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 [1]:

```
%matplotlib inline
import warnings
warnings.filterwarnings("ignore")


import sqlite3
import pandas as pd
import numpy as np
import nltk
import string
import matplotlib.pyplot as plt
```

```
import seaborn as sns
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.feature_extraction.text import TfidfVectorizer

from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics import confusion_matrix
from sklearn import metrics
from sklearn.metrics import roc_curve, auc
from nltk.stem.porter import PorterStemmer

import re
# Tutorial about Python regular expressions: https://pymotw.com/2/re/
import string
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from nltk.stem.wordnet import WordNetLemmatizer

from gensim.models import Word2Vec
from gensim.models import KeyedVectors
import pickle

from tqdm import tqdm
import os
```

In [2]:

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

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

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

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

# Give reviews with Score>3 a positive rating(1), and reviews with a score<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[2]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Tim
0	150524	0006641040	ACITT7DI6IDDL	shari zychinski	0	0	1	93934080
1	150501	0006641040	AJ46FKXOVC7NR	Nicholas A Mesiano	2	2	1	94080960

2	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Tim
	451856	B00004CXX9	AIUWLEQ1ADEG5	Elizabeth Medina	0	0	1	94409280

In [3]:

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

In [4]:

```
print(display.shape)
display.head()
```

```
(80668, 7)
```

Out[4]:

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

In [5]:

```
display[display['UserId']=='AZY10LLTJ71NX']
```

Out[5]:

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

In [6]:

```
display['COUNT(*)'].sum()
```

Out[6]:

```
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

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

In [7]:

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

Out[7]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Ti
0	78445	B000HDL1RQ	AR5J8UI46CURR	Geetha Krishnan	2	2	5	1199577(
1	138317	B000HDOPYC	AR5J8UI46CURR	Geetha Krishnan	2	2	5	1199577(
2	138277	B000HDOPYM	AR5J8UI46CURR	Geetha Krishnan	2	2	5	1199577(
3	73791	B000HDOPZG	AR5J8UI46CURR	Geetha Krishnan	2	2	5	1199577(
4	155049	B000PAQ75C	AR5J8UI46CURR	Geetha Krishnan	2	2	5	1199577(

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

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

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

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

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

In [8]:

```
#Sorting data according to ProductId in ascending order
sorted_data=filtered_data.sort_values('ProductId', axis=0, ascending=True, inplace=False, kind='qui
cksort', na_position='last')
```

In [9]:

```
#Deduplication of entries
final=sorted_data.drop_duplicates(subset={"UserId","ProfileName","Time","Text"}, keep='first', inpl
ace=False)
final.shape
```

Out[9]:

```
(364173, 10)
```

In [10]:

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

Out[10]:

```
69.25890143662969
```

Observation:- It was also seen that in two rows given below the value of HelpfulnessNumerator is greater than HelpfulnessDenominator which is not practically possible hence these two rows too are removed from calcualtions

In [11]:

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

display.head()
```

Out[11]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Ti
0	64422	B000MIDROQ	A161DK06JJMCYF	J. E. Stephens "Jeanne"	3	1	5	1224892
1	44737	B001EQ55RW	A2V0I904FH7ABY	Ram	3	2	4	1212883

In [12]:

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

In [13]:

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

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

```
(364171, 10)
```

Out[13]:

```
1    307061
0     57110
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 [14]:

```
# printing some random reviews
sent_0 = final['Text'].values[0]
print(sent_0)
print("="*50)

sent_1000 = final['Text'].values[1000]
print(sent_1000)
print("="*50)

sent_1500 = final['Text'].values[1500]
print(sent_1500)
print("="*50)

sent_4900 = final['Text'].values[4900]
print(sent_4900)
print("="*50)
```

```
this witty little book makes my son laugh at loud. i recite it in the car as we're driving along a
nd he always can sing the refrain. he's learned about whales, India, drooping roses:  i love all t
he new words this book  introduces and the silliness of it all.  this is a classic book i am
willing to bet my son will STILL be able to recite from memory when he is  in college
==================================================
I was really looking forward to these pods based on the reviews.  Starbucks is good, but I prefer
bolder taste.... imagine my surprise when I ordered 2 boxes - both were expired! One expired back
in 2005 for gosh sakes.  I admit that Amazon agreed to credit me for cost plus part of shipping, b
ut geez, 2 years expired!!!  I'm hoping to find local San Diego area shoppe that carries pods so t
hat I can try something different than starbucks.
==================================================
Great ingredients although, chicken should have been 1st rather than chicken broth, the only thing
I do not think belongs in it is Canola oil. Canola or rapeseed is not someting a dog would ever fi
nd in nature and if it did find rapeseed in nature and eat it, it would poison them. Today's Food
industries have convinced the masses that Canola oil is a safe and even better oil than olive or v
irgin coconut, facts though say otherwise. Until the late 70's it was poisonous until they figured
out a way to fix that. I still like it but it could be better.
==================================================
Originally found the syrup for sale in a tourist shop in North Georgia mountains. Could not find i
t locally and have been ordering from manufacturer ever since. I am diabetic and really think this
syrup is better than the 'full blown' stuff!
==================================================
```

In [15]:

```
# remove urls from text python: https://stackoverflow.com/a/40823105/4084039
sent_0 = re.sub(r"http\S+", "", sent_0)
sent_1000 = re.sub(r"http\S+", "", sent_1000)
```

```
sent_150 = re.sub(r"http\S+", "", sent_1500)
sent_4900 = re.sub(r"http\S+", "", sent_4900)

print(sent_0)
```

```
this witty little book makes my son laugh at loud. i recite it in the car as we're driving along a
nd he always can sing the refrain. he's learned about whales, India, drooping roses:  i love all t
he new words this book  introduces and the silliness of it all.  this is a classic book i am
willing to bet my son will STILL be able to recite from memory when he is  in college
```

In [16]:

```
# 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)
```

```
this witty little book makes my son laugh at loud. i recite it in the car as we're driving along a
nd he always can sing the refrain. he's learned about whales, India, drooping roses:  i love all t
he new words this book  introduces and the silliness of it all.  this is a classic book i am
willing to bet my son will STILL be able to recite from memory when he is  in college
==================================================
I was really looking forward to these pods based on the reviews.  Starbucks is good, but I prefer
bolder taste.... imagine my surprise when I ordered 2 boxes - both were expired! One expired back
in 2005 for gosh sakes.  I admit that Amazon agreed to credit me for cost plus part of shipping, b
ut geez, 2 years expired!!!  I'm hoping to find local San Diego area shoppe that carries pods so t
hat I can try something different than starbucks.
==================================================
Great ingredients although, chicken should have been 1st rather than chicken broth, the only thing
I do not think belongs in it is Canola oil. Canola or rapeseed is not someting a dog would ever fi
nd in nature and if it did find rapeseed in nature and eat it, it would poison them. Today's Food
industries have convinced the masses that Canola oil is a safe and even better oil than olive or v
irgin coconut, facts though say otherwise. Until the late 70's it was poisonous until they figured
out a way to fix that. I still like it but it could be better.
==================================================
Originally found the syrup for sale in a tourist shop in North Georgia mountains. Could not find i
t locally and have been ordering from manufacturer ever since. I am diabetic and really think this
syrup is better than the 'full blown' stuff!
```

In [17]:

```
# https://stackoverflow.com/a/47091490/4084039
import re

def decontracted(phrase):
    # specific
    phrase = re.sub(r"won't", "will not", phrase)
    phrase = re.sub(r"can\'t", "can not", phrase)

    # general
    phrase = re.sub(r"n\'t", " not", phrase)
    phrase = re.sub(r"\'re", " are", phrase)
    phrase = re.sub(r"\'s", " is", phrase)
    phrase = re.sub(r"\'d", " would", phrase)
    phrase = re.sub(r"\'ll", " will", phrase)
    phrase = re.sub(r"\'t", " not", phrase)
```

```
    phrase = re.sub(r"\'ve", " have", phrase)
    phrase = re.sub(r"\'m", " am", phrase)
    return phrase
```

In [18]:

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

```
Great ingredients although, chicken should have been 1st rather than chicken broth, the only thing
I do not think belongs in it is Canola oil. Canola or rapeseed is not someting a dog would ever fi
nd in nature and if it did find rapeseed in nature and eat it, it would poison them. Today is Food
industries have convinced the masses that Canola oil is a safe and even better oil than olive or v
irgin coconut, facts though say otherwise. Until the late 70 is it was poisonous until they
figured out a way to fix that. I still like it but it could be better.
==================================================
```

In [19]:

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

```
this witty little book makes my son laugh at loud. i recite it in the car as we're driving along a
nd he always can sing the refrain. he's learned about whales, India, drooping roses:  i love all t
he new words this book  introduces and the silliness of it all.  this is a classic book i am
willing to bet my son will STILL be able to recite from memory when he is  in college
```

In [20]:

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

```
Great ingredients although chicken should have been 1st rather than chicken broth the only thing I
do not think belongs in it is Canola oil Canola or rapeseed is not someting a dog would ever find
in nature and if it did find rapeseed in nature and eat it it would poison them Today is Food indu
stries have convinced the masses that Canola oil is a safe and even better oil than olive or virgi
n coconut facts though say otherwise Until the late 70 is it was poisonous until they figured out
a way to fix that I still like it but it could be better
```

In [21]:

```
# https://gist.github.com/sebleier/554280
# we are removing the words from the stop words list: 'no', 'nor', 'not'
# <br /><br /> ==> after the above steps, we are getting "br br"
# we are including them into stop words list
# instead of <br /> if we have <br/> these tags would have revmoved in the 1st step

stopwords= set(['br', 'the', 'i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "y
ou're", "you've",\
            "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his',
'himself', \
            'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them',
'their',\
            'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll",
'these', 'those', \
            'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', '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', 'how', 'all', 'any', 'both', 'e
ach', 'few', 'more',\
            'most', 'other', 'some', 'such', 'only', 'own', 'same', 'so', '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', "do
esn't", 'hadn',\
            "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn'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 [22]:

```
# 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%|████████████████████████████████████████████████████████████████████████| 364171/364171
[03:37<00:00, 1676.28it/s]
```

In [23]:

```
def word_count(x):
    return len(x.split(' '))

preprocessed_reviews_len = list(map(word_count,preprocessed_reviews))
print(preprocessed_reviews_len[:10])
```

```
[35, 44, 23, 47, 53, 16, 17, 26, 53, 15]
```

In [24]:

```
from sklearn import preprocessing

preprocessed_reviews_len_norm = preprocessing.normalize([np.array(preprocessed_reviews_len)])
```

In [25]:

```
print(preprocessed_reviews_len_norm[:10])
```

```
[[0.00106049 0.00133318 0.00069689 ... 0.00042419 0.00057569 0.00087869]]
```

In [26]:

```
data = preprocessed_reviews[:100000]
scores = final['Score'][:100000]
data_reviews_len = preprocessed_reviews_len_norm.reshape(-1,1)[:100000]

print(final['Score'][:100000].value_counts())
```

```
1    85198
0    14802
Name: Score, dtype: int64
```

In [27]:

```
from sklearn.model_selection import train_test_split

data_train,data_test,scores_train,scores_test = train_test_split(data,scores,test_size = 0.2,shuffl
e = False)
data_train_reviews_len,data_test_reviews_len,scores_train,scores_test =
train_test_split(data_reviews_len,scores,test_size = 0.2,shuffle = False)
```

```
print(len(data_train))
print(len(data_test))
print(len(data_train_reviews_len))
print(len(data_test_reviews_len))
print(len(scores_train))
print(len(scores_test))
```

```
80000
20000
80000
20000
80000
20000
```

In [28]:

```
print(scores_train.value_counts())
```

```
1    68365
0    11635
Name: Score, dtype: int64
```

In [29]:

```
dataset = {'review':data_train,'len':list(data_train_reviews_len),'score':scores_train}
data = pd.DataFrame(data = dataset)
print(data['score'].value_counts())
```

```
1    68365
0    11635
Name: score, dtype: int64
```

