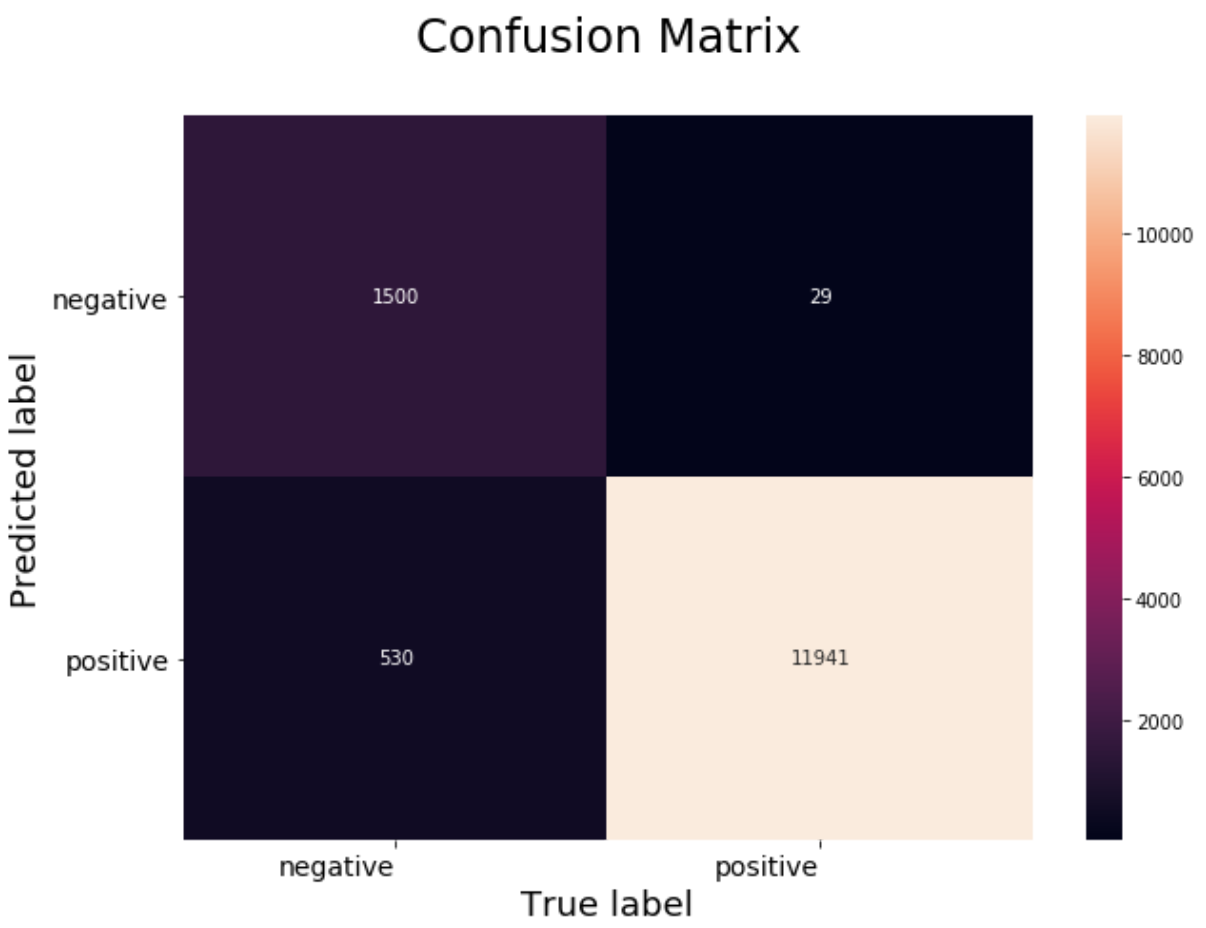
In [0]:
```
# evaluate accuracy on train data
acc = accuracy_score(Y_train, predictions1) * 100
print('\nThe Train accuracy of SVC with RBF kernel for C = %.3f is %f%%
' % (optimal_C, acc))
print('\nThe Train accuracy of SVC with RBF kernel for gamma = %.4f is
%f%%' % (optimal_gamma, acc))

# evaluate precision
```

```
acc = precision_score(Y_train, predictions1, pos_label = 1)
print('\nThe Train precision of SVC with RBF kernel for C = %.3f is %f'
 % (optimal_C, acc))
print('\nThe Train precision of SVC with RBF kernel for gamma = %.4f is
 %f' % (optimal_gamma, acc))

# evaluate recall
acc = recall_score(Y_train, predictions1, pos_label = 1)
print('\nThe Train Recall of SVC with RBF kernel for C = %.3f is %f' %
(optimal_C, acc))
print('\nThe Train Recall of SVC with RBF kernel for gamma = %.4f is %f
' % (optimal_gamma, acc))

# evaluate f1-score
acc = f1_score(Y_train, predictions1, pos_label = 1)
print('\nThe Train f1score of SVC with RBF kernel for C = %.3f is %f' %
 (optimal_C, acc))
print('\nThe Train f1score of SVC with RBF kernel for gamma = %.4f is %
f' % (optimal_gamma, acc))
```

```
The Train accuracy of SVC with RBF kernel for C = 1.000 is 96.007143%

The Train accuracy of SVC with RBF kernel for gamma = 0.0001 is 96.0071
43%

The Train precision of SVC with RBF kernel for C = 1.000 is 0.997577

The Train precision of SVC with RBF kernel for gamma = 0.0001 is 0.9975
77

The Train Recall of SVC with RBF kernel for C = 1.000 is 0.957501

The Train Recall of SVC with RBF kernel for gamma = 0.0001 is 0.957501

The Train f1score of SVC with RBF kernel for C = 1.000 is 0.977129

The Train f1score of SVC with RBF kernel for gamma = 0.0001 is 0.977129
```

In [0]:
```
# Code for drawing seaborn heatmaps on train data
```

```
class_names = ['negative','positive']
df_heatmap = pd.DataFrame(confusion_matrix(Y_train, predictions1), inde
x=class_names, columns=class_names )
fig = plt.figure(figsize=(10,7))
heatmap = sns.heatmap(df_heatmap, annot=True, fmt="d")

# Setting tick labels for heatmap
heatmap.yaxis.set_ticklabels(heatmap.yaxis.get_ticklabels(), rotation=0
, ha='right', fontsize=14)
heatmap.xaxis.set_ticklabels(heatmap.xaxis.get_ticklabels(), rotation=0
, ha='right', fontsize=14)
plt.ylabel('Predicted label',size=18)
plt.xlabel('True label',size=18)
plt.title("Confusion Matrix\n",size=24)
plt.show()
```



[5.2.3] Applying RBF SVM on AVG W2V, SET 3

In [35]:
```
# Please write all the code with proper documentation
# List of sentence in X_train text
sent_of_train=[]
for sent in X_train:
    sent_of_train.append(sent.split())

# List of sentence in X_est text
```

```
sent_of_test=[]
for sent in X_test:
    sent_of_test.append(sent.split())

# Train your own Word2Vec model using your own train text corpus
# min_count = 5 considers only words that occured atleast 5 times
w2v_model=Word2Vec(sent_of_train,min_count=5,size=50, workers=4)

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

```
number of words that occured minimum 5 times  5468
```

In [0]:

```
# compute average word2vec for each review for X_train .
train_vectors = [];
for sent in sent_of_train:
    sent_vec = np.zeros(50)
    cnt_words =0;
    for word in sent: #
        if word in w2v_words:
            vec = w2v_model.wv[word]
            sent_vec += vec
            cnt_words += 1
    if cnt_words != 0:
        sent_vec /= cnt_words
    train_vectors.append(sent_vec)

# compute average word2vec for each review for X_test .
test_vectors = [];
for sent in sent_of_test:
    sent_vec = np.zeros(50)
    cnt_words =0;
    for word in sent: #
        if word in w2v_words:
            vec = w2v_model.wv[word]
            sent_vec += vec
            cnt_words += 1
    if cnt_words != 0:
        sent_vec /= cnt_words
```

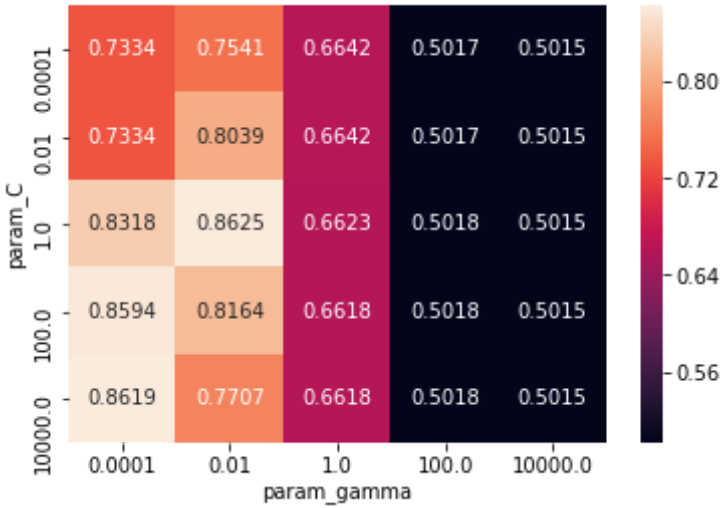
```
    test_vectors.append(sent_vec)

# Data-preprocessing: Standardizing the data
sc = StandardScaler()
X_train_vec_standardized = sc.fit_transform(train_vectors)
X_test_vec_standardized = sc.transform(test_vectors)
```

In [37]:

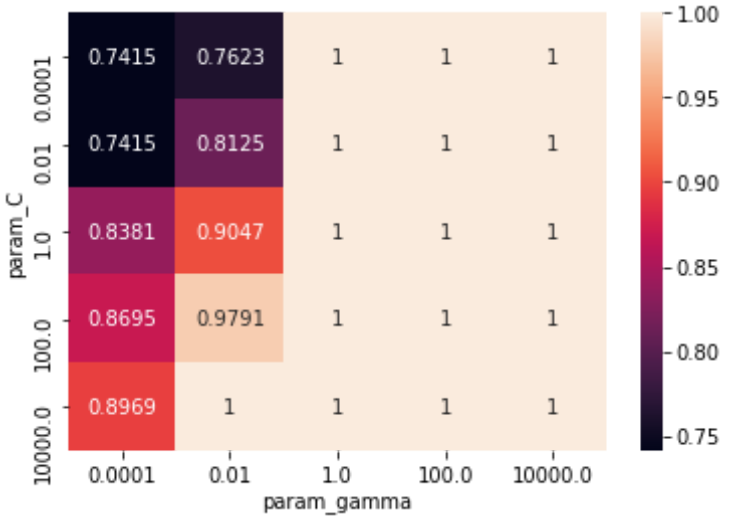
```
# Please write all the code with proper documentation
param_grid = {'C':[10**-4, 10**-2, 10**0, 10**2, 10**4],'gamma':[10**-4
, 10**-2, 10**0, 10**2, 10**4]}
model = GridSearchCV(SVC(kernel = "rbf", class_weight = "balanced"), pa
ram_grid, scoring = 'roc_auc', cv=3 , n_jobs = -1,pre_dispatch=2)
model.fit(X_train_vec_standardized, Y_train)
print("Model with best parameters :\n",model.best_estimator_)
print("Accuracy of the model : ",model.score(X_test_vec_standardized, Y
_test))

print("\nOptimal C:", model.best_estimator_.C)
print("\nOptimal Gamma:", model.best_estimator_.gamma)
print("\nBest kernel:", model.best_estimator_.kernel)
print("\nBest Score:", model.best_score_)
df_gridsearch = pd.DataFrame(model.cv_results_)
max_scores = df_gridsearch.groupby(['param_C','param_gamma']).max()
max_scores = max_scores.unstack()[['mean_test_score', 'mean_train_scor
e']]
sns.heatmap(max_scores.mean_test_score, annot=True, fmt='.4g')
plt.show()
```

```
Model with best parameters :
 SVC(C=1, cache_size=200, class_weight='balanced', coef0=0.0,
  decision_function_shape='ovr', degree=3, gamma=0.01, kernel='rbf',
  max_iter=-1, probability=False, random_state=None, shrinking=True,
  tol=0.001, verbose=False)
Accuracy of the model :  0.8563020867261634

Optimal C: 1

Optimal Gamma: 0.01
```

```
Best kernel: rbf

Best Score: 0.8625034027238583
```



In [38]:
```
df_gridsearch = pd.DataFrame(model.cv_results_)
max_scores = df_gridsearch.groupby(['param_C','param_gamma']).max()
max_scores = max_scores.unstack()[['mean_test_score', 'mean_train_scor
e']]
sns.heatmap(max_scores.mean_train_score, annot=True, fmt='.4g')
plt.show()
```



In [0]:
```
optimal_C = model.best_estimator_.C
print("The optimal value of C is : ",optimal_C)

optimal_gamma = model.best_estimator_.gamma
print("The optimal value of C is : ",optimal_gamma)
```

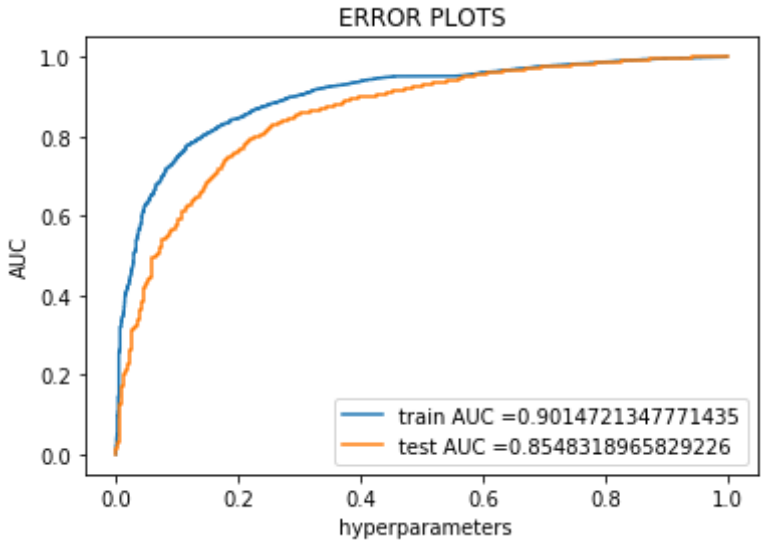
```
The optimal value of C is :  1
The optimal value of C is :  0.01
```

In [0]:
```
# SVC with RBF kernel with Optimal value of C
svc = SVC(C=optimal_C,gamma=optimal_gamma,kernel = "rbf", class_weight
= "balanced", probability = True)
svc.fit(X_train_vec_standardized,Y_train)
predictions = svc.predict(X_test_vec_standardized)
predictions1 = svc.predict(X_train_vec_standardized)

