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 [3]:

```
# using SQLite Table to read data.
con = sqlite3.connect('database.sqlite')

# filtering only positive and negative reviews i.e.
# not taking into consideration those reviews with Score=3
# SELECT * FROM Reviews WHERE Score != 3 LIMIT 500000, will give top 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""", 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 (525814, 10)
```

Out[3]:

	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	COU
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	COU
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	(
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
sample_data = filtered_data.sample(n=50000,random_state=0)
sorted_data=sample_data.sort_values('ProductId', axis=0, ascending=True
, inplace=False, kind='quicksort', na_position='last')
```

In [5]:
```
#Deduplication of entries
final=sorted_data.drop_duplicates(subset={"UserId","ProfileName","Time"
,"Text"}, keep='first', inplace=False)
final.shape
```

Out[5]:
```
(46100, 10)
```

In [6]:
```
#Checking to see how much % of data still remains
(final['Id'].size*1.0)/(filtered_data['Id'].size*1.0)*100
```

Out[6]:
```
8.767358799879805
```

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	Helpfuln
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-REFILL/dp/B00004RBDY<
br />http://www.amazon.com/VICTOR-FLY-MAGNET-BAIT-REFILL/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]:

```python
# https://stackoverflow.com/a/47091490/4084039
import re
from bs4 import BeautifulSoup

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/18082370/4084039
```

```
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 [9]:

```
# 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%|██████████| 46100/46100 [00:21<00:00, 2141.91it/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-grams
# count_vect = CountVectorizer(ngram_range=(1,2))
# please do read the CountVectorizer documentation http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

# you can choose these numebrs min_df=10, max_features=5000, of your choice
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_shape())
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
```

[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 values
# To use this code-snippet, download "GoogleNews-vectors-negative300.bin"
# from https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit
# it's 1.9GB in size.


# http://kavita-ganesan.com/gensim-word2vec-tutorial-starter-code/#.W17SRFAzZPY
# 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_train_w2v = True, to train your own w2v ")
```

```
[('snack', 0.9951335191726685), ('calorie', 0.9946465492248535), ('wonderful', 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 9: Random Forests

1. **Apply Random Forests & GBDT on these feature sets**

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

2. **The hyper paramter tuning (Consider two hyperparameters: n_estimators & max_depth)**

    - 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

3. **Feature importance**

- Get top 20 important features and represent them in a word cloud. Do this for BOW & TFIDF.

4. **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.

5. **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 with X-axis as **n_estimators**, Y-axis as **max_depth**, and Z-axis as **AUC Score** , we have given the notebook which explains how to plot this 3d plot, you can find it in the same drive *3d_scatter_plot.ipynb*

(or)

- 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 seaborn heat maps with rows as **n_estimators**, columns as **max_depth**, and values inside the cell representing **AUC Score**
- You choose either of the plotting techniques out of 3d plot or heat map
- 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.

6. **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.

[5.1] Applying RF

In [0]:
```
final["Clean_text"] = preprocessed_reviews
```

In [0]:
```
final = final.sort_values('Time', axis=0, ascending=True, inplace=False
, kind='quicksort', na_position='last')
#split data in train, test and cv before using it to avoid data leakage
from sklearn.model_selection import train_test_split

X = final['Clean_text']
y = final['Score']

X_train,X_test,y_train_,y_test_ = train_test_split(X,y,test_size=.5,ran
dom_state=0,shuffle=False)
```

```
X_cv,X_test,y_cv_,y_test_ = train_test_split(X_test,y_test_,test_size=.
5,random_state=0,shuffle=False)
```

In [0]:

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import roc_auc_score
from sklearn.metrics import confusion_matrix
import sklearn
from tqdm import tqdm
import matplotlib.pyplot as plt
import numpy as np
#function to find an optimal value of k,AUC,ROC,confusion matrix
def best_para(x_train,x_cv,x_test,y_train,y_cv,y_test):
        n_estimators = [1, 4, 16, 64, 100, 200]
        depth = [1, 2, 4, 8, 16, 32]
        param_grid = {'max_depth':[1, 2, 4, 8, 16, 32],'n_estimators':[
1, 4, 16, 64, 100, 200]}
        grid_search = GridSearchCV(estimator = RandomForestClassifier
(),param_grid=param_grid ,cv = 5,n_jobs = -1,return_train_score=True)
        grid_search.fit(x_train, y_train)

        auc_cv = grid_search.cv_results_['mean_test_score'].reshape(6,6
)
        auc = grid_search.cv_results_['mean_train_score'].reshape(6,6)
        print('best depth = ',grid_search.best_estimator_ .max_depth)
        print('best n_estimators = ',grid_search.best_estimator_ .n_est
imators)



        # matrix for auc
        confusion_matr = pd.DataFrame(auc,index=n_estimators,columns=de
pth)
        sns.heatmap(confusion_matr,annot=True,cmap='viridis')
        plt.title("AUC matrix for train")
        plt.xlabel("depth")
        plt.ylabel("n_estimators")
        plt.show()
        confusion_matr_cv = pd.DataFrame(auc_cv,index=n_estimators,columns=depth)
        sns.heatmap(confusion_matr_cv,annot=True,cmap='viridis')
        plt.title("AUC matrix for cv")
        plt.xlabel("depth")
        plt.ylabel("n_estimators")
        plt.show()



def test(x_train,x_cv,x_test,y_train,y_cv,y_test,depth,n_estimators ):
        clf = RandomForestClassifier(max_depth = depth, n_estimators  = n_estimators , class_weight = 'balanced')
        clf.fit(x_train,y_train)
        prob_test = clf.predict_proba(x_test)
        prob_test = prob_test[:,1]

        prob_train = clf.predict_proba(x_train)
        prob_train = prob_train[:,1]
        print("AUC Score: {}".format(roc_auc_score(y_test,prob_test)))

        #ROC curve
        fpr_tr,tpr_tr,thres_tr = roc_curve(y_train,prob_train)
        fpr,tpr,thres = roc_curve(y_test,prob_test)
        plt.plot([0,0],[1,1],linestyle='--')
        plt.plot(fpr,tpr,'r',marker='.',label='test')
        plt.plot(fpr_tr,tpr_tr,'b',marker='.',label='train')
        plt.legend(loc='upper right')
        plt.title("ROC curve")
        plt.show()

        #confusion matrix fortrain and test
        print("Confusion matrix for train data")
        predict_tr = clf.predict(x_train)
        confu_metrix_(y_train,predict_tr)

        print("Confusion matrix for test data")
        predict_te = clf.predict(x_test)
        confu_metrix_(y_test,predict_te)
```

```
def confu_metrix_(y,predict):
    confu_metrix = confusion_matrix(y,predict)
    confu_df = pd.DataFrame(confu_metrix,index=["-ve","+ve"],columns=[
"-ve","+ve"])
    sns.heatmap(confu_df,annot=True,fmt='d',cmap='viridis')
    plt.title("Confusion matrix")
    plt.xlabel("predicted label")
    plt.ylabel("True label")
    plt.show()
```

[5.1.1] Applying Random Forests on BOW, SET 1

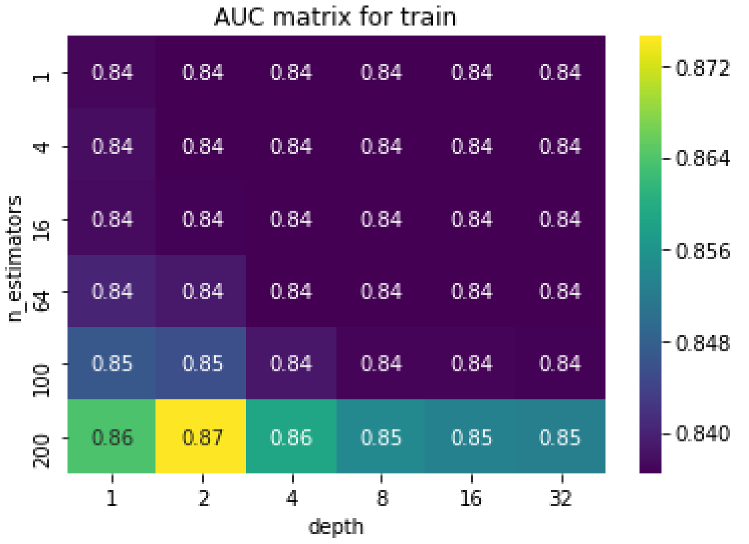
In [0]:

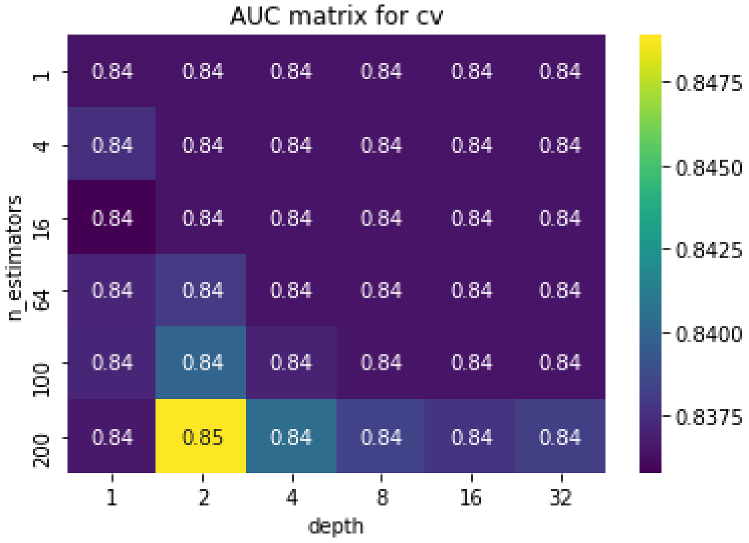
```
# Please write all the code with proper documentation
import pickle
bow_cv,bow_test,bow_train = pickle.load(open("bow.pkl",'rb'))
y_cv,y_test,y_train = pickle.load(open("label.pkl",'rb'))
```

In [0]:

```
best_para(bow_train,bow_cv,bow_test,y_train,y_cv,y_test)
```

```
best depth =  32
best n_estimators =  4
```

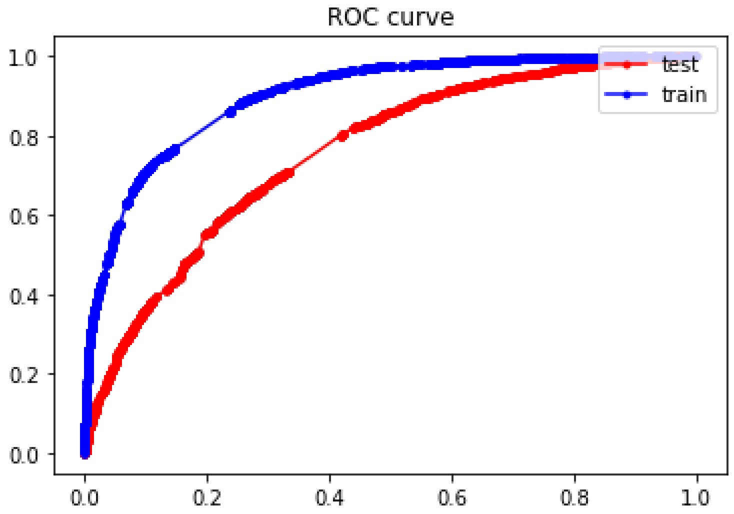




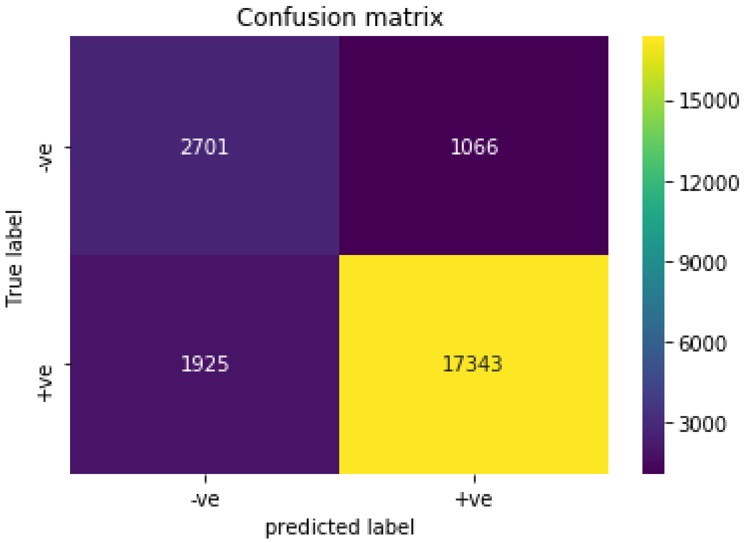
In [0]:

```
test(bow_train,bow_cv,bow_test,y_train,y_cv,y_test,32,4)
```

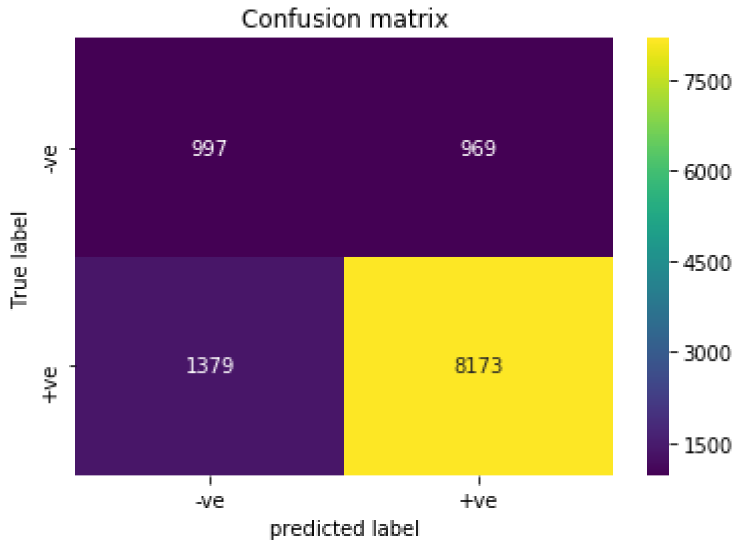
```
AUC Score: 0.7592166442163343
```



```
Confusion matrix for train data
```



```
Confusion matrix for test data
```



[5.1.2] Wordcloud of top 20 important features from SET 1

In [0]:

```
pip install wordcloud
```

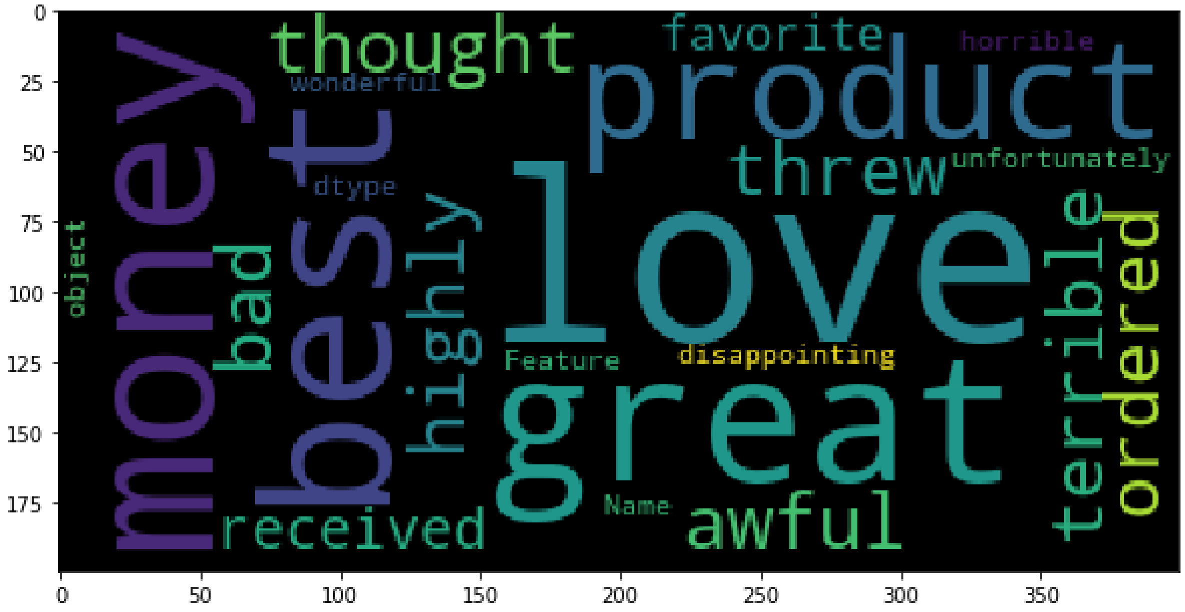
```
Requirement already satisfied: wordcloud in /usr/local/lib/python3.6/di
st-packages (1.5.0)
Requirement already satisfied: numpy>=1.6.1 in /usr/local/lib/python3.
6/dist-packages (from wordcloud) (1.16.3)
Requirement already satisfied: pillow in /usr/local/lib/python3.6/dist-
packages (from wordcloud) (4.3.0)
Requirement already satisfied: olefile in /usr/local/lib/python3.6/dist
-packages (from pillow->wordcloud) (0.46)
```

In [0]:

```
# Please write all the code with proper documentation
cnt_vec = CountVectorizer()
p = cnt_vec.fit_transform(X_train)

clf = RandomForestClassifier(max_depth = 32, n_estimators  = 4,class_we
ight='balanced')
clf.fit(p,y_train_)
```