In [30]:

```
from sklearn.utils import resample

# Duplicating the minority data points
data_pos = data[data.score == 1]
data_neg = data[data.score == 0]
# Upsample minority class
data_neg_upsampled = resample(data_neg,replace=True,n_samples=len(data_pos),random_state=123)

# Combine majority class with upsampled minority class
data_upsampled = pd.concat([data_pos, data_neg_upsampled])

# Display new class counts
data_upsampled.score.value_counts()
```

Out[30]:

```
1    68365
0    68365
Name: score, dtype: int64
```

In [31]:

```
data_train = data_upsampled['review']
data_train_reviews_len = data_upsampled['len']
scores_train = data_upsampled['score']

print(len(data_train))
print(len(data_train_reviews_len))
print(len(scores_train))
```

```
136730
136730
136730
```

[3.2] Preprocessing Review Summary

In [32]:

```
## Similartly you can do preprocessing for review summary also.
```

[4] Featurization

[4.1] BAG OF WORDS

In [33]:

```
#BoW
count_vect = CountVectorizer(max_features = 5000) #in scikit-learn
data_train_bow = count_vect.fit_transform(data_train)
data_test_bow = count_vect.transform(data_test)

print(data_train_bow.shape)
print(data_test_bow.shape)
```

```
(136730, 5000)
(20000, 5000)
```

[4.2] Bi-Grams and n-Grams.

In [30]:

```
#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_s
hape()[1])
```

```
the type of count vectorizer  <class 'scipy.sparse.csr.csr_matrix'>
the shape of out text BOW vectorizer  (364171, 5000)
the number of unique words including both unigrams and bigrams  5000
```

[4.3] TF-IDF

In [34]:

```
tf_idf_vect = TfidfVectorizer(ngram_range=(1,2), min_df=10)
data_train_tfidf = tf_idf_vect.fit_transform(data_train)
data_test_tfidf = tf_idf_vect.transform(data_test)

print(data_train_tfidf.shape)
print(data_test_tfidf.shape)
```

```
(136730, 111853)
(20000, 111853)
```

[4.4] Word2Vec

In [ ]:

```
# 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 [ ]:

```
# 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=Tr
ue)
        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 ")
```

In [ ]:

```
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])
```

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

[4.4.1.1] Avg W2v

In [ ]:

```
# 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, 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 sentence/review
    for word in sent: # for each word in a review/sentence
        if word in w2v_words:
            vec = w2v_model.wv[word]
            sent_vec += vec
            cnt_words += 1
```

```
    if cnt_words != 0:
        sent_vec /= cnt_words
    sent_vectors.append(sent_vec)
print(len(sent_vectors))
print(len(sent_vectors[0]))
```

[4.4.1.2] TFIDF weighted W2v

In [ ]:

```
# 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 value
dictionary = dict(zip(model.get_feature_names(), list(model.idf_)))
```

In [ ]:

```
# TF-IDF weighted Word2Vec
tfidf_feat = model.get_feature_names() # tfidf words/col-names
# final_tf_idf is the sparse matrix with row= sentence, col=word and cell_val = tfidf

tfidf_sent_vectors = []; # the tfidf-w2v for each sentence/review is stored 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/review
    for word in sent: # for each word in a review/sentence
        if word in w2v_words and word in tfidf_feat:
            vec = w2v_model.wv[word]
#             tf_idf = tf_idf_matrix[row, tfidf_feat.index(word)]
            # to reduce the computation we are
            # dictionary[word] = idf value of word in whole courpus
            # sent.count(word) = tf valeus of word in this review
            tf_idf = dictionary[word]*(sent.count(word)/len(sent))
            sent_vec += (vec * tf_idf)
            weight_sum += tf_idf
    if weight_sum != 0:
        sent_vec /= weight_sum
    tfidf_sent_vectors.append(sent_vec)
    row += 1
```

[5] Assignment 4: Apply Naive Bayes

1. **Apply Multinomial NaiveBayes 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)

2. **The hyper paramter tuning(find best Alpha)**

   - Find the best hyper parameter which will give the maximum AUC value
   - Consider a wide range of alpha values for hyperparameter tuning, start as low as 0.00001
   - 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**

   - Find the top 10 features of positive class and top 10 features of negative class for both feature sets Set 1 and Set 2 using values of `feature_log_prob_` parameter of MultinomialNB and print their corresponding feature names

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. Here on X-axis you will have alpha values, since they have a wide range, just to represent those alpha values on the graph, apply log function on those alpha values.
    - 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.

Applying Multinomial Naive Bayes

In [35]:

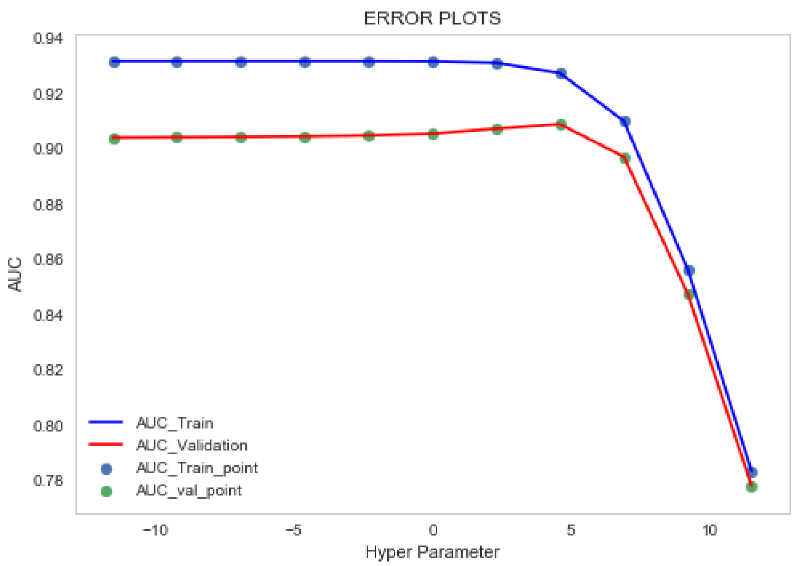
```
from sklearn.naive_bayes import MultinomialNB
from sklearn.model_selection import GridSearchCV