avg_w2v_gridrbf_C = optimal_C
avg_w2v_gridrbf_gamma = optimal_gamma
avg_w2v_gridrbf_train_acc = model.score(X_test_vec_standardized, Y_test
)*100
avg_w2v_gridrbf_test_acc = accuracy_score(Y_test, predictions) * 100
```

In [0]:
```
calib = CalibratedClassifierCV(svc, method = "sigmoid", cv = "prefit")
calib.fit(X_train_vec_standardized, Y_train)
```

Out[0]:
```
CalibratedClassifierCV(base_estimator=SVC(C=1, cache_size=200, class_we
ight='balanced', coef0=0.0,
  decision_function_shape='ovr', degree=3, gamma=0.01, kernel='rbf',
  max_iter=-1, probability=True, random_state=None, shrinking=True,
  tol=0.001, verbose=False),
            cv='prefit', method='sigmoid')
```

In [0]:
```
train_fpr, train_tpr, thresholds = roc_curve(Y_train, calib.predict_pro
ba(X_train_vec_standardized)[:,1])
test_fpr, test_tpr, thresholds = roc_curve(Y_test, calib.predict_proba(
X_test_vec_standardized)[:,1])

plt.plot(train_fpr, train_tpr, label="train AUC ="+str(auc(train_fpr, t
rain_tpr)))
plt.plot(test_fpr, test_tpr, label="test AUC ="+str(auc(test_fpr, test_
tpr)))
plt.legend()
plt.xlabel(" hyperparameters")
plt.ylabel("AUC")
plt.title("ERROR PLOTS")
plt.show()
```



In [0]:
```
# evaluate accuracy on test data
acc = accuracy_score(Y_test, predictions) * 100
print('\nThe Test Accuracy of SVC with RBF kernel for C = %.3f is %f%%'
 % (optimal_C, acc))
print('\nThe Test Accuracy of SVC with RBF kernel for gamma = %.4f is %
f%%' % (optimal_gamma, acc))

# evaluate precision
acc = precision_score(Y_test, predictions, pos_label = 1)
print('\nThe Test Precision of SVC with RBF kernel for C = %.3f is %f'
% (optimal_C, acc))
print('\nThe Test Precision of SVC with RBF kernel for gammma = %.4f is
 %f' % (optimal_gamma, acc))

# evaluate recall
acc = recall_score(Y_test, predictions, pos_label = 1)
print('\nThe Test Recall of SVC with RBF kernel for C = %.3f is %f' % (
optimal_C, acc))
print('\nThe Test Recall of SVC with RBF kernel for gamma = %.4f is %f'
 % (optimal_gamma, acc))

# evaluate f1-score
```

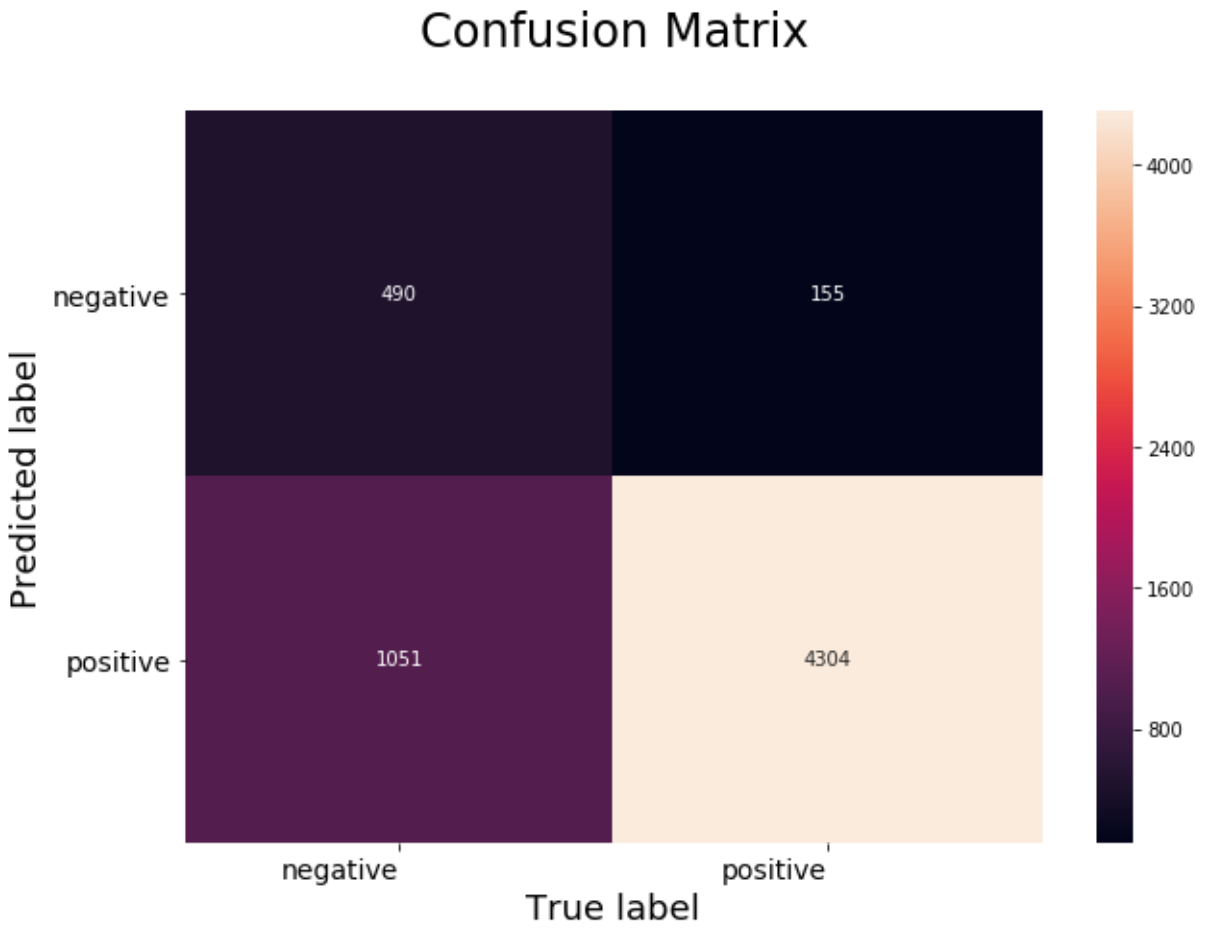
```
acc = f1_score(Y_test, predictions, pos_label = 1)
print('\nThe Test F1-Score of SVC with RBF kernel for C = %.3f is %f' %
 (optimal_C, acc))
print('\nThe Test F1-Score of SVC with RBF kernel for gamma = %.4f is %
f' % (optimal_gamma, acc))
```