```
feat_log_prob = clf.feature_importances_

p = pd.DataFrame(feat_log_prob.T,columns=['+ve'])
p["Feature"] = cnt_vec.get_feature_names()
p = p.sort_values(by = '+ve',kind = 'quicksort',ascending= False)
```

In [0]:
```
from wordcloud import WordCloud

wordcloud = WordCloud().generate(str(p[:20]['Feature']))
plt.figure(figsize=(10,10))
plt.imshow(wordcloud)
plt.show()
```

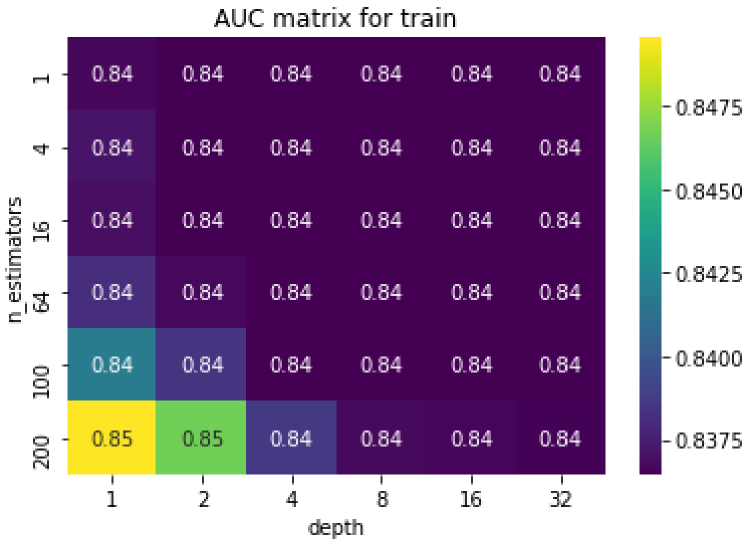


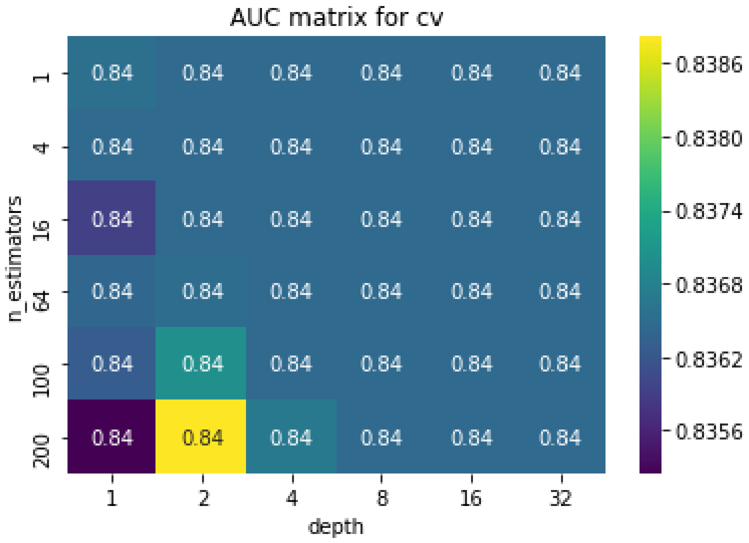
[5.1.3] Applying Random Forests on TFIDF, SET 2

In [0]:
```
# Please write all the code with proper documentation
tfidf_cv,tfidf_test,tfidf_train = pickle.load(open("tfidf.pkl",'rb'))
```

```
In [0]: best_para(tfidf_train,tfidf_cv,tfidf_test,y_train,y_cv,y_test)
```

```
best depth =  32
best n_estimators =  4
```

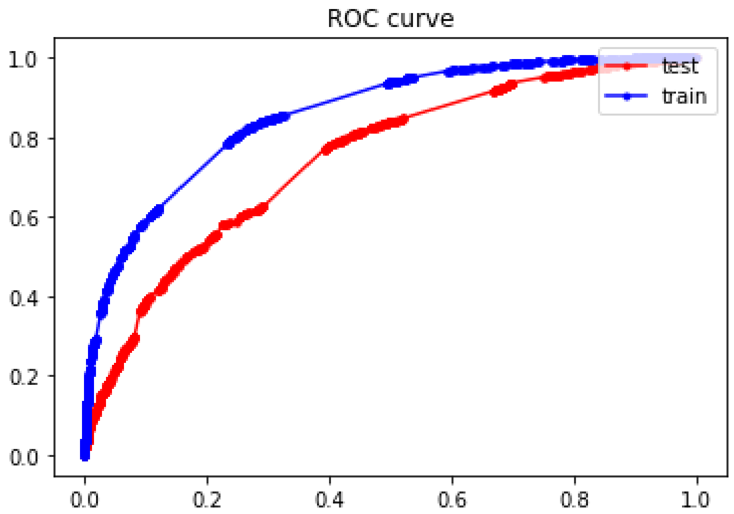




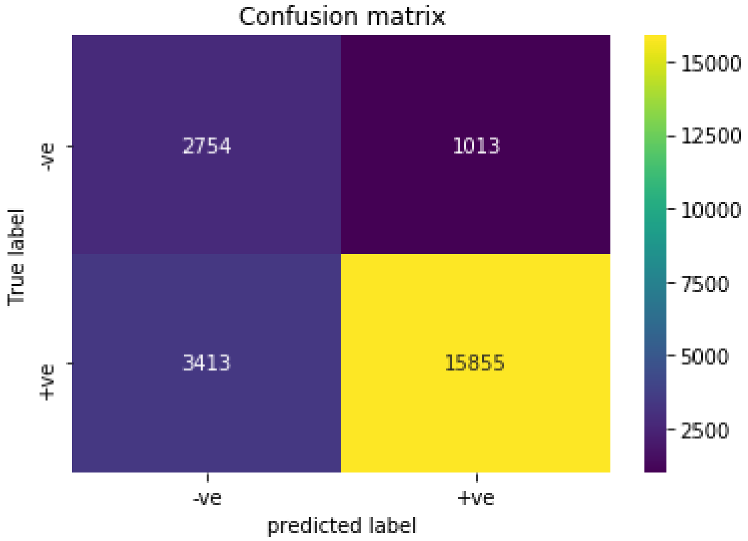
In [0]:

```
test(tfidf_train,tfidf_cv,tfidf_test,y_train,y_cv,y_test,32,4)
```

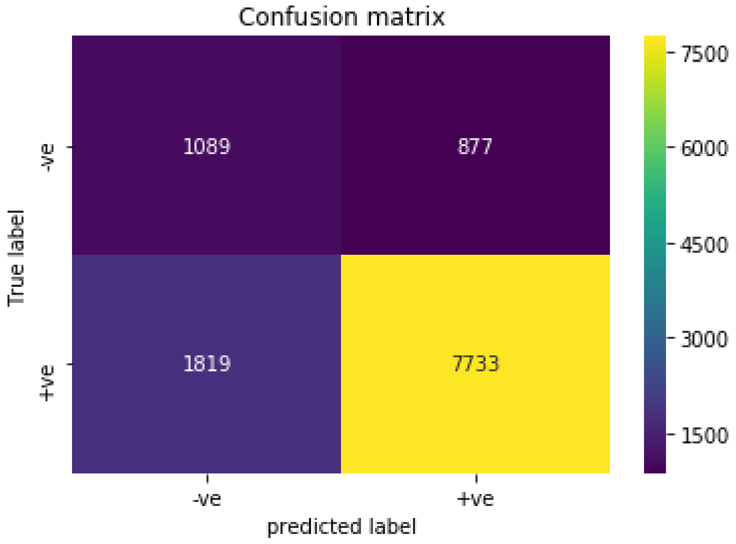
```
AUC Score: 0.7482731721936232
```