def get_AUC(X_train,y_train,dict_alpha_range):
    """This function apply the knn model with given algorithm
       on train and cv data and return AUC values for train and cross validation"""
    MNB = MultinomialNB()
    clf = GridSearchCV(MNB,dict_alpha_range,cv=10,scoring='roc_auc')
    clf.fit(X_train,y_train)

    train_auc = clf.cv_results_['mean_train_score']
    cv_auc  = clf.cv_results_['mean_test_score']

    return train_auc,cv_auc

def plot_AUC_Curves(auc_train,auc_cv,alpha_range):
    """This function plots the auc curves for the given auc values and k_range"""
    sns.set_style("whitegrid",{'axes.grid' : False})
    plt.plot(np.log(alpha_range),auc_train,"b-",label = "AUC_Train")
    plt.plot(np.log(alpha_range),auc_cv,"r-",label = "AUC_Validation")
    plt.scatter(np.log(alpha_range),auc_train,label = "AUC_Train_point")
    plt.scatter(np.log(alpha_range),auc_cv,label = "AUC_val_point")
    plt.legend()
    plt.xlabel("Hyper Parameter")
    plt.ylabel("AUC")
    plt.title("ERROR PLOTS")
    plt.show()

def apply_roc_curve(X_train,y_train,X_test,y_test,optimal_alpha):
    """This function apply knn model on train and predict labels for test data
      and also find FPR and TPR for train and test data.
      Returns the predicted labels,FPR and TPR values"""
    MNB = MultinomialNB(alpha = optimal_alpha,class_prior = [0.5,0.5])
    MNB.fit(X_train,y_train)
    prob_train = MNB.predict_proba(X_train)
    fpr_train, tpr_train, threshold = roc_curve(y_train, prob_train[:, 1])
    prob_test = MNB.predict_proba(X_test)
    fpr_test, tpr_test, threshold = roc_curve(y_test, prob_test[:, 1])

    # predict the class labels
    pred_train = MNB.predict(X_train)
    pred_test = MNB.predict(X_test)
```

```
    feature_log_prob = MNB.feature_log_prob_

    return fpr_train,tpr_train,fpr_test,tpr_test,pred_train,pred_test,feature_log_prob

def plot_roc_curve(fpr_train,tpr_train,fpr_test,tpr_test):
    """This function plot the roc curves for train and test data"""
    # plot ROC curves for train and test data
    plt.plot(fpr_train,tpr_train,"g-",label = "AUC_Train : "+str(auc(fpr_train, tpr_train)))
    plt.plot(fpr_test,tpr_test,"r-",label = "AUC_Test : "+str(auc(fpr_test, tpr_test)))
    plt.plot([0,1],[0,1],"b-")
    plt.legend(loc="lower right")
    plt.xlabel("False Positive Rate")
    plt.ylabel("True Positive Rate")
    plt.show()

def plot_Confusion_Matrix(actual_labels,predict_labels,title):
    """This function plot the confusion matrix"""
    # Reference : https://seaborn.pydata.org/generated/seaborn.heatmap.html
    cm = confusion_matrix(actual_labels, predict_labels)
    classNames = ['NO','YES']
    cm_data = pd.DataFrame(cm,index = classNames,
                  columns = classNames)
    plt.figure(figsize = (5,4))
    sns.heatmap(cm_data, annot=True,fmt="d")
    plt.title(title)
    plt.ylabel('Actual label')
    plt.xlabel('Predicted label')
    plt.show()
```

[5.1] Applying Naive Bayes on BOW, SET 1

In [36]:

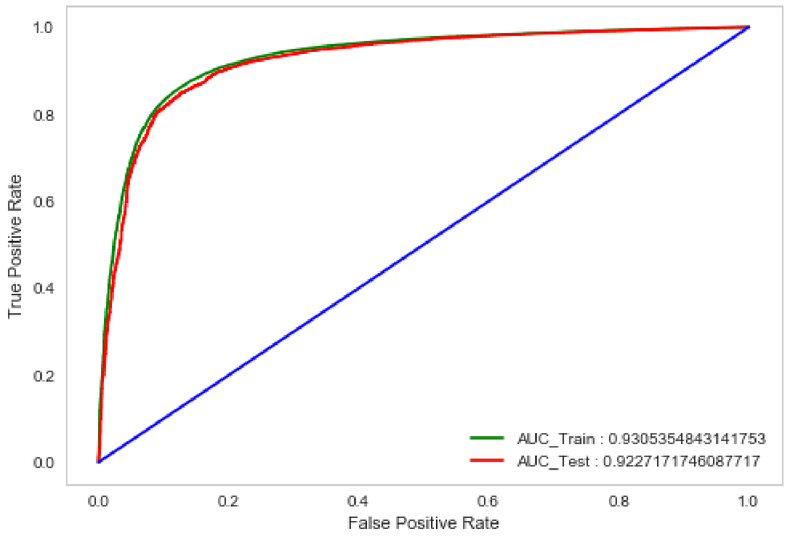
```
# Please write all the code with proper documentation
# dictionary alpha
hyper_parameter = {"alpha" : [10**-5,10**-4,10**-3,10**-2,10**-1,10**0,10**1,10**2,10**3,10**4,10**
5]}
# AUC
train_auc,cv_auc = get_AUC(data_train_bow,scores_train,hyper_parameter)

# alpha range
alpha = hyper_parameter.get("alpha")
# plot auc
plot_AUC_Curves(train_auc,cv_auc,alpha)
```

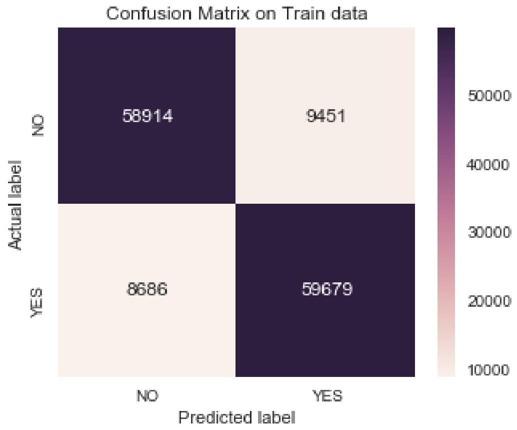


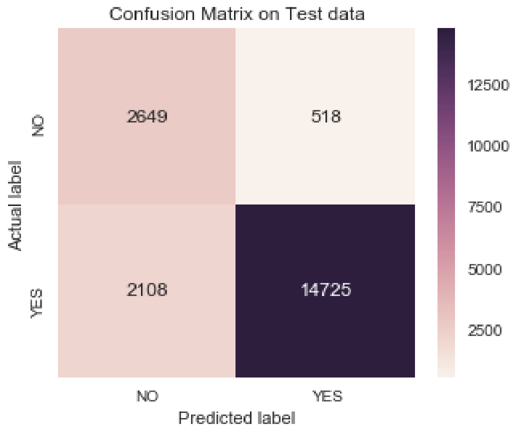
In [37]:

```
# optimal alpha
#optimal_alpha_bow = alpha[cv_auc.tolist().index(max(cv_auc.tolist()))]
optimal_alpha_bow = 0.01
# roc
```

```
# roc
fpr_train,tpr_train,fpr_test,tpr_test,pred_train,pred_test,feature_log_prob =
apply_roc_curve(data_train_bow,scores_train,data_test_bow,scores_test,optimal_alpha_bow)
# auc
auc_bow = auc(fpr_test,tpr_test)
# plot roc
plot_roc_curve(fpr_train,tpr_train,fpr_test,tpr_test)
# plot confusion matrix
plot_Confusion_Matrix(scores_train,pred_train,"Confusion Matrix on Train data")
plot_Confusion_Matrix(scores_test,pred_test,"Confusion Matrix on Test data")
print(optimal_alpha_bow)
```








```
0.01
```

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

In [38]:

```
# Please write all the code with proper documentation
possitive_class_prob_sorted = feature_log_prob[1,:].argsort()[::-1]
rank = np.array(range(1,11))
top10_positive_features =
np.take(np.array(count_vect.get_feature_names()),possitive_class_prob_sorted[:10])
prob = np.take(feature_log_prob[1,:], possitive_class_prob_sorted[:10])
top10_positive_features_details = pd.DataFrame(data = {'Rank' : rank,'Feature' : top10_positive_fea
tures,'Probability' : prob})
print(top10_positive_features_details)
```

```
   Feature  Probability  Rank
0      not    -3.681410     1
1      tea    -4.543857     2
2     like    -4.569027     3
3     good    -4.642667     4
4    great    -4.648669     5
5      one    -4.828825     6
6    taste    -4.972970     7
7     love    -5.022142     8
8  product    -5.026688     9
9   flavor    -5.109130    10
```

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

In [39]:

```
# Please write all the code with proper documentation
negative_class_prob_sorted = feature_log_prob[0,:].argsort()[::-1]
rank = np.array(range(1,11))
top10_negative_features = np.take(count_vect.get_feature_names(),negative_class_prob_sorted[:10])
prob = np.take(feature_log_prob[1,:], negative_class_prob_sorted[:10])
top10_negative_features_details = pd.DataFrame(data = {'Rank' : rank,'Feature' : top10_negative_fea
tures,'Probability' : prob})
print(top10_negative_features_details)
```

```
   Feature  Probability  Rank
0      not    -3.681410     1
1     like    -4.569027     2
2  product    -5.026688     3
3    would    -5.116650     4
4    taste    -4.972970     5
5      one    -4.828825     6
6       no    -5.346240     7
7      tea    -4.543857     8
8     good    -4.642667     9
9     food    -5.444529    10
```

In [40]:

```
print(len(data_test_reviews_len))
```

```
20000
```

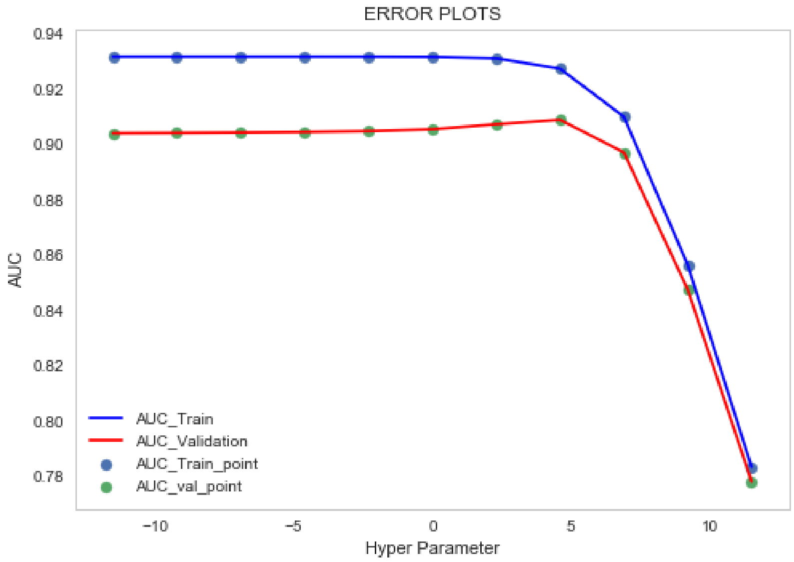
In [41]:

```
# Feature Enginerring
# Taken review length as new feature
from scipy.sparse import hstack,coo_matrix
data_train_reviews_len_bow =
hstack([data_train_bow,coo_matrix(np.array(data_train_reviews_len.tolist())[:None])])
data_test_reviews_len_bow = hstack([data_test_bow,coo_matrix(np.array(data_test_reviews_len)[:None
])])

print(data_train_reviews_len_bow.shape)
print(data_test_reviews_len_bow.shape)
```