```
The Test Accuracy of SVC with RBF kernel for C = 1.000 is 79.900000%

The Test Accuracy of SVC with RBF kernel for gamma = 0.0100 is 79.90000
0%

The Test Precision of SVC with RBF kernel for C = 1.000 is 0.965239

The Test Precision of SVC with RBF kernel for gammma = 0.0100 is 0.9652
39

The Test Recall of SVC with RBF kernel for C = 1.000 is 0.803735

The Test Recall of SVC with RBF kernel for gamma = 0.0100 is 0.803735

The Test F1-Score of SVC with RBF kernel for C = 1.000 is 0.877114

The Test F1-Score of SVC with RBF kernel for gamma = 0.0100 is 0.877114
```

In [0]:

```
# Code for drawing seaborn heatmaps on test data
class_names = ['negative','positive']
df_heatmap = pd.DataFrame(confusion_matrix(Y_test, predictions), index=
class_names, columns=class_names )
fig = plt.figure(figsize=(10,7))
heatmap = sns.heatmap(df_heatmap, annot=True, fmt="d")

# Setting tick labels for heatmap
heatmap.yaxis.set_ticklabels(heatmap.yaxis.get_ticklabels(), rotation=0
, ha='right', fontsize=14)
heatmap.xaxis.set_ticklabels(heatmap.xaxis.get_ticklabels(), rotation=0
, ha='right', fontsize=14)
plt.ylabel('Predicted label',size=18)
plt.xlabel('True label',size=18)
```

```
plt.title("Confusion Matrix\n",size=24)
plt.show()
```



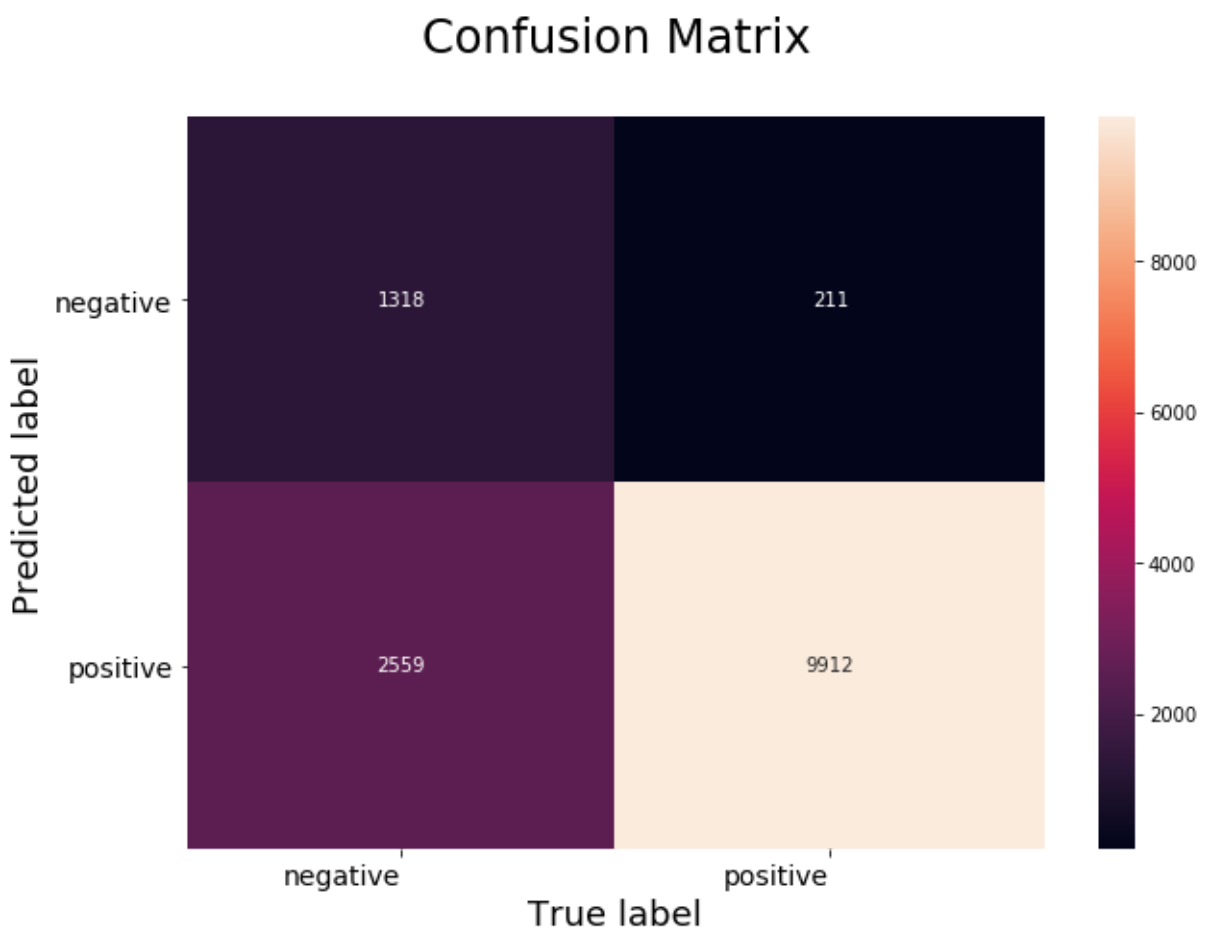
In [0]:
```
# evaluate accuracy on train data
acc = accuracy_score(Y_train, predictions1) * 100
print('\nThe Train accuracy of SVC with RBF kernel for C = %.3f is %f%%
' % (optimal_C, acc))
print('\nThe Train accuracy of SVC with RBF kernel for gamma = %.4f is
%f%%' % (optimal_gamma, acc))

# evaluate precision
```

```
acc = precision_score(Y_train, predictions1, pos_label = 1)
print('\nThe Train precision of SVC with RBF kernel for C = %.3f is %f'
 % (optimal_C, acc))
print('\nThe Train precision of SVC with RBF kernel for gamma = %.4f is
 %f' % (optimal_gamma, acc))

# evaluate recall
acc = recall_score(Y_train, predictions1, pos_label = 1)
print('\nThe Train Recall of SVC with RBF kernel for C = %.3f is %f' %
(optimal_C, acc))
print('\nThe Train Recall of SVC with RBF kernel for gamma = %.4f is %f
' % (optimal_gamma, acc))

# evaluate f1-score
acc = f1_score(Y_train, predictions1, pos_label = 1)
print('\nThe Train f1score of SVC with RBF kernel for C = %.3f is %f' %
 (optimal_C, acc))
print('\nThe Train f1score of SVC with RBF kernel for gamma = %.4f is %
f' % (optimal_gamma, acc))
```

```
The Train accuracy of SVC with RBF kernel for C = 1.000 is 80.214286%

The Train accuracy of SVC with RBF kernel for gamma = 0.0100 is 80.2142
86%

The Train precision of SVC with RBF kernel for C = 1.000 is 0.979156

The Train precision of SVC with RBF kernel for gamma = 0.0100 is 0.9791
56

The Train Recall of SVC with RBF kernel for C = 1.000 is 0.794804

The Train Recall of SVC with RBF kernel for gamma = 0.0100 is 0.794804

The Train f1score of SVC with RBF kernel for C = 1.000 is 0.877401

The Train f1score of SVC with RBF kernel for gamma = 0.0100 is 0.877401
```

In [0]:
```
# Code for drawing seaborn heatmaps on train data
```

```
class_names = ['negative','positive']
df_heatmap = pd.DataFrame(confusion_matrix(Y_train, predictions1), inde
x=class_names, columns=class_names )
fig = plt.figure(figsize=(10,7))
heatmap = sns.heatmap(df_heatmap, annot=True, fmt="d")