```
Confusion matrix for train data
```



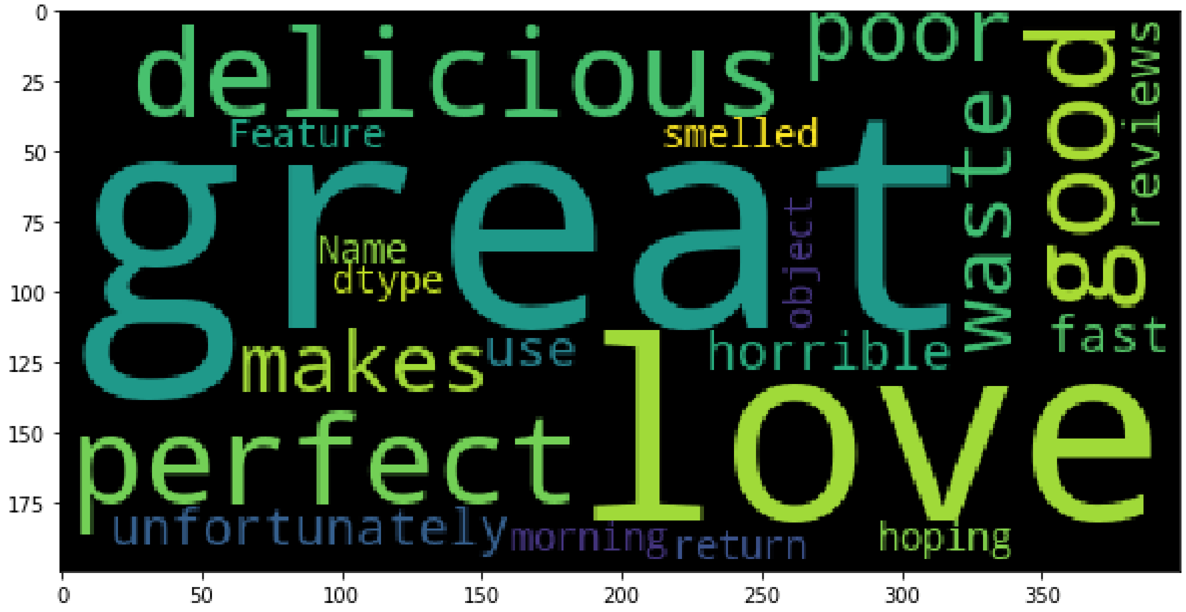
```
Confusion matrix for test data
```



[5.1.4] Wordcloud of top 20 important features from SET 2

In [0]:
```
# Please write all the code with proper documentation
cnt_vec = TfidfVectorizer()
p = cnt_vec.fit_transform(X_train)

clf = RandomForestClassifier(max_depth = 32, n_estimators  = 4,class_we
ight='balanced')
clf.fit(p,y_train_)
feat_log_prob = clf.feature_importances_

p = pd.DataFrame(feat_log_prob.T,columns=['+ve'])
p["Feature"] = cnt_vec.get_feature_names()
p = p.sort_values(by = '+ve',kind = 'quicksort',ascending= False)
```

In [0]:
```
from wordcloud import WordCloud

wordcloud = WordCloud().generate(str(p[:20]['Feature']))
plt.figure(figsize=(10,10))
plt.imshow(wordcloud)
plt.show()
```



[5.1.5] Applying Random Forests on AVG W2V, SET 3

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

In [14]:
```
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 ")
```

```
[('excellent', 0.8381518721580505), ('awesome', 0.814774751663208), ('g
ood', 0.8101101517677307), ('fantastic', 0.8042314052581787), ('wonderf
ul', 0.7979162931442261), ('terrific', 0.7955546975135803), ('amazing',
0.761521577835083), ('delicious', 0.7269079685211182), ('decent', 0.724
6918082237244), ('perfect', 0.7103558778762817)]
==================================================
[('disgusting', 0.794817328453064), ('closest', 0.7763254642486572),
('experienced', 0.7750902771949768), ('unmatched', 0.7720858454704285),
('musty', 0.7682808637619019), ('remember', 0.7680240869522095), ('surp
asses', 0.760034978389 74), ('smoothest', 0.7555502653121948), ('ive',
0.7524262070655823), ('tastiest', 0.7452026009559631)]
```

In [15]:

```
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  9345
sample words  ['happens', 'say', 'name', 'three', 'times', 'michael',
'stars', 'comedy', 'two', 'live', 'old', 'story', 'house', 'coming', 'b
ack', 'supply', 'store', 'couple', 'suddenly', 'get', 'caught', 'insid
e', 'broken', 'start', 'lake', 'board', 'got', 'hopes', 'small', 'dog',
'steps', 'car', 'starts', 'slide', 'waters', 'minutes', 'later', 'fin
d', 'home', 'somehow', 'light', 'fireplace', 'done', 'magic', 'weird',
'looking', 'dead', 'guy', 'known', 'way']
```

In [0]:

```
def vectorize_W2V(data):
    sent_vectors = []; # the avg-w2v for each sentence/review is stored
 in this list
    for sent in tqdm(data): # for each review/sentence
        sent_vec = np.zeros(50) # as word vectors are of zero length 5
0, you 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 sentenc
e/review
        for word in sent.split(): # 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)
    return sent_vectors
```

In [17]:

```
# vectorize all train,test and cv data
avg_w2v_train = vectorize_W2V(X_train)
avg_w2v_cv = vectorize_W2V(X_cv)
avg_w2v_test = vectorize_W2V(X_test)
```

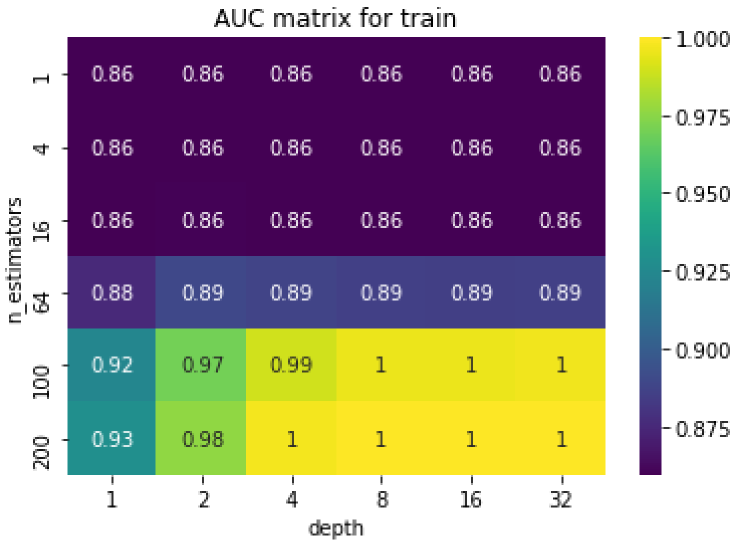
```
100%|██████████| 23050/23050 [00:35<00:00, 640.64it/s]
100%|██████████| 11525/11525 [00:18<00:00, 613.99it/s]
100%|██████████| 11525/11525 [00:19<00:00, 598.14it/s]
```

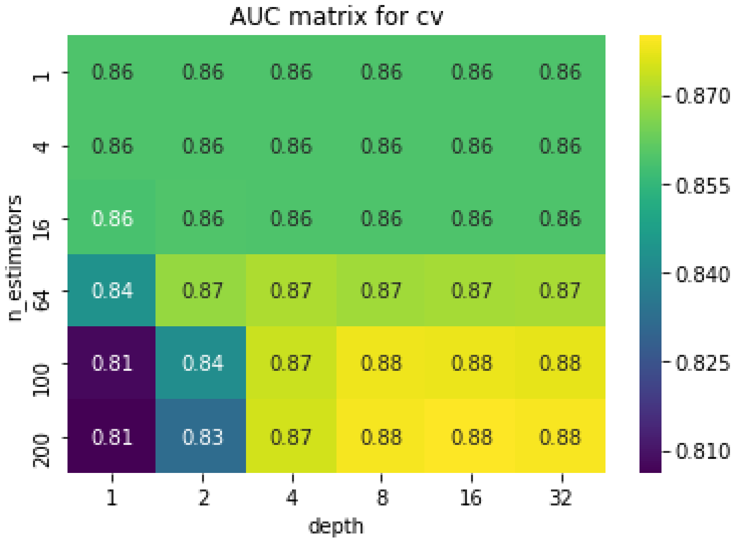
In [0]:

```
best_para(avg_w2v_train,avg_w2v_cv,avg_w2v_test,y_train_,y_cv_,y_test_)
```

```
/usr/local/lib/python3.6/dist-packages/joblib/externals/loky/process_ex
ecutor.py:700: UserWarning: A worker stopped while some jobs were given
to the executor. This can be caused by a too short worker timeout or by
a memory leak.
  "timeout or by a memory leak.", UserWarning
/usr/local/lib/python3.6/dist-packages/joblib/externals/loky/process_ex
ecutor.py:700: UserWarning: A worker stopped while some jobs were given
to the executor. This can be caused by a too short worker timeout or by
a memory leak.
  "timeout or by a memory leak.", UserWarning
```

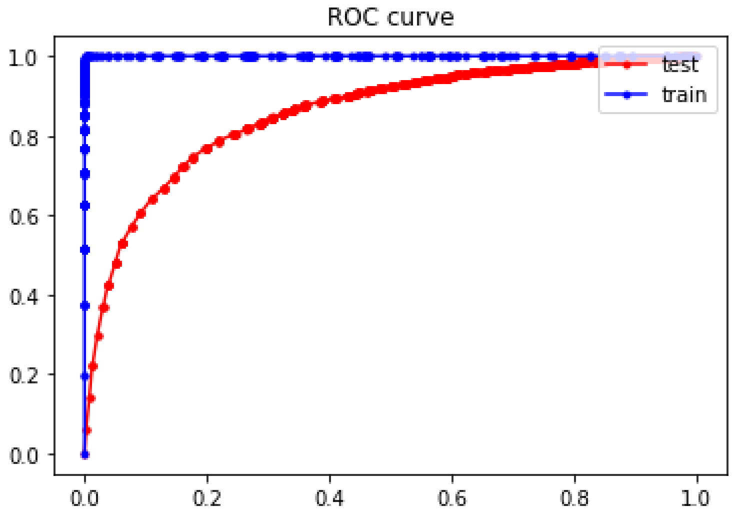
```
best depth =  32
best n_estimators =  100
```



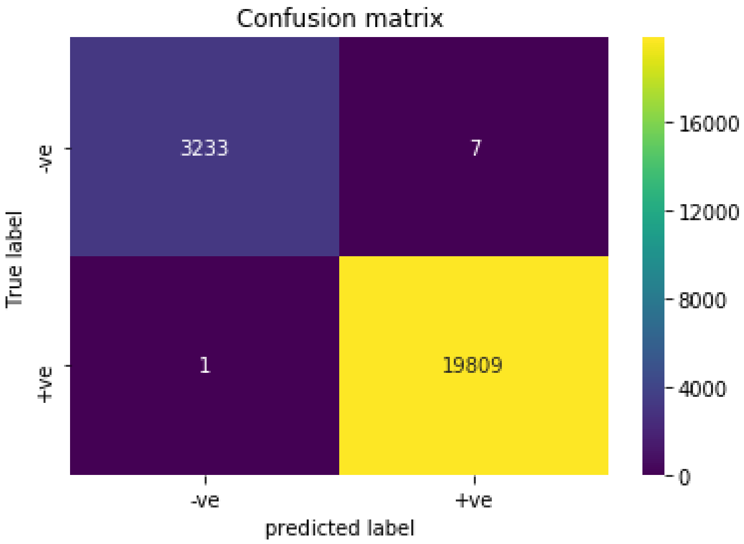


```
In [0]: test(avg_w2v_train,avg_w2v_cv,avg_w2v_test,y_train_,y_cv_,y_test_,32,100)
```

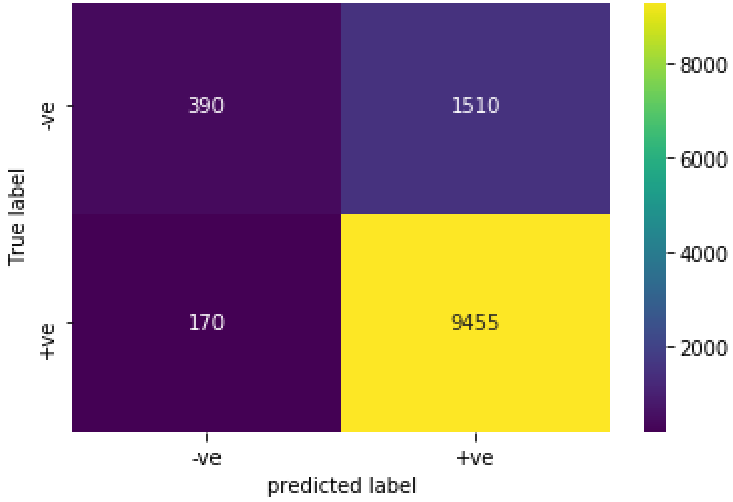
```
AUC Score: 0.857815010252905
```



```
Confusion matrix for train data
```



```
Confusion matrix for test data
```



[5.1.6] Applying Random Forests on TFIDF W2V, SET 4

In [0]:
```
# Please write all the code with proper documentation
model_ = TfidfVectorizer(ngram_range=(1,2), min_df=10,max_features=500)
tf_idf_matrix = model_.fit_transform(X_train)
# 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
def vectorizer_W2V_tfidf(data):
    tfidf_sent_vectors = []; # the tfidf-w2v for each sentence/review i
s stored in this list
    row=0;
    for sent in tqdm(data): # 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 senten
ce/review
```

```
        for word in sent.split(): # for each word in a review/sentence
            if word in w2v_words and word in tfidf_feat:
                vec = w2v_model.wv[word]
    #             tf_idf = tf_idf_matrix[row, tfidf_feat.index(word)]
                # to reduce the computation we are
                # dictionary[word] = idf value of word in whole courpus
                # sent.count(word) = tf valeus of word in this review
                tf_idf = dictionary[word]*(sent.count(word)/len(sent))
                sent_vec += (vec * tf_idf)
                weight_sum += tf_idf
        if weight_sum != 0:
            sent_vec /= weight_sum
        tfidf_sent_vectors.append(sent_vec)
        row += 1
    return tfidf_sent_vectors
```

In [20]:

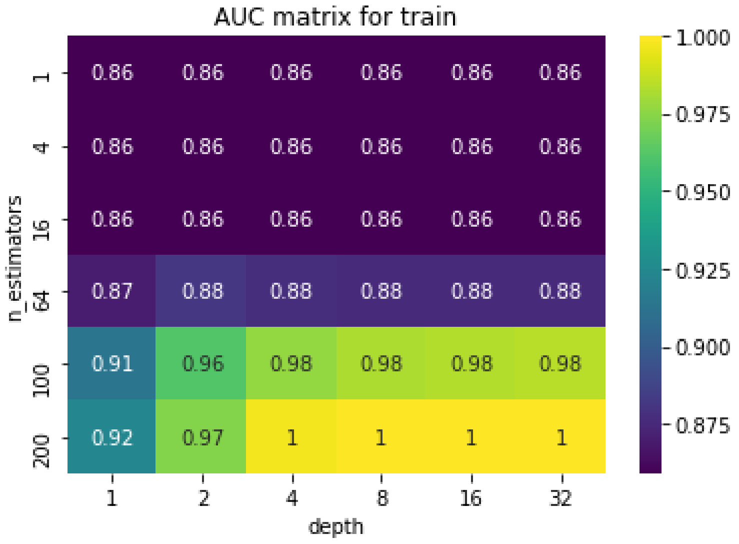
```
tfidf_w2v_train = vectorizer_W2V_tfidf(X_train)
tfidf_w2v_cv = vectorizer_W2V_tfidf(X_cv)
tfidf_w2v_test = vectorizer_W2V_tfidf(X_test)
```

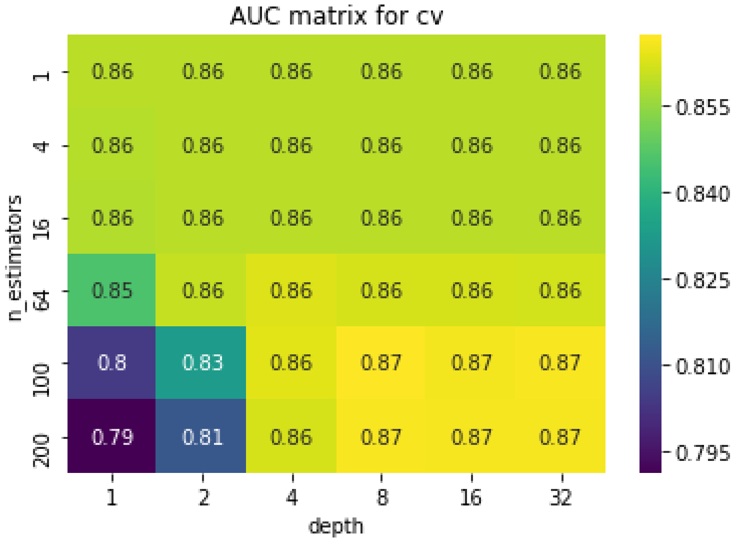
```
100%|██████████| 23050/23050 [00:47<00:00, 489.93it/s]
100%|██████████| 11525/11525 [00:24<00:00, 464.11it/s]
100%|██████████| 11525/11525 [00:24<00:00, 461.60it/s]
```

In [0]:

```
best_para(tfidf_w2v_train,tfidf_w2v_cv,tfidf_w2v_test,y_train_,y_cv_,y_
test_)
```

```
best depth =  16
best n_estimators =  64
```

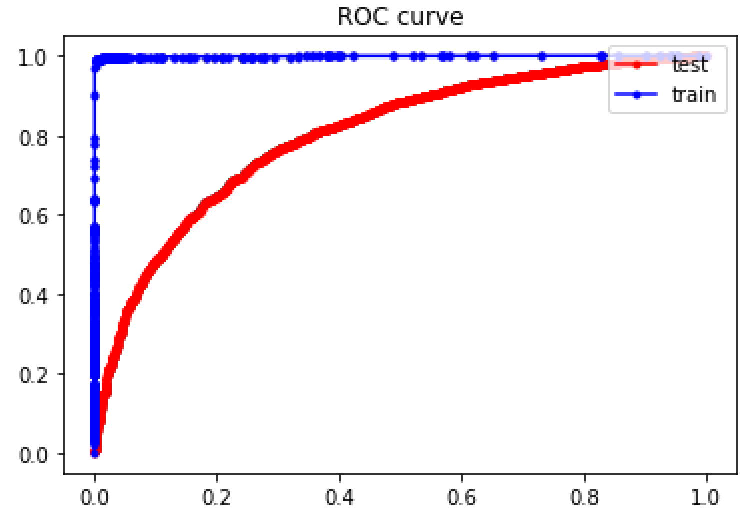




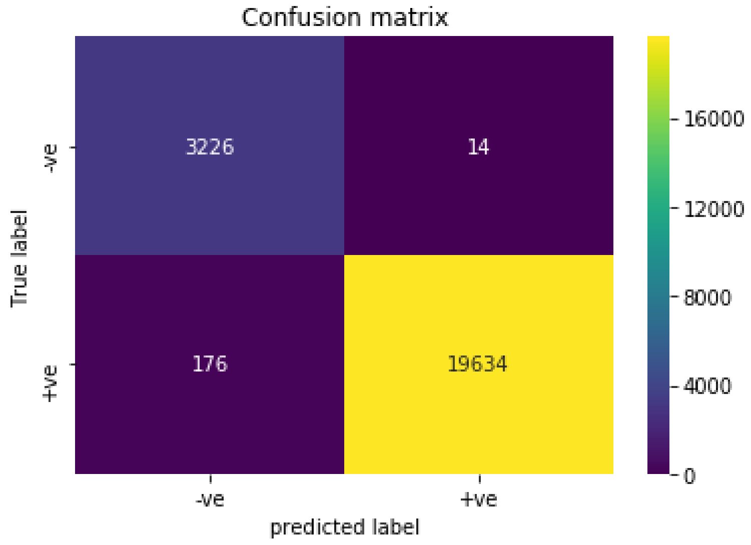
In [0]:

```
test(tfidf_w2v_train,tfidf_w2v_cv,tfidf_w2v_test,y_train_,y_cv_,y_test_
,16,64)
```

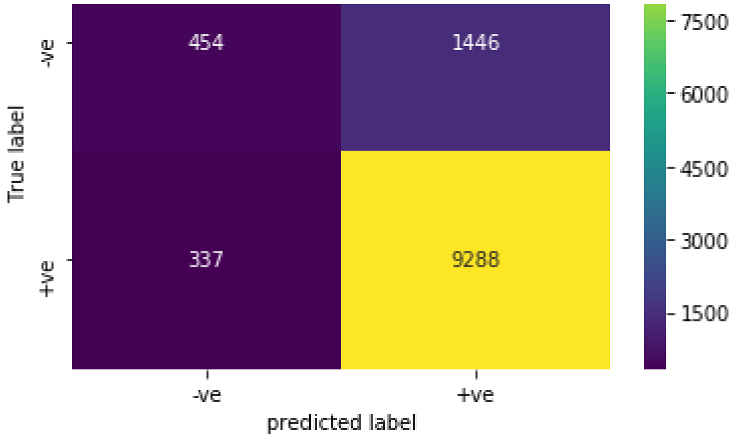
```
AUC Score: 0.802383021189337
```



```
Confusion matrix for train data
```



```
Confusion matrix for test data
```



[5.2] Applying GBDT using XGBOOST

In [0]:

```
pip install xgboost
```

```
Requirement already satisfied: xgboost in /usr/local/lib/python3.6/dist
-packages (0.82)
Requirement already satisfied: scipy in /usr/local/lib/python3.6/dist-p
ackages (from xgboost) (1.3.0)
Requirement already satisfied: numpy in /usr/local/lib/python3.6/dist-p
ackages (from xgboost) (1.16.3)
```

In [0]:

```
from xgboost import XGBClassifier
#function to find an optimal value of k,AUC,ROC,confusion matrix
def best_para_xgb(x_train,x_cv,x_test,y_train,y_cv,y_test):
        n_estimators = [1, 4, 16, 64, 100, 200]
        depth = [1, 2, 4, 8, 16, 32]
        param_grid = {'max_depth':[1, 2, 4, 8, 16, 32],'n_estimators':[
1, 4, 16, 64, 100, 200]}
        clf = XGBClassifier(booster='gbtree')
        grid_search = GridSearchCV(estimator = clf ,param_grid=param_gr
id ,cv = 3,n_jobs = -1,return_train_score=True)
        grid_search.fit(x_train, y_train,verbose=True)
        auc_cv = grid_search.cv_results_['mean_test_score'].reshape(6,6
)
        auc = grid_search.cv_results_['mean_train_score'].reshape(6,6)
        print('best depth = ',grid_search.best_estimator_ .max_depth)
        print('best n_estimators = ',grid_search.best_estimator_ .n_est
imators)



        # matrix for auc
        confusion_matr = pd.DataFrame(auc,index=n_estimators,columns=de
pth)
        sns.heatmap(confusion_matr,annot=True,cmap='viridis')
        plt.title("AUC matrix for train")
        plt.xlabel("depth")
        plt.ylabel("n_estimators")
        plt.show()

        confusion_matr_cv = pd.DataFrame(auc_cv,index=n_estimators,colu
mns=depth)
        sns.heatmap(confusion_matr_cv,annot=True,cmap='viridis')
        plt.title("AUC matrix for cv")
        plt.xlabel("depth")
        plt.ylabel("n_estimators")
        plt.show()

from scipy import sparse
def test_xgb(x_train,x_cv,x_test,y_train,y_cv,y_test,depth,n_estimators
 ):
        clf = XGBClassifier(max_depth = depth, n_estimators  = n_estima
tors , class_weight = 'balanced')
        clf.fit(x_train,y_train)
        prob_test = clf.predict_proba(x_test)
        prob_test = prob_test[:,1]

        prob_train = clf.predict_proba(x_train)
        prob_train = prob_train[:,1]
        print("AUC Score: {}".format(roc_auc_score(y_test,prob_test)))
        #ROC curve
        fpr_tr,tpr_tr,thres_tr = roc_curve(y_train,prob_train)
        fpr,tpr,thres = roc_curve(y_test,prob_test)
        plt.plot([0,0],[1,1],linestyle='--')
        plt.plot(fpr,tpr,'r',marker='.',label='test')
        plt.plot(fpr_tr,tpr_tr,'b',marker='.',label='train')
        plt.legend(loc='upper right')
        plt.title("ROC curve")
        plt.show()

        #confusion matrix fortrain and test
        print("Confusion matrix for train data")
        predict_tr = clf.predict(x_train)
        confu_metrix_(y_train,predict_tr)

        print("Confusion matrix for test data")
        predict_te = clf.predict(x_test)
        confu_metrix_(y_test,predict_te)
```

[5.2.1] Applying XGBOOST on BOW, SET 1

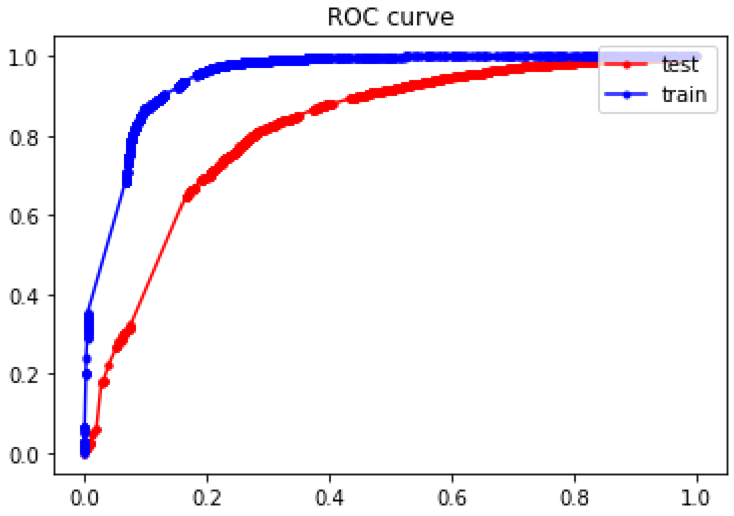
In [0]:

```
# Please write all the code with proper documentation
best_para_xgb(bow_train,bow_cv,bow_test,y_train,y_cv,y_test)
```

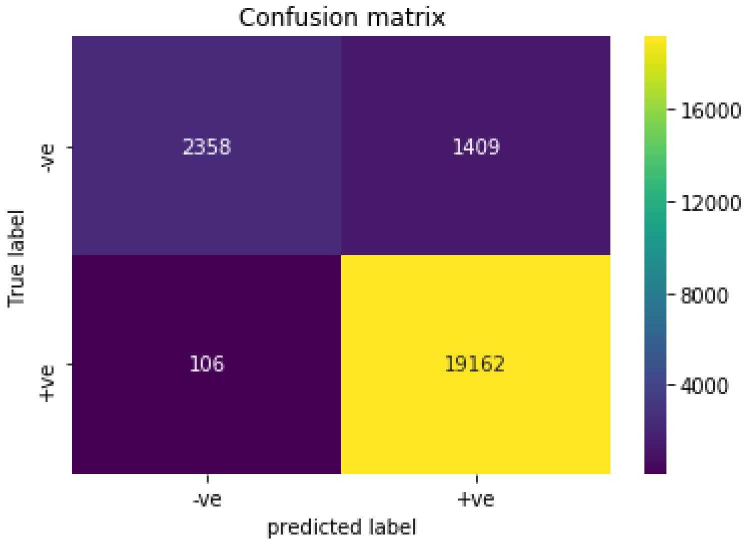
In [0]:

```
test_xgb(bow_train,bow_cv,bow_test,y_train,y_cv,y_test,32,4)
```

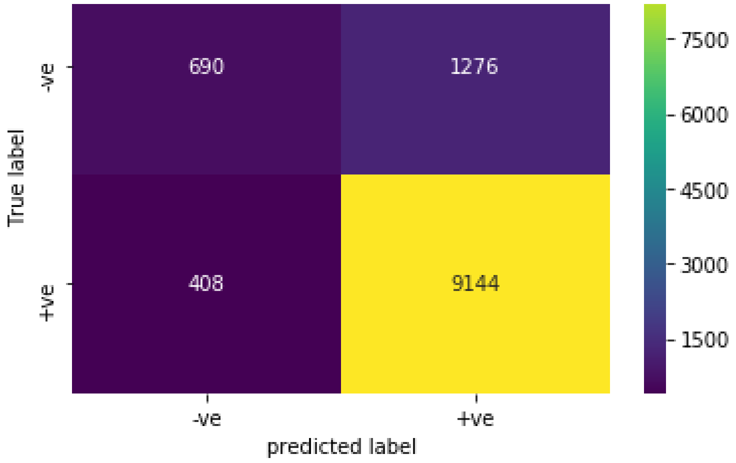
```
AUC Score: 0.8174851878926678
```



```
Confusion matrix for train data
```



```
Confusion matrix for test data
```



[5.2.2] Applying XGBOOST on TFIDF, SET 2

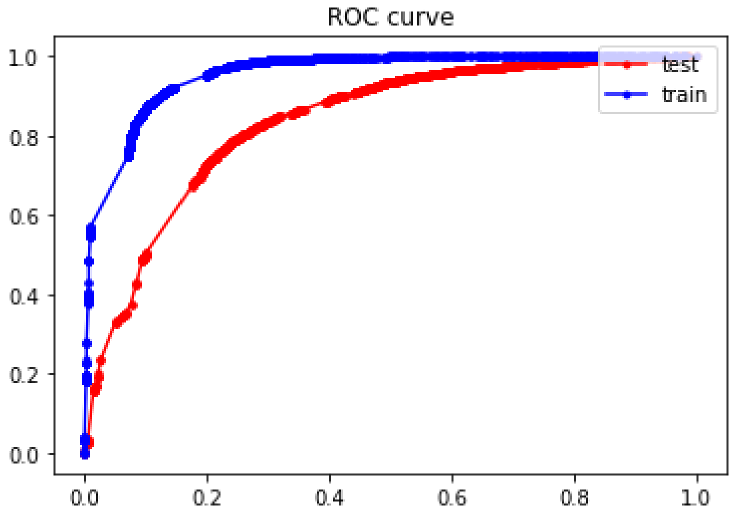
In [0]:

```
# Please write all the code with proper documentation
best_para_xgb(tfidf_train,tfidf_cv,tfidf_test,y_train,y_cv,y_test)
```

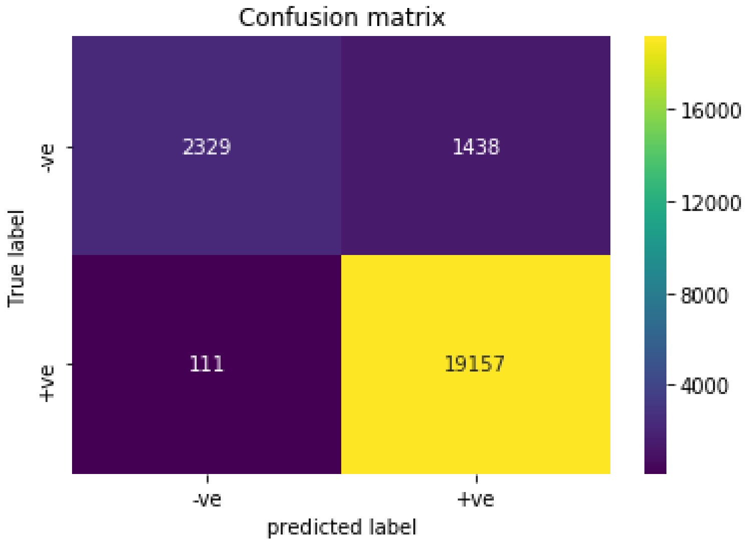
In [0]:

```
test_xgb(tfidf_train,tfidf_cv,tfidf_test,y_train,y_cv,y_test,32,4)
```

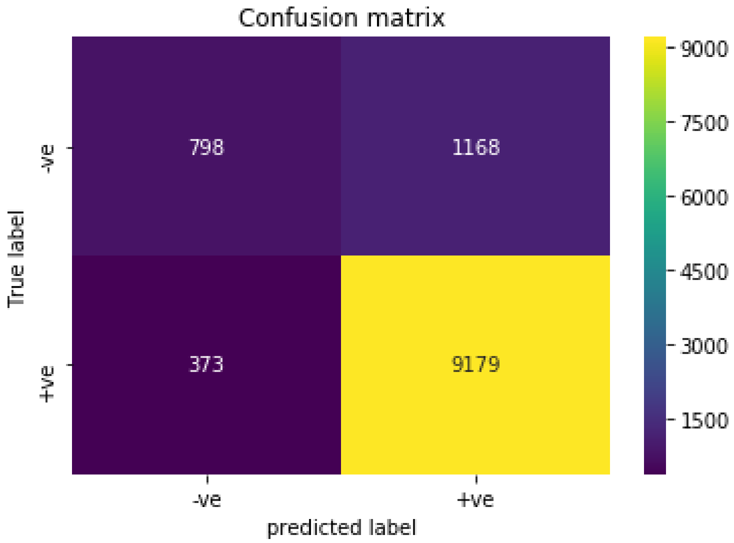
```
AUC Score: 0.8362670262553868
```



```
Confusion matrix for train data
```



```
Confusion matrix for test data
```



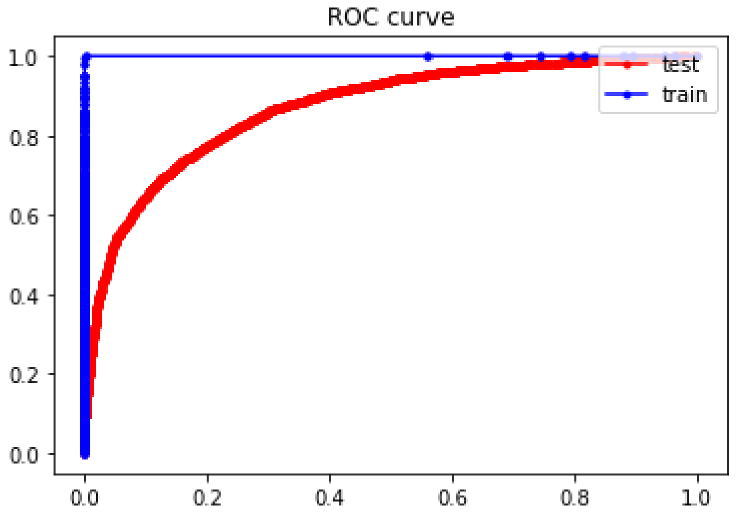
[5.2.3] Applying XGBOOST on AVG W2V, SET 3

```
In [0]: # Please write all the code with proper documentation
        best_para_xgb(avg_w2v_train,avg_w2v_cv,avg_w2v_test,y_train_,y_cv_,y_te
        st_)
```

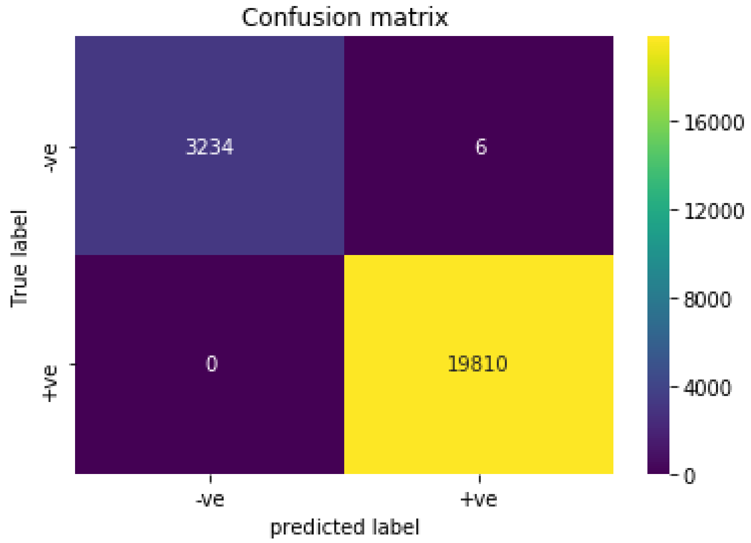
```
In [0]: avg_w2v_train,avg_w2v_cv,avg_w2v_test = sparse.csr_matrix(avg_w2v_train
        ),sparse.csr_matrix(avg_w2v_cv),sparse.csr_matrix(avg_w2v_test)
```

```
In [33]: test_xgb(avg_w2v_train,avg_w2v_cv,avg_w2v_test,y_train_,y_cv_,y_test_,3
         2,100)
```

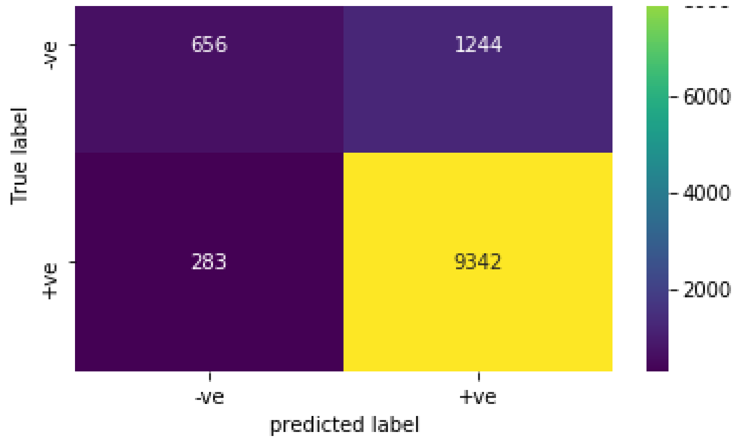
```
AUC Score: 0.8709808612440191
```



```
Confusion matrix for train data
```



```
Confusion matrix for test data
```



[5.2.4] Applying XGBOOST on TFIDF W2V, SET 4

In [0]:

```
# Please write all the code with proper documentation
best_para_xgb(tfidf_w2v_train,tfidf_w2v_cv,tfidf_w2v_test,y_train_,y_cv
_,y_test_)
```

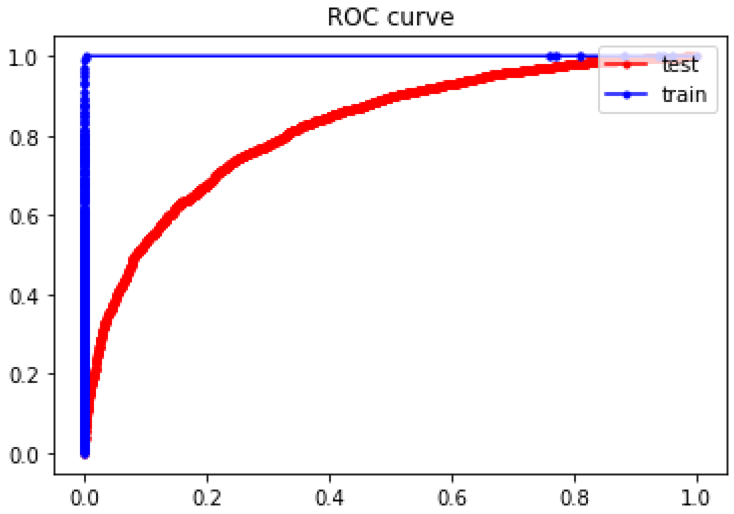
In [0]:

```
tfidf_w2v_train,tfidf_w2v_cv,tfidf_w2v_test = sparse.csr_matrix(tfidf_w
2v_train),sparse.csr_matrix(tfidf_w2v_cv),sparse.csr_matrix(tfidf_w2v_t
est)
```

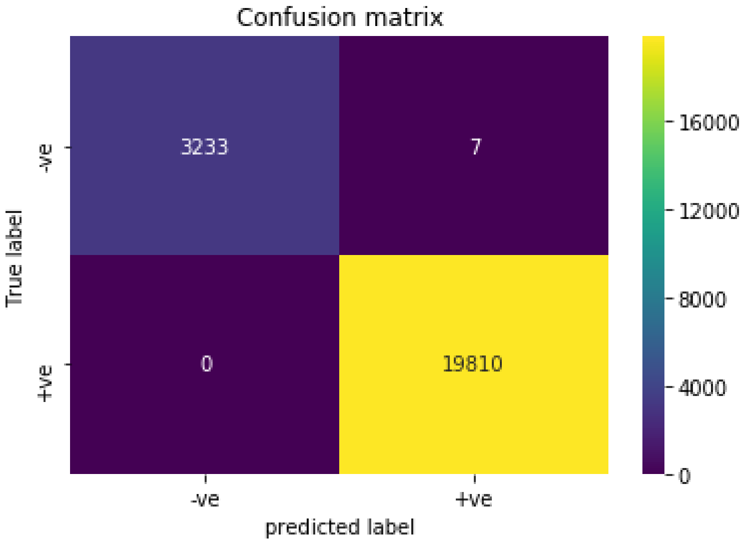
In [35]:

```
test_xgb(tfidf_w2v_train,tfidf_w2v_cv,tfidf_w2v_test,y_train_,y_cv_,y_t
est_,16,64)
```

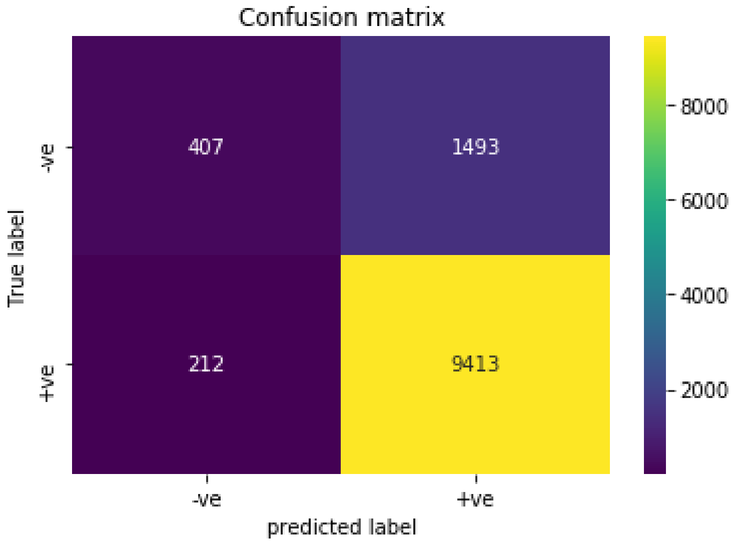
```
AUC Score: 0.8209905126452495
```



```
Confusion matrix for train data
```



```
Confusion matrix for test data
```



[6] Conclusions

In [36]:

```
# Please compare all your models using Prettytable library
df = pd.DataFrame({"Model":[" Random Forests on BOW", "Random Forests o
n TFIDF",

        "Random Forests on AVG W2V", "Random Forests on TFIDF W2V", "XG
BOOST on BOW","XGBOOST on TFIDF", "XGBOOST on AVG W2V", "XGBOOST on TFI
DF W2V"],
                   "Hyper parameter(Depth)":[32,32,32,16,32,32,32,16],
                                         "Hyper parameter(n_estimators
 )":[4,4,100,64,4,4,100,64],
                   "AUC":[0.7592166442163343, 0.7482731721936232, 0.857
815010252905, 0.802383021189337, 0.8174851878926678, 0.8362670262553868
, 0.8709808612440191, 0.8209905126452495]}
                  ,columns=["Model","Hyper parameter(Depth)","Hyper par
ameter(n_estimators )","AUC"])
df.sort_values(by="AUC",ascending=False)
```

Out[36]:

	Model	Hyper parameter(Depth)	Hyper parameter(n_estimators)	AUC
6	XGBOOST on AVG W2V	32	100	0.870981
2	Random Forests on AVG W2V	32	100	0.857815
5	XGBOOST on TFIDF	32	4	0.836267
7	XGBOOST on TFIDF W2V	16	64	0.820991
4	XGBOOST on BOW	32	4	0.817485
3	Random Forests on TFIDF W2V	16	64	0.802383
0	Random Forests on BOW	32	4	0.759217
1	Random Forests on TFIDF	32	4	0.748273

From above results we can say that avg_w2v models performes better than other models. and XGBoost models performs slitly better than random forest models.