```
(136730, 5001)
(20000, 5001)
```

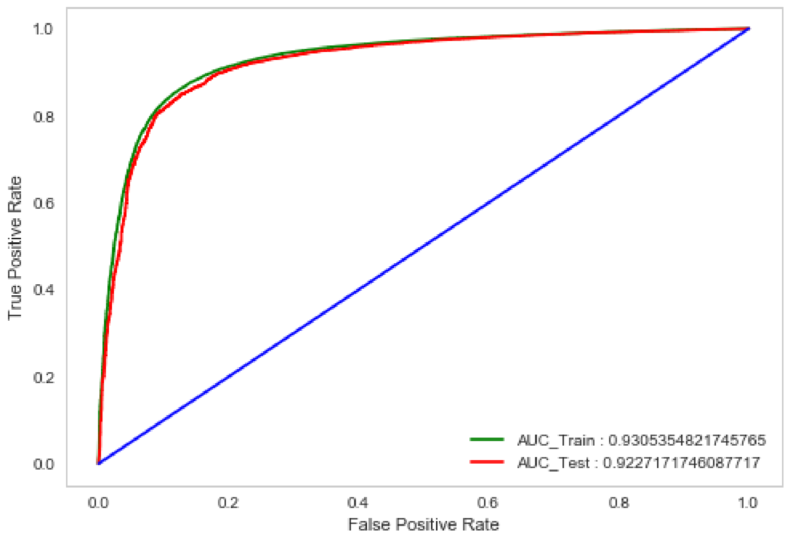
In [42]:

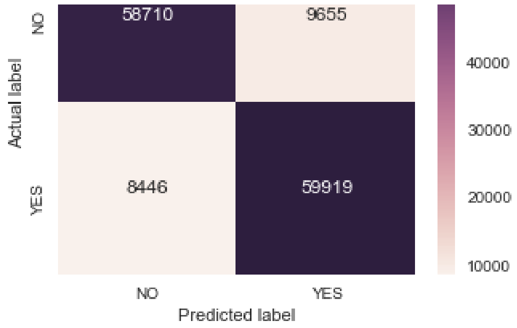
```
# Please write all the code with proper documentation
# alpha
hyper_parameter = {"alpha" : [10**-5,10**-4,10**-3,10**-2,10**-1,10**0,10**1,10**2,10**3,10**4,10**
5]}
# auc
train_auc,cv_auc = get_AUC(data_train_reviews_len_bow,scores_train,hyper_parameter)
alpha = hyper_parameter.get("alpha")
# plot auc
plot_AUC_Curves(train_auc,cv_auc,alpha)
```

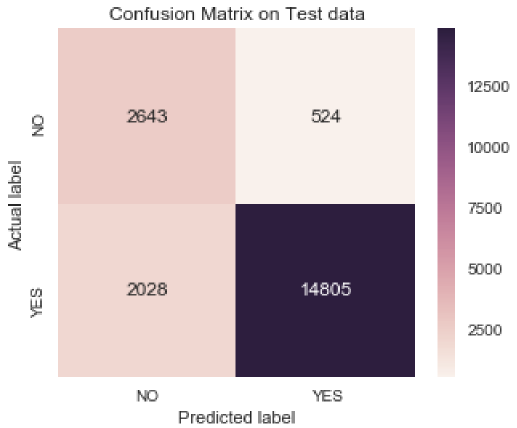


In [43]:

```
#optimal_alpha_bow_FE = alpha[cv_auc.tolist().index(max(cv_auc.tolist()))]
optimal_alpha_bow_FE = 0.01
#roc
fpr_train,tpr_train,fpr_test,tpr_test,pred_train,pred_test,feature_log_prob =
apply_roc_curve(data_train_reviews_len_bow,scores_train,data_test_reviews_len_bow,scores_test,opti
mal_alpha_bow_FE)
auc_bow_FE = auc(fpr_test,tpr_test)
plot_roc_curve(fpr_train,tpr_train,fpr_test,tpr_test)
# confusion matrix
plot_Confusion_Matrix(scores_train,pred_train,"Confusion Matrix on Train data")
plot_Confusion_Matrix(scores_test,pred_test,"Confusion Matrix on Test data")
print(optimal_alpha_bow_FE)
```





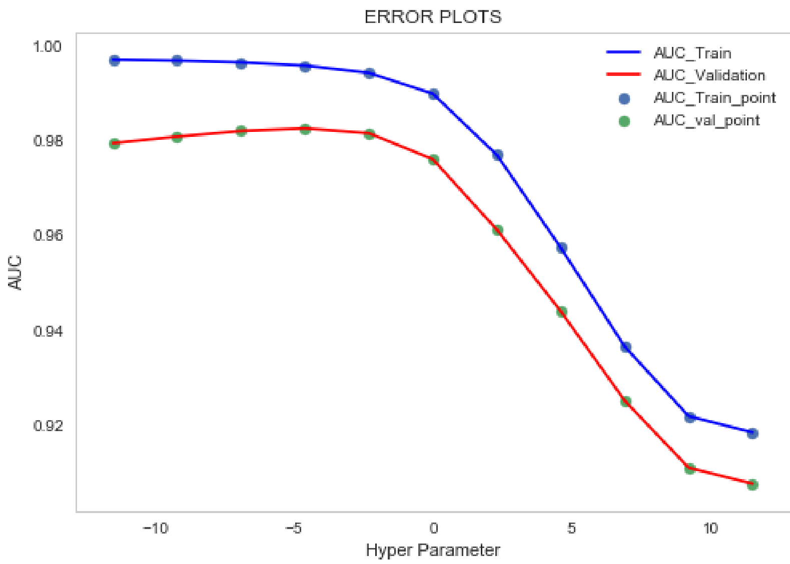