# Setting tick labels for heatmap
heatmap.yaxis.set_ticklabels(heatmap.yaxis.get_ticklabels(), rotation=0
, ha='right', fontsize=14)
heatmap.xaxis.set_ticklabels(heatmap.xaxis.get_ticklabels(), rotation=0
, ha='right', fontsize=14)
plt.ylabel('Predicted label',size=18)
plt.xlabel('True label',size=18)
plt.title("Confusion Matrix\n",size=24)
plt.show()
```



[5.2.4] Applying RBF SVM on TFIDF W2V, SET 4

In [0]:
```
# Please write all the code with proper documentation
# TF-IDF weighted Word2Vec
tf_idf_vect = TfidfVectorizer()

# final_tf_idf1 is the sparse matrix with row= sentence, col=word and c
ell_val = tfidf
final_tf_idf1 = tf_idf_vect.fit_transform(X_train)
```

```
# tfidf words/col-names
tfidf_feat = tf_idf_vect.get_feature_names()

# compute TFIDF Weighted Word2Vec for each review for X_test .
tfidf_test_vectors = [];
row=0;
for sent in sent_of_test:
    sent_vec = np.zeros(50)
    weight_sum =0;
    for word in sent:
        if word in w2v_words:
            vec = w2v_model.wv[word]
            # obtain the tf_idfidf of a word in a sentence/review
            tf_idf = final_tf_idf1[row, tfidf_feat.index(word)]
            sent_vec += (vec * tf_idf)
            weight_sum += tf_idf
    if weight_sum != 0:
        sent_vec /= weight_sum
    tfidf_test_vectors.append(sent_vec)
    row += 1
```

In [0]:
```
# compute TFIDF Weighted Word2Vec for each review for X_train .
tfidf_train_vectors = [];
row=0;
for sent in sent_of_train:
    sent_vec = np.zeros(50)
    weight_sum =0;
    for word in sent:
        if word in w2v_words:
            vec = w2v_model.wv[word]
            # obtain the tf_idfidf of a word in a sentence/review
            tf_idf = final_tf_idf1[row, tfidf_feat.index(word)]
            sent_vec += (vec * tf_idf)
            weight_sum += tf_idf
    if weight_sum != 0:
        sent_vec /= weight_sum
    tfidf_train_vectors.append(sent_vec)
    row += 1
```

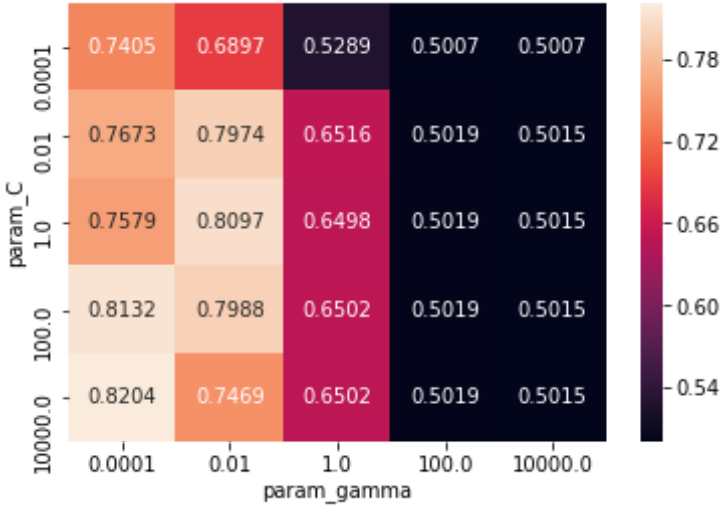
```
# Data-preprocessing: Standardizing the data
sc = StandardScaler()
X_train_vec_standardized = sc.fit_transform(tfidf_train_vectors)
X_test_vec_standardized = sc.transform(tfidf_test_vectors)
```

In [41]:

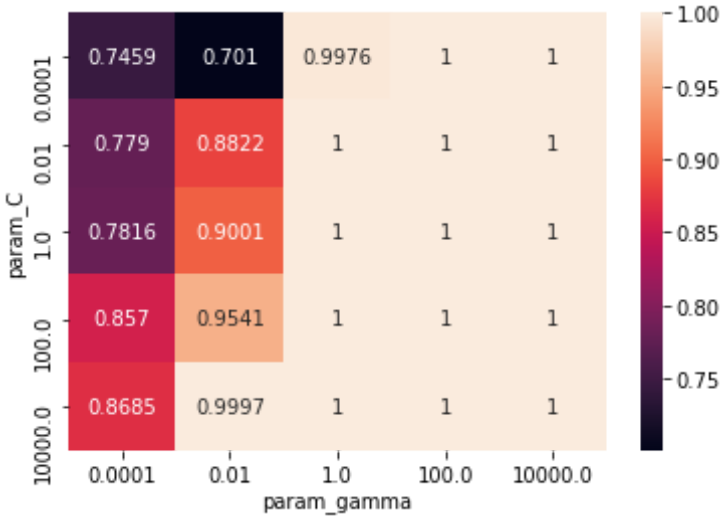
```
# Please write all the code with proper documentation
param_grid = {'C':[10**-4, 10**-2, 10**0, 10**2, 10**4],'gamma':[10**-4
, 10**-2, 10**0, 10**2, 10**4]}
model = GridSearchCV(SVC(kernel = "rbf"), param_grid, scoring = 'roc_au
c', cv=3 , n_jobs = -1,pre_dispatch=2)
model.fit(X_train_vec_standardized, Y_train)
print("Model with best parameters :\n",model.best_estimator_)
print("Accuracy of the model : ",model.score(X_test_vec_standardized, Y
_test))

print("\nOptimal C:", model.best_estimator_.C)
print("\nOptimal Gamma:", model.best_estimator_.gamma)
print("\nBest kernel:", model.best_estimator_.kernel)
print("\nBest Score:", model.best_score_)
df_gridsearch = pd.DataFrame(model.cv_results_)
max_scores = df_gridsearch.groupby(['param_C','param_gamma']).max()
max_scores = max_scores.unstack()[['mean_test_score', 'mean_train_scor
e']]
sns.heatmap(max_scores.mean_test_score, annot=True, fmt='.4g')
plt.show()
```

```
Model with best parameters :
 SVC(C=10000, cache_size=200, class_weight=None, coef0=0.0,
  decision_function_shape='ovr', degree=3, gamma=0.0001, kernel='rbf',
  max_iter=-1, probability=False, random_state=None, shrinking=True,
  tol=0.001, verbose=False)
Accuracy of the model :  0.574779927474866

Optimal C: 10000

Optimal Gamma: 0.0001

Best kernel: rbf
```

```
Best Score: 0.8204366383287528
```



In [42]:
```
df_gridsearch = pd.DataFrame(model.cv_results_)
max_scores = df_gridsearch.groupby(['param_C','param_gamma']).max()
max_scores = max_scores.unstack()[['mean_test_score', 'mean_train_scor
e']]
sns.heatmap(max_scores.mean_train_score, annot=True, fmt='.4g')
plt.show()
```



In [0]:
```
optimal_C = model.best_estimator_.C
print("The optimal value of C is : ",optimal_C)

optimal_gamma = model.best_estimator_.gamma
print("The optimal value of C is : ",optimal_gamma)
```

```
The optimal value of C is :  10000
The optimal value of C is :  0.0001
```

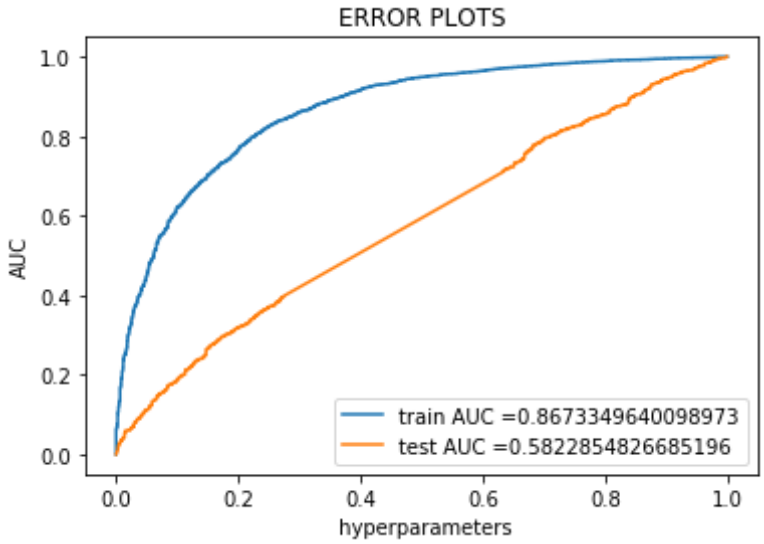
In [0]:
```
# SVC with RBF kernel with Optimal value of C
svc = SVC(C=optimal_C,gamma=optimal_gamma,kernel = "rbf", probability =
 True)
svc.fit(X_train_vec_standardized,Y_train)
predictions = svc.predict(X_test_vec_standardized)
predictions1 = svc.predict(X_train_vec_standardized)

tfidf_w2v_gridrbf_C = optimal_C
tfidf_w2v_gridrbf_train_acc = model.score(X_test_vec_standardized, Y_te
st)*100
tfidf_w2v_gridrbf_test_acc = accuracy_score(Y_test, predictions) * 100
```

In [0]:
```
tfidf_w2v_gridrbf_gamma = optimal_gamma
```

In [0]:
```
calib = CalibratedClassifierCV(svc, method = "sigmoid", cv = "prefit")
calib.fit(X_train_vec_standardized, Y_train)
```

Out[0]:
```
CalibratedClassifierCV(base_estimator=SVC(C=10000, cache_size=200, clas
s_weight=None, coef0=0.0,
  decision_function_shape='ovr', degree=3, gamma=0.0001, kernel='rbf',
  max_iter=-1, probability=True, random_state=None, shrinking=True,
  tol=0.001, verbose=False),
            cv='prefit', method='sigmoid')
```

In [0]:
```
train_fpr, train_tpr, thresholds = roc_curve(Y_train, calib.predict_pro
ba(X_train_vec_standardized)[:,1])
test_fpr, test_tpr, thresholds = roc_curve(Y_test, calib.predict_proba(
X_test_vec_standardized)[:,1])

plt.plot(train_fpr, train_tpr, label="train AUC ="+str(auc(train_fpr, t
rain_tpr)))
plt.plot(test_fpr, test_tpr, label="test AUC ="+str(auc(test_fpr, test_
tpr)))
plt.legend()
plt.xlabel(" hyperparameters")
plt.ylabel("AUC")
plt.title("ERROR PLOTS")
plt.show()
```



In [0]:
```
# evaluate accuracy on test data
acc = accuracy_score(Y_test, predictions) * 100
print('\nThe Test Accuracy of SVC with RBF kernel for C = %.3f is %f%%'
 % (optimal_C, acc))
print('\nThe Test Accuracy of SVC with RBF kernel for gamma = %.4f is %
f%%' % (optimal_gamma, acc))

# evaluate precision
acc = precision_score(Y_test, predictions, pos_label = 1)
print('\nThe Test Precision of SVC with RBF kernel for C = %.3f is %f'
% (optimal_C, acc))
print('\nThe Test Precision of SVC with RBF kernel for gammma = %.4f is
 %f' % (optimal_gamma, acc))

# evaluate recall
acc = recall_score(Y_test, predictions, pos_label = 1)
print('\nThe Test Recall of SVC with RBF kernel for C = %.3f is %f' % (
optimal_C, acc))
print('\nThe Test Recall of SVC with RBF kernel for gamma = %.4f is %f'
 % (optimal_gamma, acc))

# evaluate f1-score
```

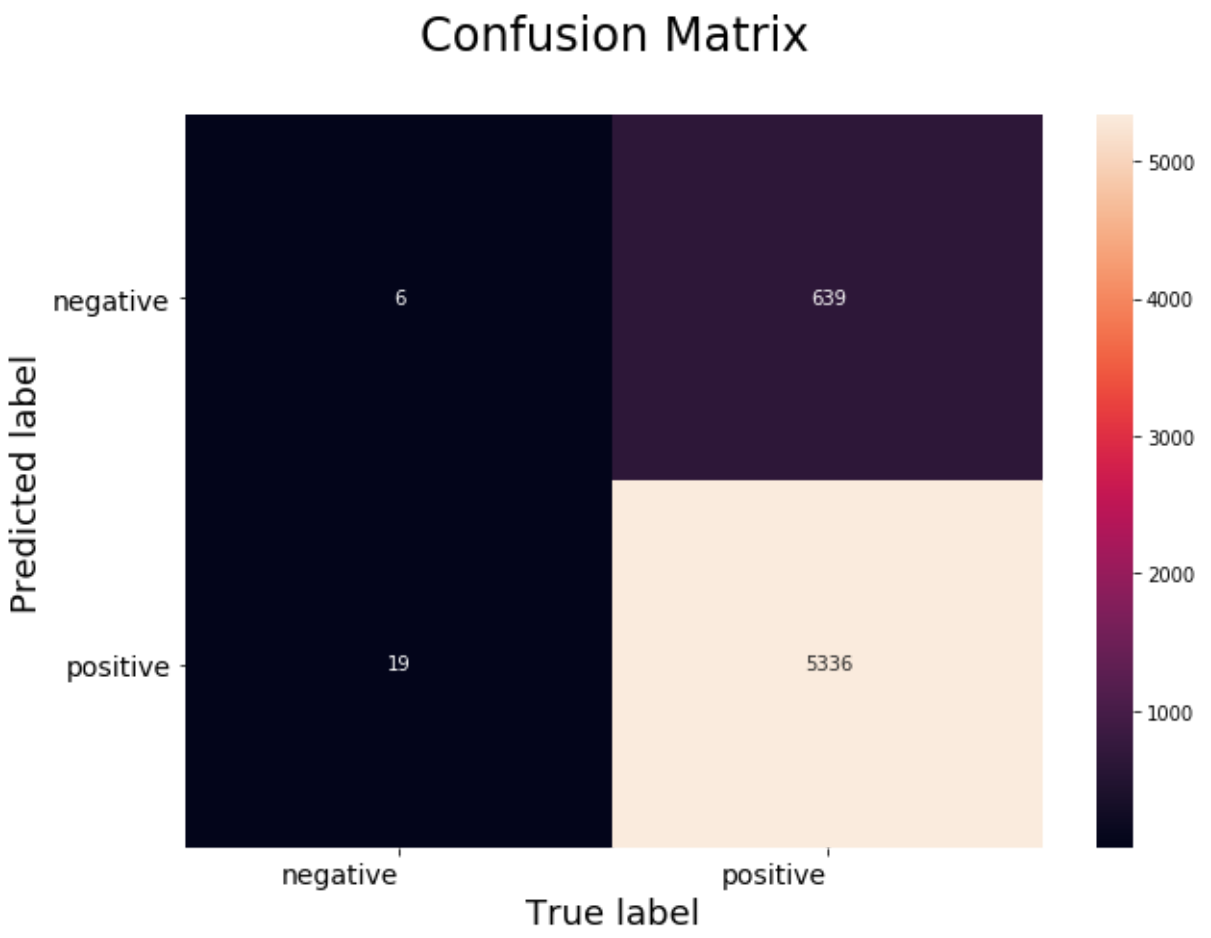
```
acc = f1_score(Y_test, predictions, pos_label = 1)
print('\nThe Test F1-Score of SVC with RBF kernel for C = %.3f is %f' %
 (optimal_C, acc))
print('\nThe Test F1-Score of SVC with RBF kernel for gamma = %.4f is %
f' % (optimal_gamma, acc))
```

```
The Test Accuracy of SVC with RBF kernel for C = 10000.000 is 89.03333
3%

The Test Accuracy of SVC with RBF kernel for gamma = 0.0001 is 89.03333
3%

The Test Precision of SVC with RBF kernel for C = 10000.000 is 0.893054

The Test Precision of SVC with RBF kernel for gammma = 0.0001 is 0.8930
54

The Test Recall of SVC with RBF kernel for C = 10000.000 is 0.996452

The Test Recall of SVC with RBF kernel for gamma = 0.0001 is 0.996452

The Test F1-Score of SVC with RBF kernel for C = 10000.000 is 0.941924

The Test F1-Score of SVC with RBF kernel for gamma = 0.0001 is 0.941924
```

In [0]:

```
# Code for drawing seaborn heatmaps on test data
class_names = ['negative','positive']
df_heatmap = pd.DataFrame(confusion_matrix(Y_test, predictions), index=
class_names, columns=class_names )
fig = plt.figure(figsize=(10,7))
heatmap = sns.heatmap(df_heatmap, annot=True, fmt="d")

# Setting tick labels for heatmap
heatmap.yaxis.set_ticklabels(heatmap.yaxis.get_ticklabels(), rotation=0
, ha='right', fontsize=14)
heatmap.xaxis.set_ticklabels(heatmap.xaxis.get_ticklabels(), rotation=0
, ha='right', fontsize=14)
plt.ylabel('Predicted label',size=18)
plt.xlabel('True label',size=18)
```

```
plt.title("Confusion Matrix\n",size=24)
plt.show()
```

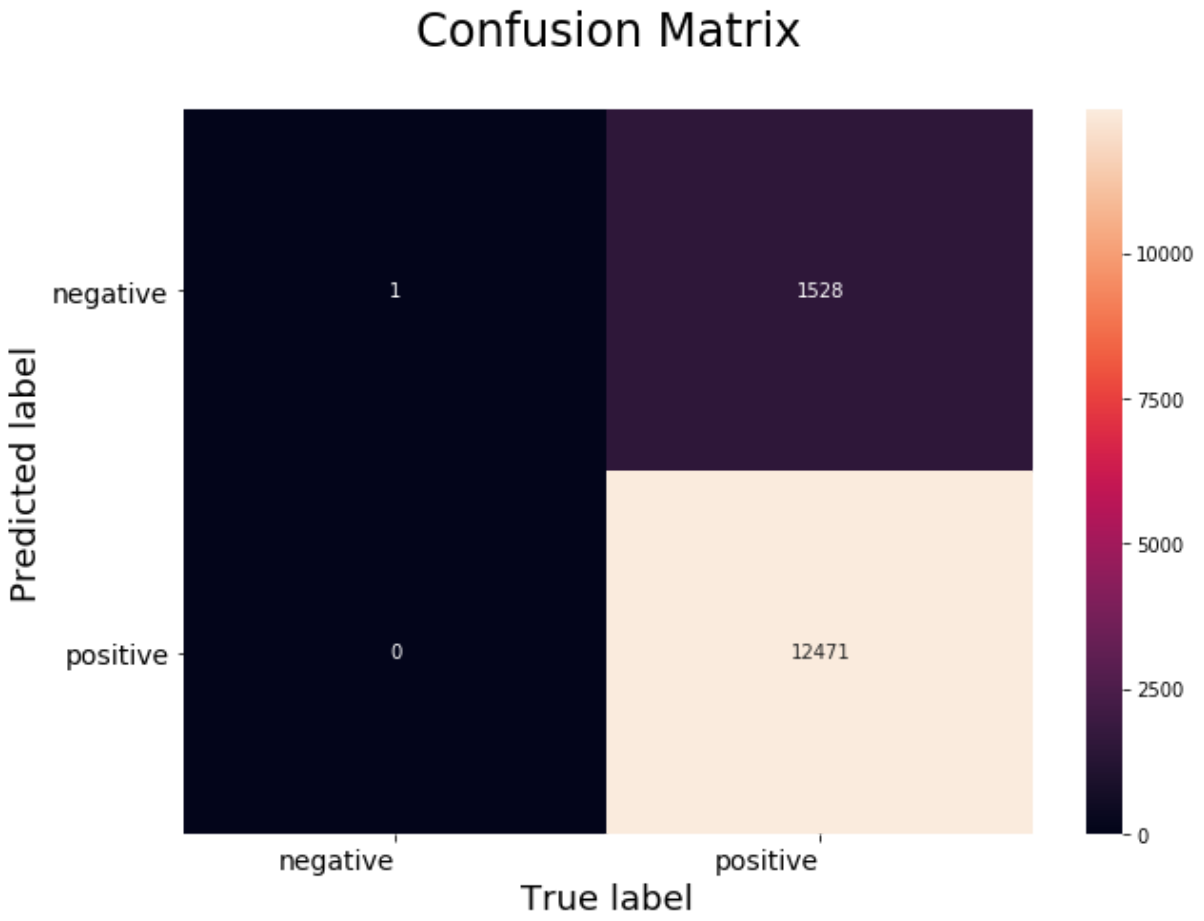


In [0]:
```
# evaluate accuracy on train data
acc = accuracy_score(Y_train, predictions1) * 100
print('\nThe Train accuracy of SVC with RBF kernel for C = %.3f is %f%%
' % (optimal_C, acc))
print('\nThe Train accuracy of SVC with RBF kernel for gamma = %.4f is
%f%%' % (optimal_gamma, acc))

# evaluate precision
```

```
acc = precision_score(Y_train, predictions1, pos_label = 1)
print('\nThe Train precision of SVC with RBF kernel for C = %.3f is %f'
 % (optimal_C, acc))
print('\nThe Train precision of SVC with RBF kernel for gamma = %.4f is
 %f' % (optimal_gamma, acc))

# evaluate recall
acc = recall_score(Y_train, predictions1, pos_label = 1)
print('\nThe Train Recall of SVC with RBF kernel for C = %.3f is %f' %
(optimal_C, acc))
print('\nThe Train Recall of SVC with RBF kernel for gamma = %.4f is %f
' % (optimal_gamma, acc))

# evaluate f1-score
acc = f1_score(Y_train, predictions1, pos_label = 1)
print('\nThe Train f1score of SVC with RBF kernel for C = %.3f is %f' %
 (optimal_C, acc))
print('\nThe Train f1score of SVC with RBF kernel for gamma = %.4f is %
f' % (optimal_gamma, acc))
```

```
The Train accuracy of SVC with RBF kernel for C = 10000.000 is 89.08571
4%

The Train accuracy of SVC with RBF kernel for gamma = 0.0001 is 89.0857
14%

The Train precision of SVC with RBF kernel for C = 10000.000 is 0.89084
9

The Train precision of SVC with RBF kernel for gamma = 0.0001 is 0.8908
49

The Train Recall of SVC with RBF kernel for C = 10000.000 is 1.000000

The Train Recall of SVC with RBF kernel for gamma = 0.0001 is 1.000000

The Train f1score of SVC with RBF kernel for C = 10000.000 is 0.942274

The Train f1score of SVC with RBF kernel for gamma = 0.0001 is 0.942274
```

In [0]:
```
# Code for drawing seaborn heatmaps on train data
class_names = ['negative','positive']
df_heatmap = pd.DataFrame(confusion_matrix(Y_train, predictions1), inde
x=class_names, columns=class_names )
fig = plt.figure(figsize=(10,7))
heatmap = sns.heatmap(df_heatmap, annot=True, fmt="d")

# Setting tick labels for heatmap
heatmap.yaxis.set_ticklabels(heatmap.yaxis.get_ticklabels(), rotation=0
, ha='right', fontsize=14)
heatmap.xaxis.set_ticklabels(heatmap.xaxis.get_ticklabels(), rotation=0
, ha='right', fontsize=14)
plt.ylabel('Predicted label',size=18)
plt.xlabel('True label',size=18)
plt.title("Confusion Matrix\n",size=24)
plt.show()
```



procedure

- STEP 1 :- Text Preprocessing
- STEP 2:- Time-based splitting of whole dataset into train_data and test_data
- STEP 3:- Training the vectorizer on train_data and later applying same vectorizer on both train_data and test_data to transform them into vectors
- STEP 4:- Using SGD classifier/SVC as an estimator in 10-Fold Cross-Validation in order to find optimal value of alpha .

- STEP 5:- Draw various plots auc's vs aplha,k
- STEP 6:- Once , we get optimal value of alpha then train SGD classifier/SVC again with this optimal alpha and make predictions on test_data
- STEP 7:- Find important features per class
- STEP 8 :- Evaluate : Accuracy , F1-Score , Precision , Recall , TPR , FPR , TNR , FNR
- STEP 9:- Draw Seaborn Heatmap for Confusion Matrix .

[6] Conclusions

In [0]:
```
# Please compare all your models using Prettytable library
# Creating table using PrettyTable library
from prettytable import PrettyTable

# Names of models
names = ['SGD in GridSearchCV for BoW',\
         'SGD in GridSearchCV for TFIDF',\
         'SGD in GridSearchCV for Avg_Word2Vec',\
         'SGD in GridSearchCV for tfidf_Word2Vec']

# Optimal values of alpha i.e. (1/C)
optimal_alpha = [bow_grid_alpha,tfidf_grid_alpha,avg_w2v_grid_alpha,tfi
df_avg_w2v_grid_alpha]

# Training accuracies
train_acc = [81.85, 79.14, 100,88.25]

# Test accuracies
test_acc = [bow_grid_test_acc,tfidf_grid_test_acc,avg_w2v_grid_test_acc
,tfidf_avg_w2v_grid_test_acc]

numbering = [1,2,3,4]

# Initializing prettytable
ptable = PrettyTable()

# Adding columns
```

```
ptable.add_column("S.NO.",numbering)
ptable.add_column("MODEL",names)
ptable.add_column("Best Alpha(1/C)",optimal_alpha)
ptable.add_column("Training Accuracy",train_acc)
ptable.add_column("Test Accuracy",test_acc)

# Printing the Table
print(ptable)
```

```
+-------+----------------------------------------+-----------------+---
---------------+-------------------+
| S.NO. |                 MODEL                  | Best Alpha(1/C) | Tr
aining Accuracy |   Test Accuracy   |
+-------+----------------------------------------+-----------------+---
---------------+-------------------+
|   1   |      SGD in GridSearchCV for BoW       |        1        |
    81.85       | 87.57333333333334 |
|   2   |     SGD in GridSearchCV for TFIDF      |        1        |
    79.14       | 86.54666666666667 |
|   3   |  SGD in GridSearchCV for Avg_Word2Vec  |       0.01      |
     100        |       78.02       |
|   4   | SGD in GridSearchCV for tfidf_Word2Vec |       0.1       |
    88.25       |       88.08       |
+-------+----------------------------------------+-----------------+---
---------------+-------------------+
```

In [0]:
```
# Please compare all your models using Prettytable library
# Creating table using PrettyTable library
from prettytable import PrettyTable

# Names of models
names = ['SVC in GridSearchCV for BoW',\
         'SVC in GridSearchCV for TFIDF',\
         'SVC in GridSearchCV for Avg_Word2Vec',\
         'SVC in GridSearchCV for tfidf_Word2Vec']

# Optimal values of alpha i.e. (1/C)
optimal_C = [bow_gridrbf_C,tfidf_gridrbf_C,avg_w2v_gridrbf_C,tfidf_w2v_
gridrbf_C]
```

```
optimal_gamma = [bow_gridrbf_gamma,tfidf_gridrbf_gamma,avg_w2v_gridrbf_
gamma,tfidf_w2v_gridrbf_gamma]

# Training accuracies
train_acc = [95.48 ,96.00,80.4,89.08]

# Test accuracies
test_acc = [bow_gridrbf_test_acc,tfidf_gridrbf_test_acc,avg_w2v_gridrbf
_test_acc,tfidf_w2v_gridrbf_test_acc]

numbering = [1,2,3,4]

# Initializing prettytable
ptable = PrettyTable()

# Adding columns
ptable.add_column("S.NO.",numbering)
ptable.add_column("MODEL",names)
ptable.add_column("Best C",optimal_C)
ptable.add_column("Best gamma",optimal_gamma)
ptable.add_column("Training Accuracy",train_acc)
ptable.add_column("Test Accuracy",test_acc)

# Printing the Table
print(ptable)
```

```
+-------+---------------------------------------+--------+------------
+------------------+------------------+
| S.NO. |                 MODEL                 | Best C | Best gamma
| Training Accuracy |   Test Accuracy   |
+-------+---------------------------------------+--------+------------
+------------------+------------------+
|   1   |      SVC in GridSearchCV for BoW      |   1    |   0.0001
|       95.48       | 89.48333333333333 |
|   2   |     SVC in GridSearchCV for TFIDF     |   1    |   0.0001
|        96.0       | 89.13333333333333 |
|   3   |  SVC in GridSearchCV for Avg_Word2Vec |   1    |    0.01
|        80.4       |        79.9       |
```

```
|   4   | SVC in GridSearchCV for tfidf_Word2Vec | 10000  |   0.0001
|       89.08       | 89.03333333333333 |
+-------+----------------------------------------+--------+------------
+------------------+------------------+
```