```
0.01
```

[5.2] Applying Naive Bayes on TFIDF, SET 2

In [44]:

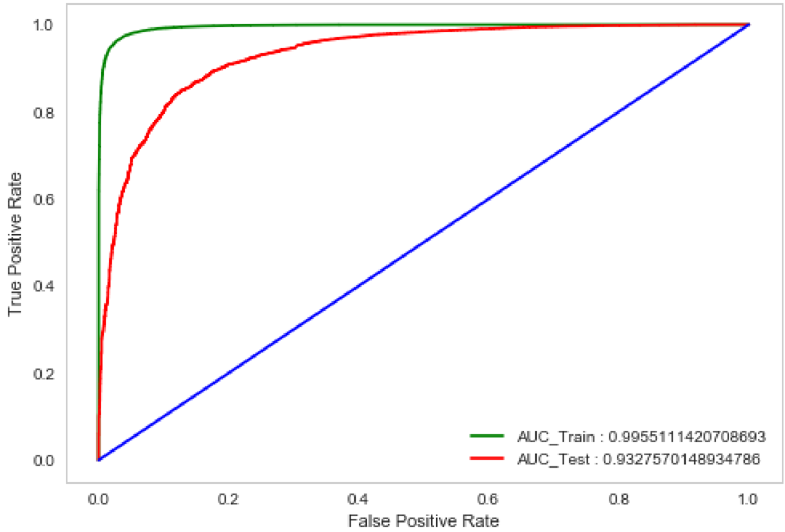
```
# Please write all the code with proper documentation
# dict alpha
hyper_parameter = {"alpha" : [10**-5,10**-4,10**-3,10**-2,10**-1,10**0,10**1,10**2,10**3,10**4,10**
5]}
# AUC
train_auc,cv_auc = get_AUC(data_train_tfidf,scores_train,hyper_parameter)
# alpha range
alpha = hyper_parameter.get("alpha")
# plot auc
plot_AUC_Curves(train_auc,cv_auc,alpha)
```

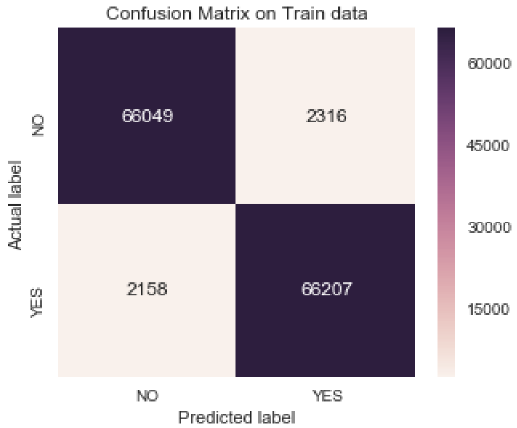


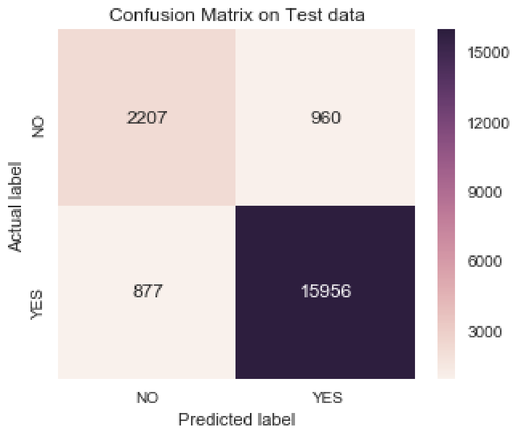
In [45]:

```
# optimal alph
```

```
# optimal alph
optimal_alpha_tf_idf = alpha[cv_auc.tolist().index(max(cv_auc.tolist()))]
# roc
fpr_train,tpr_train,fpr_test,tpr_test,pred_train,pred_test,feature_log_prob =
apply_roc_curve(data_train_tfidf,scores_train,data_test_tfidf,scores_test,optimal_alpha_tf_idf)
# auc
auc_TF_IDF = auc(fpr_test,tpr_test)
# plot confusion matrix
plot_roc_curve(fpr_train,tpr_train,fpr_test,tpr_test)
plot_Confusion_Matrix(scores_train,pred_train,"Confusion Matrix on Train data")
plot_Confusion_Matrix(scores_test,pred_test,"Confusion Matrix on Test data")
```







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

In [46]:

```
# Please write all the code with proper documentation
possitive_class_prob_sorted = feature_log_prob[1,:].argsort()[::-1]
```

```
rank = np.array(range(1,11))
top10_positive_features = np.take(tf_idf_vect.get_feature_names(),possitive_class_prob_sorted[:10]
)
prob = np.take(feature_log_prob[1,:], possitive_class_prob_sorted[:10])
top10_positive_features_details = pd.DataFrame(data = {'Rank' : rank,'Feature' : top10_negative_fea
tures,'Probability' : prob})
print(top10_positive_features_details)
```

```
   Feature  Probability  Rank
0      not    -5.524384     1
1     like    -5.611215     2
2  product    -5.639279     3
3    would    -5.812261     4
4    taste    -5.935018     5
5      one    -5.986849     6
6       no    -6.143524     7
7      tea    -6.164836     8
8     good    -6.194363     9
9     food    -6.200062    10
```

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

In [47]:

```
# Please write all the code with proper documentation
negative_class_prob_sorted = feature_log_prob[0,:].argsort()[::-1]
rank = np.array(range(1,11))
top10_negative_features = np.take(tf_idf_vect.get_feature_names(),negative_class_prob_sorted[:10])
prob = np.take(feature_log_prob[1,:], negative_class_prob_sorted[:10])
top10_negative_features_details = pd.DataFrame(data = {'Rank' : rank,'Feature' : top10_negative_fea
tures,'Probability' : prob})
print(top10_negative_features_details)
```

```
   Feature  Probability  Rank
0      not    -5.524384     1
1     like    -5.986849     2
2  product    -6.143524     3
3    taste    -6.200062     4
4    would    -6.457109     5
5      one    -6.164836     6
6      tea    -5.611215     7
7       no    -6.564834     8
8     good    -5.812261     9
9   flavor    -6.219524    10
```

In [48]:

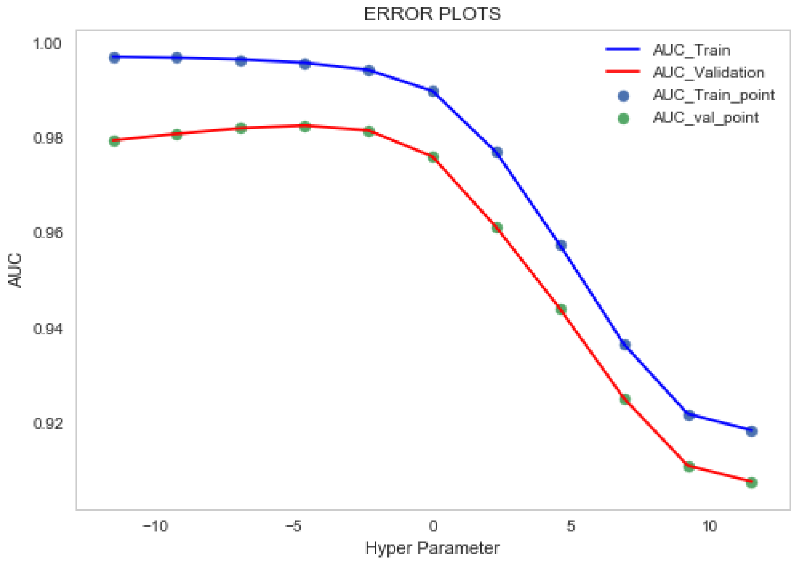
```
# Feature Enginerring
# Taken review length as new feature
from scipy.sparse import hstack,coo_matrix

data_train_review_len_tfidf = hstack([data_train_tfidf,coo_matrix(np.array(data_train_reviews_len.
tolist())[:None]])])
data_test_review_len_tfidf = hstack([data_test_tfidf,coo_matrix(np.array(data_test_reviews_len)[:N
one])])

print(data_train_review_len_tfidf.shape)
print(data_test_review_len_tfidf.shape)
```

```
(136730, 111854)
(20000, 111854)
```

In [49]:

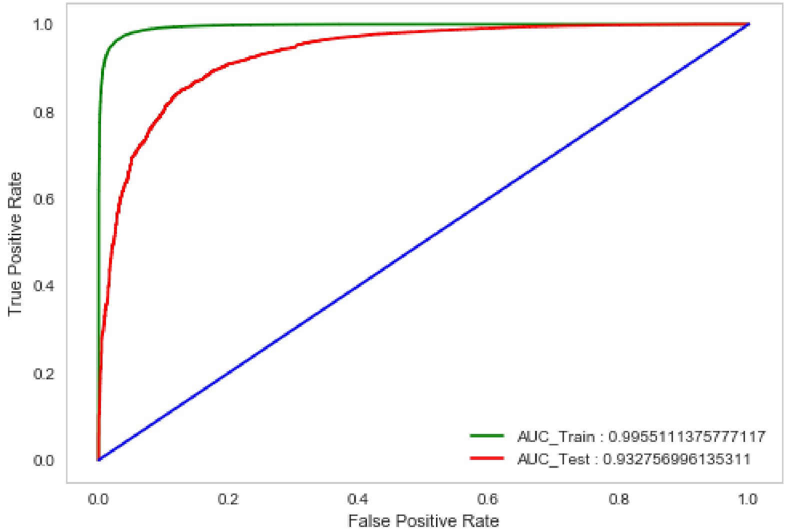
```
# Please write all the code with proper documentation
# dict alpha
hyper_parameter = {"alpha" : [10**-5,10**-4,10**-3,10**-2,10**-1,10**0,10**1,10**2,10**3,10**4,10**
5]}
# AUC
train_auc,cv_auc = get_AUC(data_train_review_len_tfidf,scores_train,hyper_parameter)
# alpha range
```

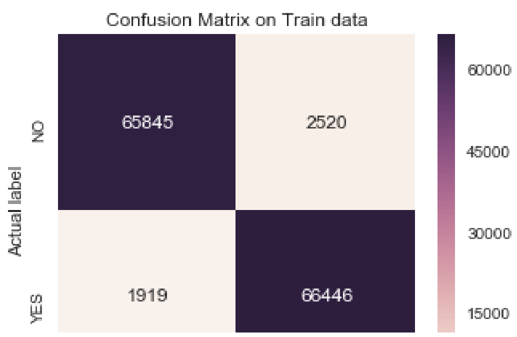
```
alpha = hyper_parameter.get("alpha")
# plot auc
plot_AUC_Curves(train_auc,cv_auc,alpha)
```

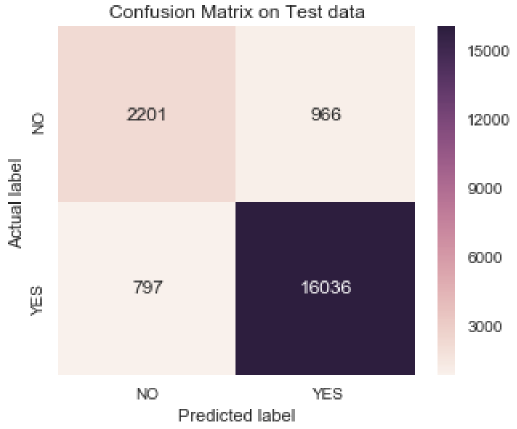


In [50]:

```
# optimal alph
optimal_alpha_tf_idf_FE = alpha[cv_auc.tolist().index(max(cv_auc.tolist()))]
# roc
fpr_train,tpr_train,fpr_test,tpr_test,pred_train,pred_test,feature_log_prob =
apply_roc_curve(data_train_review_len_tfidf,scores_train,data_test_review_len_tfidf,scores_test,op
timal_alpha_tf_idf_FE)
# auc
auc_tf_idf_FE = auc(fpr_test,tpr_test)
# plot confusion matrix
plot_roc_curve(fpr_train,tpr_train,fpr_test,tpr_test)
plot_Confusion_Matrix(scores_train,pred_train,"Confusion Matrix on Train data")
plot_Confusion_Matrix(scores_test,pred_test,"Confusion Matrix on Test data")
```







[6] Conclusions

In [51]:

```
# Please compare all your models using Prettytable library

from prettytable import PrettyTable

table = PrettyTable()
table.field_names = ["Vectorizer","Hyper parameter","AUC"]
table.add_row(["BOW",optimal_alpha_bow,round(auc_bow,2)])
table.add_row(["TFIDF",optimal_alpha_tf_idf,round(auc_TF_IDF,2)])
table.add_row(["BOW_FE",optimal_alpha_bow_FE,round(auc_bow_FE,2)])
table.add_row(["TFIDF_FE",optimal_alpha_tf_idf_FE,round(auc_tf_idf_FE,2)])

print(table.get_string(title="Results"))
```

```
+---------------------------------------+
|                Results                |
+------------+-----------------+--------+
| Vectorizer | Hyper parameter |  AUC   |
+------------+-----------------+--------+
|    BOW     |       0.01      |  0.92  |
|   TFIDF    |       0.01      |  0.93  |
|   BOW_FE   |       0.01      |  0.92  |
|  TFIDF_FE  |       0.01      |  0.93  |
+------------+-----------------+--------+
```

In [ ]: