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
```

```
D:\AAnaconda\lib\site-packages\gensim\utils.py:1212: UserWarning: detected Windows; aliasing chunkize t
o chunkize_serial
  warnings.warn("detected Windows; aliasing chunkize to chunkize_serial")
```

In [2]:

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

# filtering only positive and negative reviews i.e.
# not taking into consideration those reviews with Score=3
# SELECT * FROM Reviews WHERE Score != 3 LIMIT 500000, will give top 500000 data 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""", con)
# for tsne assignment you can take 5k data points

filtered_data = pd.read_sql_query(""" SELECT * FROM Reviews WHERE Score != 3 LIMIT 100000""", 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 (100000, 10)
```

Out[2]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Ti
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	1	1303862

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Ti
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	0	1346976
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1	1	1219017

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	B005ZBZLT4	Breyton	1331510400	2	Overall its just OK when considering the price...	2
1	#oc-R11D9D7SHXIJB9	B005HG9ESG	Louis E. Emory "hoppy"	1342396800	5	My wife has recurring extreme muscle spasms, u...	3
2	#oc-R11DNU2NBKQ23Z	B005ZBZLT4	Kim Cieszykowski	1348531200	1	This coffee is horrible and unfortunately not ...	2
3	#oc-R11O5J5ZVQE25C	B005HG9ESG	Penguin Chick	1346889600	5	This will be the bottle that you grab from the...	3
4	#oc-R12KPBODL2B5ZD	B007OSBEV0	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	B001ATMQK2	undertheshrine "undertheshrine"	1296691200	5	I bought this 6 pack because for the price tha...	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 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	
0	78445	B000HDL1RQ	AR5J8UI46CURR	Geetha Krishnan	2	2	5	11995
1	138317	B000HDOPYC	AR5J8UI46CURR	Geetha Krishnan	2	2	5	11995
2	138277	B000HDOPYM	AR5J8UI46CURR	Geetha Krishnan	2	2	5	11995
3	73791	B000HDOPZG	AR5J8UI46CURR	Geetha Krishnan	2	2	5	11995
4	155049	B000PAQ75C	AR5J8UI46CURR	Geetha Krishnan	2	2	5	11995

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='quickso
rt', na_position='last')
```

In [9]:

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

Out[9]:

```
(87775, 10)
```

In [10]:

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

Out[10]:

```
87.775
```

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

In [11]:

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

display.head()
```

Out[11]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	
0	64422	B000MIDROQ	A161DK06JJMCYF	J. E. Stephens "Jeanne"	3	1	5	12248
1	44737	B001EQ55RW	A2V0I904FH7ABY	Ram	3	2	4	12128

In [12]:

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

In [13]:

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

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

```
(87773, 10)
```

Out[13]:

```
1    73592
0    14181
Name: Score, dtype: int64
```

[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)
```

```
My dogs loves this chicken but its a product from China, so we wont be buying it anymore.  Its very har
d to find any chicken products made in the USA but they are out there, but this one isnt.  Its too bad
too because its a good product but I wont take any chances till they know what is going on with the chi
na imports.
==================================================
The Candy Blocks were a nice visual for the Lego Birthday party but the candy has little taste to it.
Very little of the 2 lbs that I bought were eaten and I threw the rest away.  I would not buy the candy
again.
==================================================
was way to hot for my blood, took a bite and did a jig  lol
==================================================
My dog LOVES these treats. They tend to have a very strong fish oil smell. So if you are afraid of the
fishy smell, don't get it. But I think my dog likes it because of the smell. These treats are really sm
all in size. They are great for training. You can give your dog several of these without worrying about
him over eating. Amazon's price was much more reasonable than any other retailer. You can buy a 1 pound
bag on Amazon for almost the same price as a 6 ounce bag at other retailers. It's definitely worth it t
o buy a big bag if your dog eats them a lot.
==================================================
```

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)
```

```
My dogs loves this chicken but its a product from China, so we wont be buying it anymore.  Its very har
d to find any chicken products made in the USA but they are out there, but this one isnt.  Its too bad
too because its a good product but I wont take any chances till they know what is going on with the chi
na imports.
```

In [16]:

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

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

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

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

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

```
My dogs loves this chicken but its a product from China, so we wont be buying it anymore.  Its very har
d to find any chicken products made in the USA but they are out there, but this one isnt.  Its too bad
too because its a good product but I wont take any chances till they know what is going on with the chi
na imports.
==================================================
The Candy Blocks were a nice visual for the Lego Birthday party but the candy has little taste to it.
Very little of the 2 lbs that I bought were eaten and I threw the rest away.  I would not buy the candy
again.
==================================================
was way to hot for my blood, took a bite and did a jig  lol
==================================================
My dog LOVES these treats. They tend to have a very strong fish oil smell. So if you are afraid of the
fishy smell, don't get it. But I think my dog likes it because of the smell. These treats are really sm
all in size. They are great for training. You can give your dog several of these without worrying about
him over eating. Amazon's price was much more reasonable than any other retailer. You can buy a 1 pound
bag on Amazon for almost the same price as a 6 ounce bag at other retailers. It's definitely worth it t
o buy a big bag if your dog eats them a lot.
```

In [17]:

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

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

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

In [18]:

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

```
was way to hot for my blood, took a bite and did a jig  lol
==================================================
```

In [19]:

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

```
My dogs loves this chicken but its a product from China, so we wont be buying it anymore.  Its very har
d to find any chicken products made in the USA but they are out there, but this one isnt.  Its too bad
too because its a good product but I wont take any chances till they know what is going on with the chi
na imports.
```

In [20]:

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

```
was way to hot for my blood took a bite and did a jig lol
```

In [21]:

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

stopwords= set(['br', 'the', 'i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you'r
e", "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', 't
heir',\
            'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these',
'those', \
            'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'd
o', 'does', \
            'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'whil
e', 'of', \
            'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during', 'bef
ore', 'after',\
            'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'a
gain', 'further',\
            'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each
', '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', "doesn
'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%|██████████| 87773/87773 [00:41<00:00, 2135.28it/s]
```

In [23]:

```
y = final['Score']
final['CleanedText'] = preprocessed_reviews
preprocessed_reviews[1500]
```

Out[23]:

```
'way hot blood took bite jig lol'
```

[3.2] Preprocessing Review Summary

In [66]:

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

[4] Featurization

[4.1] BAG OF WORDS

In [25]:

```
#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', 'aaa', 'aaaa', 'aaaaa', 'aaaaaaaaaaaa', 'aaaaaaaaaaaaaa', 'aaaaaaahhhhhh',
'aaaaaaarrrrrggghhh', 'aaaaaawwwwwwwwww', 'aaaaah']
==================================================
the type of count vectorizer  <class 'scipy.sparse.csr.csr_matrix'>
the shape of out text BOW vectorizer  (87773, 54904)
the number of unique words  54904
```

[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/skl
earn.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.get_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', 'able find', 'able get', 'absolute
', 'absolutely', 'absolutely delicious', '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), ('e
xcellent', 0.9944332838058472), ('especially', 0.9941144585609436), ('baked', 0.9940600395202637), ('sa
lted', 0.994047224521637), ('alternative', 0.9937226176261902), ('tasty', 0.9936816692352295), ('health
y', 0.9936649799346924)]
==================================================
[('varieties', 0.9994194507598877), ('become', 0.9992934465408325), ('popcorn', 0.9992750883102417), ('
de', 0.9992610216140747), ('miss', 0.9992451071739197), ('melitta', 0.999218761920929), ('choice', 0.99
92102384567261), ('american', 0.9991837739944458), ('beef', 0.9991780519485474), ('finish', 0.999156713
4857178)]
```

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', 'stinky', 'right', 'nearby', 'used'
, 'ca', 'not', 'beat', 'great', 'received', 'shipment', 'could', 'hardly', 'wait', 'try', 'love', 'call
', 'instead', 'removed', 'easily', 'daughter', 'designed', 'printed', 'use', 'car', 'windows', 'beautif
ully', 'shop', 'program', 'going', 'lot', 'fun', 'everywhere', 'like', 'tv', 'computer', 'really', 'goo
d', 'idea', 'final', 'outstanding', 'window', 'everybody', 'asks', 'bought', 'made']
```

[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, you might need to change this to 3
00 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]))
```

```
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 value
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 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
```

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

[5] Assignment 8: Decision Trees

1. **Apply Decision Trees 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 (best `depth` in range [1, 5, 10, 50, 100, 500, 100], and the best `min_samples_split` in range [5, 10, 100, 500])**
    - 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. **Graphviz**
    - Visualize your decision tree with Graphviz. It helps you to understand how a decision is being made, given a new vector.
    - Since feature names are not obtained from word2vec related models, visualize only BOW & TFIDF decision trees using Graphviz
    - Make sure to print the words in each node of the decision tree instead of printing its index.
    - Just for visualization purpose, limit max_depth to 2 or 3 and either embed the generated images of graphviz in your notebook, or directly upload them as .png files.

4. **Feature importance**

   - Find the top 20 important features from both feature sets Set 1 and Set 2 using `feature_importances_` method of Decision Tree Classifier and print their corresponding feature names

5. **Feature engineering**

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

6. **Representation of results**

   - You need to plot the performance of model both on train data and cross validation data for each hyper parameter, like shown in the figure.
   - Once after you found the best hyper parameter, you need to train your model with it, and find the AUC on test data and plot the ROC curve on both train and test.
   - Along with plotting ROC curve, you need to print the confusion matrix with predicted and original labels of test data points. Please visualize your confusion matrices using seaborn heatmaps.

7. **Conclusion**

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

Note: Data Leakage

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

In [24]:

```
#Splitting data into train and test:

from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.model_selection import cross_val_score
from collections import Counter
from sklearn.metrics import accuracy_score
from sklearn import model_selection

X_train, X_test, y_train, y_test = train_test_split(final, y, test_size=0.2)
print(X_train.shape, y_train.shape)
print(X_test.shape, y_test.shape)

#Splitting train data into train and cv(60:20)
X_tr, X_cv, y_tr, y_cv = train_test_split(X_train, y_train, test_size=0.2)
print(X_tr.shape, y_tr.shape)
print(X_cv.shape, y_cv.shape)
```

```
(70218, 11) (70218,)
(17555, 11) (17555,)
(56174, 11) (56174,)
(14044, 11) (14044,)
```

In [25]:

```
#Applying BoW
model = CountVectorizer()
model.fit(X_tr['CleanedText'])
train_bow = model.transform(X_tr['CleanedText'])
cv_bow = model.transform(X_cv['CleanedText'])
test_bow = model.transform(X_test['CleanedText'])
print(test_bow.shape)
print(cv_bow.shape)
print(train_bow.shape)
```

```
(17555, 44464)
(14044, 44464)
(56174, 44464)
```

In [26]:

```
#Applying tf_idf vectorization
tf_idf_vect = TfidfVectorizer(ngram_range=(1,2),min_df=5)
tf_idf_vect.fit(X_tr['Text'])
train_tf_idf = tf_idf_vect.transform(X_tr['Text'])
test_tf_idf = tf_idf_vect.transform(X_test['Text'])
cv_tf_idf = tf_idf_vect.transform(X_cv['Text'])

print(test_tf_idf.shape)
print(train_tf_idf.shape)
print(cv_tf_idf.shape)
```

```
(17555, 112229)
(56174, 112229)
(14044, 112229)
```

In [27]:

```
# Word2Vec model for train/test and cv dataset
i=0
list_of_sent=[]
for sent in X_tr['CleanedText'].values:
    list_of_sent.append(sent.split())

print(X_tr['CleanedText'].values[0])
print("*****************************************************************")
print(list_of_sent[0])


# Word2Vec model for test and CV
i=0
list_of_sent_cv=[]
for sent in X_cv['CleanedText'].values:
    list_of_sent_cv.append(sent.split())

print(X_cv['CleanedText'].values[0])
print("*****************************************************************")
print(list_of_sent_cv[0])


i=0
list_of_sent_test=[]
for sent in X_test['CleanedText'].values:
    list_of_sent_test.append(sent.split())

print(X_test['CleanedText'].values[0])
print("*****************************************************************")
print(list_of_sent_test[0])


w2v_model_train=Word2Vec(list_of_sent,min_count=5,size=50, workers=5)
w2v_model_test=Word2Vec(list_of_sent_test,min_count=5,size=50, workers=5)
w2v_model_cv=Word2Vec(list_of_sent_cv,min_count=5,size=50, workers=5)


w2v_words = list(w2v_model_train.wv.vocab)
print("number of words that occured minimum 5 times ",len(w2v_words))
print("sample words ", w2v_words[0:50])


# average Word2Vec
# compute average word2vec for each review Train dataset
sent_vectors = []; # the avg-w2v for each sentence/review is stored in this list
for sent in tqdm(list_of_sent): # for each review/sentence
    sent_vec = np.zeros(50) # as word vectors are of zero length
    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
    for word in sent: # for each word in a review/sentence
        if word in w2v_words:
            vec = w2v_model_train.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]))


# average Word2Vec
# compute average word2vec for each review - test dataset
sent_vectors_test = []; # the avg-w2v for each sentence/review is stored in this list
for sent in tqdm(list_of_sent_test): # for each review/sentence
    sent_vec_test = np.zeros(50) # as word vectors are of zero length
    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_train.wv[word]
            sent_vec += vec
            cnt_words += 1
    if cnt_words != 0:
        sent_vec /= cnt_words
    sent_vectors_test.append(sent_vec)
print(len(sent_vectors_test))
print(len(sent_vectors_test[0]))


# average Word2Vec
# compute average word2vec for each review - cv dataset
sent_vectors_cv = []; # the avg-w2v for each sentence/review is stored in this list
for sent in tqdm(list_of_sent_cv): # for each review/sentence
    sent_vec_cv = np.zeros(50) # as word vectors are of zero length
    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_train.wv[word]
            sent_vec += vec
            cnt_words += 1
    if cnt_words != 0:
        sent_vec /= cnt_words
    sent_vectors_cv.append(sent_vec)
print(len(sent_vectors_cv))
print(len(sent_vectors_cv[0]))
```

```
cookies delicious light crisp tangy not readily available local groceries searched online happily surpr
ised find could buy amazon family takes turns buying pack boxes sharing
*****************************************************************
['cookies', 'delicious', 'light', 'crisp', 'tangy', 'not', 'readily', 'available', 'local', 'groceries'
, 'searched', 'online', 'happily', 'surprised', 'find', 'could', 'buy', 'amazon', 'family', 'takes', 't
urns', 'buying', 'pack', 'boxes', 'sharing']
soy chai latte blew mind driving missoula montana yearned years capture taste tried chai chai no luck c
hai comes close like spicey chai without sugar milk premixed may chai
*****************************************************************
['soy', 'chai', 'latte', 'blew', 'mind', 'driving', 'missoula', 'montana', 'yearned', 'years', 'capture
', 'taste', 'tried', 'chai', 'chai', 'no', 'luck', 'chai', 'comes', 'close', 'like', 'spicey', 'chai',
'without', 'sugar', 'milk', 'premixed', 'may', 'chai']
mo old daughter eats earth best products loves except not figure finally concluded smells little like c
atfood tried feeding mango chicken flavor different times would eat bites gave donated rest food bank h
oping someone else child would like good experience subscribe save products general no broken jars quic
k delivery great find cheaper whole foods even coupons flavor earth best baby food daughter dislikes wo
rth suggest skipping one trying variety pack first
*****************************************************************
['mo', 'old', 'daughter', 'eats', 'earth', 'best', 'products', 'loves', 'except', 'not', 'figure', 'fin
ally', 'concluded', 'smells', 'little', 'like', 'catfood', 'tried', 'feeding', 'mango', 'chicken', 'fla
vor', 'different', 'times', 'would', 'eat', 'bites', 'gave', 'donated', 'rest', 'food', 'bank', 'hoping
', 'someone', 'else', 'child', 'would', 'like', 'good', 'experience', 'subscribe', 'save', 'products',
'general', 'no', 'broken', 'jars', 'quick', 'delivery', 'great', 'find', 'cheaper', 'whole', 'foods', '
even', 'coupons', 'flavor', 'earth', 'best', 'baby', 'food', 'daughter', 'dislikes', 'worth', 'suggest'
, 'skipping', 'one', 'trying', 'variety', 'pack', 'first']
number of words that occured minimum 5 times  14236
sample words  ['cookies', 'delicious', 'light', 'crisp', 'tangy', 'not', 'readily', 'available', 'local
', 'groceries', 'searched', 'online', 'happily', 'surprised', 'find', 'could', 'buy', 'amazon', 'family
', 'takes', 'turns', 'buying', 'pack', 'boxes', 'sharing', 'great', 'mouth', 'feel', 'juicy', 'chewy',
'sweet', 'without', 'overpoweringly', 'super', 'antioxidant', 'ingredients', 'like', 'acai', 'goji', 'b
lueberry', 'pomegranate', 'seen', 'bars', 'type', 'natural', 'no', 'artificial', 'sweeteners', 'gluten'
, 'free']
```

```
100%|██████████| 56174/56174 [02:05<00:00, 447.26it/s]
```

```
56174
50
```

```
100%|██████████| 17555/17555 [00:39<00:00, 440.78it/s]
```

```
17555
50
```

```
100%|██████████| 14044/14044 [00:31<00:00, 442.17it/s]
```

```
14044
50
```

In [28]:

```python
# S = ["abc def pqr", "def def def abc", "pqr pqr def"]
model = TfidfVectorizer()
tf_idf_matrix = model.fit_transform(X_tr['CleanedText'].values)
# 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 [29]:

```python
# TF-IDF weighted Word2Vec for train dataset
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_sent): # 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_train.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%|██████████| 56174/56174 [34:48<00:00, 26.89it/s]
```

In [30]:

```python
# TF-IDF weighted Word2Vec for test dataset
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_test = []; # the tfidf-w2v for each sentence/review is stored in this list
row=0;
for sent in tqdm(list_of_sent_test): # for each review/sentence
    sent_vec = np.zeros(50) # as word vectors are of zero length
    weight_sum =0; # num of words with a valid vector in the sentence/review
    for word in sent: # for each word in a review/sentence
        if word in w2v_words and word in tfidf_feat:
            vec = w2v_model_train.wv[word]
#              tf_idf = tf_idf_matrix[row, tfidf_feat.index(word)]
#               tf_idf = tf_idf_matrix[row, tfidf_feat.index(word)]
            # to reduce the computation we are
            # dictionary[word] = idf value of word in whole courpus
            # sent.count(word) = tf valeus of word in this review
            tf_idf = dictionary[word]*(sent.count(word)/len(sent))
            sent_vec += (vec * tf_idf)
            weight_sum += tf_idf
    if weight_sum != 0:
        sent_vec /= weight_sum
    tfidf_sent_vectors_test.append(sent_vec)
    row += 1


# TF-IDF weighted Word2Vec for cv dataset
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_cv = []; # the tfidf-w2v for each sentence/review is stored in this list
row=0;
for sent in tqdm(list_of_sent_cv): # 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_train.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_cv.append(sent_vec)
    row += 1
```

```
100%|██████████| 17555/17555 [10:30<00:00, 27.83it/s]
100%|██████████| 14044/14044 [39:24<00:00,  5.94it/s]
```

Applying Decision Trees

[5.1] Applying Decision Trees on BOW, SET 1

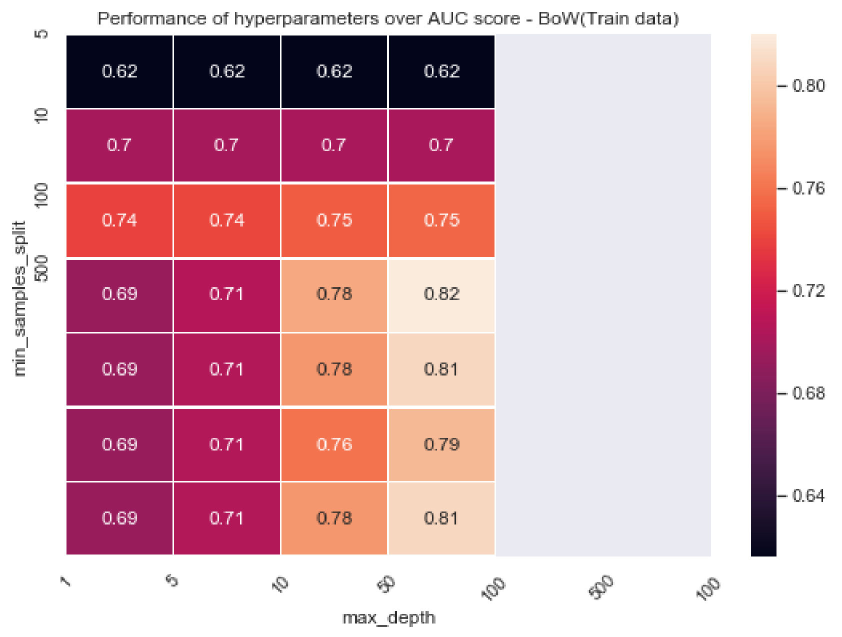
In [31]:

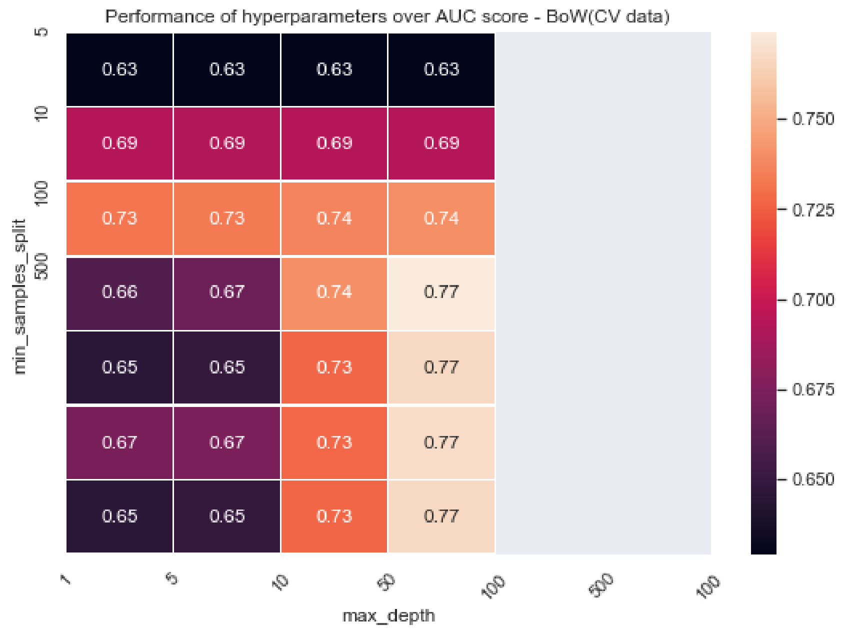
```
# Train data
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import GridSearchCV
from sklearn import linear_model
param_grid = {"max_depth": [1, 5, 10, 50, 100, 500, 100], "min_samples_split":[5, 10, 100, 500]}
scoring = {'AUC': 'roc_auc'}
clf=DecisionTreeClassifier(random_state=0)
grid = GridSearchCV(clf,param_grid=param_grid,scoring = scoring, refit = 'AUC')
#train_bow = pickle.load(fp)
grid.fit(train_bow, y_tr)
print(grid)
# summarize the results of the grid search
(grid.best_score_)
print(grid.best_estimator_.max_depth)
print(grid.best_estimator_.min_samples_split)
results_tr_bow = grid.cv_results_
    #print(results)


# CV Data
grid.fit(cv_bow, y_cv)
print(grid)
# summarize the results of the grid search
(grid.best_score_)
print(grid.best_estimator_.max_depth)
print(grid.best_estimator_.min_samples_split)
results_cv_bow = grid.cv_results_
    #print(results)
```

```
GridSearchCV(cv='warn', error_score='raise-deprecating',
       estimator=DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=0,
            splitter='best'),
       fit_params=None, iid='warn', n_jobs=None,
       param_grid={'max_depth': [1, 5, 10, 50, 100, 500, 100], 'min_samples_split': [5, 10, 100, 500]},
       pre_dispatch='2*n_jobs', refit='AUC', return_train_score='warn',
       scoring={'AUC': 'roc_auc'}, verbose=0)
50
500
GridSearchCV(cv='warn', error_score='raise-deprecating',
       estimator=DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=0,
            splitter='best'),
       fit_params=None, iid='warn', n_jobs=None,
       param_grid={'max_depth': [1, 5, 10, 50, 100, 500, 100], 'min_samples_split': [5, 10, 100, 500]},
       pre_dispatch='2*n_jobs', refit='AUC', return_train_score='warn',
       scoring={'AUC': 'roc_auc'}, verbose=0)
50
500
```

In [32]:

```
#Heatmap for Train dataset
import seaborn as sns
max_depth = [1, 5, 10, 50, 100, 500, 100]
min_samples_split = [5, 10, 100, 500]
scores = results_tr_bow['mean_test_AUC'].reshape(len(max_depth),len(min_samples_split))
## Plotting Function
#plt.figure(figsize=(8, 6))
sns.set()
#plt.subplots_adjust(left=.2, right=0.95, bottom=0.15, top=0.95)
f, ax = plt.subplots(figsize=(9, 6))
sns.heatmap(scores, annot=True, linewidths=.5, ax=ax)
#plt.imshow(scores, interpolation='nearest', cmap=plt.cm.hot)
plt.xlabel('max_depth')
plt.ylabel('min_samples_split')
#plt.colorbar()
plt.xticks(np.arange(len(max_depth)), max_depth, rotation=45)
plt.yticks(np.arange(len(min_samples_split)), min_samples_split)
plt.title('Performance of hyperparameters over AUC score - BoW(Train data)')
plt.show()


#Heatmap for Test dataset
import seaborn as sns
max_depth = [1, 5, 10, 50, 100, 500, 100]
min_samples_split = [5, 10, 100, 500]
scores = results_cv_bow['mean_test_AUC'].reshape(len(max_depth),len(min_samples_split))
## Plotting Function
#plt.figure(figsize=(8, 6))
sns.set()
#plt.subplots_adjust(left=.2, right=0.95, bottom=0.15, top=0.95)
f, ax = plt.subplots(figsize=(9, 6))
sns.heatmap(scores, annot=True, linewidths=.5, ax=ax)
#plt.imshow(scores, interpolation='nearest', cmap=plt.cm.hot)
plt.xlabel('max_depth')
plt.ylabel('min_samples_split')
#plt.colorbar()
plt.xticks(np.arange(len(max_depth)), max_depth, rotation=45)
plt.yticks(np.arange(len(min_samples_split)), min_samples_split)
plt.title('Performance of hyperparameters over AUC score - BoW(CV data)')
plt.show()
```





In [33]:

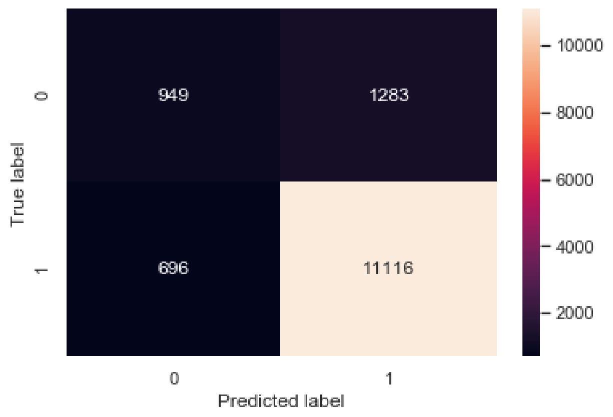
```
#Applying Decision tree for CV dataset
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import GridSearchCV
from sklearn import linear_model
clf = DecisionTreeClassifier(max_depth =50, min_samples_split=500)
scoring = {'AUC': 'roc_auc'}
cclf = clf.fit(train_bow,y_tr).predict(cv_bow)
#Caliberate the classifier.
#clf_calibrated=CalibratedClassifierCV(clf, cv='prefit', method='isotonic')
#clf_calibrated.fit(cv_bow, y_cv)
pred_cv = clf.predict_proba(cv_bow)[:,1]
fpr, tpr, thresholds = roc_curve(y_cv,pred_cv)
roc_auc_cv = auc(fpr, tpr)
print('Area under the ROC curve : %f', + roc_auc_cv)

#Plotting confusion matrix
import seaborn as sns
conf_mat = confusion_matrix(y_cv, cclf)
print(conf_mat)
#conf_normalized = conf_mat.astype('int') / conf_mat.sum(axis=1)[:, np.newaxis]
sns.heatmap(conf_mat, annot=True, fmt ='g')
plt.ylabel('True label')
plt.xlabel('Predicted label')
```

```
Area under the ROC curve : %f 0.8231869176234119
[[  949  1283]
```

```
 [  696 11116]]
```

Out[33]:

```
Text(0.5, 12.5, 'Predicted label')
```



In [34]:

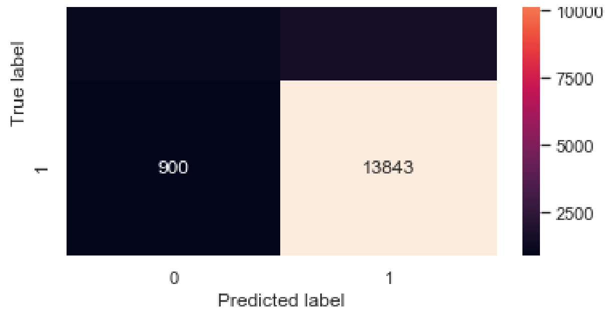
```python
#Applying Decision Tree for test dataset
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import GridSearchCV
from sklearn import linear_model
clf = DecisionTreeClassifier(max_depth =50, min_samples_split=500)
scoring = {'AUC': 'roc_auc', 'Accuracy': 'accuracy', 'Precision' : 'precision', 'Recall' : 'recall'}
cclf = clf.fit(train_bow,y_tr).predict(test_bow)
pred_test = clf.predict_proba(test_bow)[:,1]
fpr, tpr, thresholds = roc_curve(y_test,pred_test)
roc_auc_test = auc(fpr, tpr)
print('Area under the ROC curve : %f', + roc_auc_test)

#Plotting confusion matrix
import seaborn as sns
conf_mat = confusion_matrix(y_test, cclf)
print(conf_mat)
#conf_normalized = conf_mat.astype('int') / conf_mat.sum(axis=1)[:, np.newaxis]
sns.heatmap(conf_mat, annot=True, fmt ='g')
plt.ylabel('True label')
plt.xlabel('Predicted label')

#Plot ROC Curve
# calculate the fpr and tpr for all thresholds of the classification
#https://stackoverflow.com/questions/25009284/how-to-plot-roc-curve-in-python
plt.figure(0).clf()
fpr, tpr, thresholds = roc_curve(y_test,pred_test)
roc_auc_test_tfidf = auc(fpr, tpr)
plt.plot(fpr,tpr,label="Test Data, auc="+str(roc_auc_test))

fpr, tpr, thresh = roc_curve(y_cv, pred_cv)
roc_auc_cv_tfidf = auc(fpr, tpr)
plt.plot(fpr,tpr,label="Train Data, auc="+str(roc_auc_cv))
plt.title('ROC curve for Train and Test Dataset - BOW')
plt.xlabel('True Positive Rate')
plt.ylabel('False Positive Rate')
plt.legend(loc=0)
```

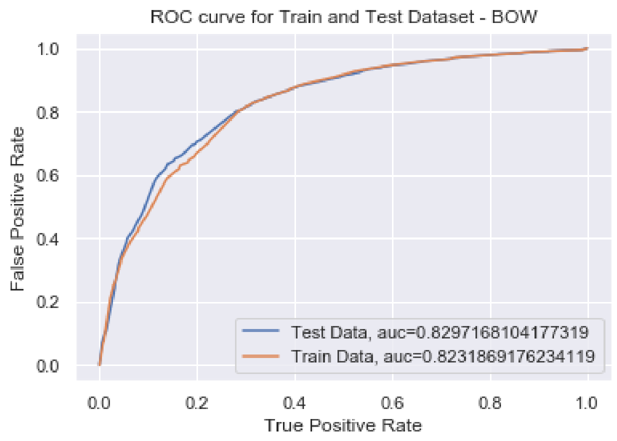
```
Area under the ROC curve : %f 0.8297168104177319
[[ 1189  1623]
 [  900 13843]]
```

Out[34]:

```
<matplotlib.legend.Legend at 0x475c3d68>
```

1189
1623
0
12500

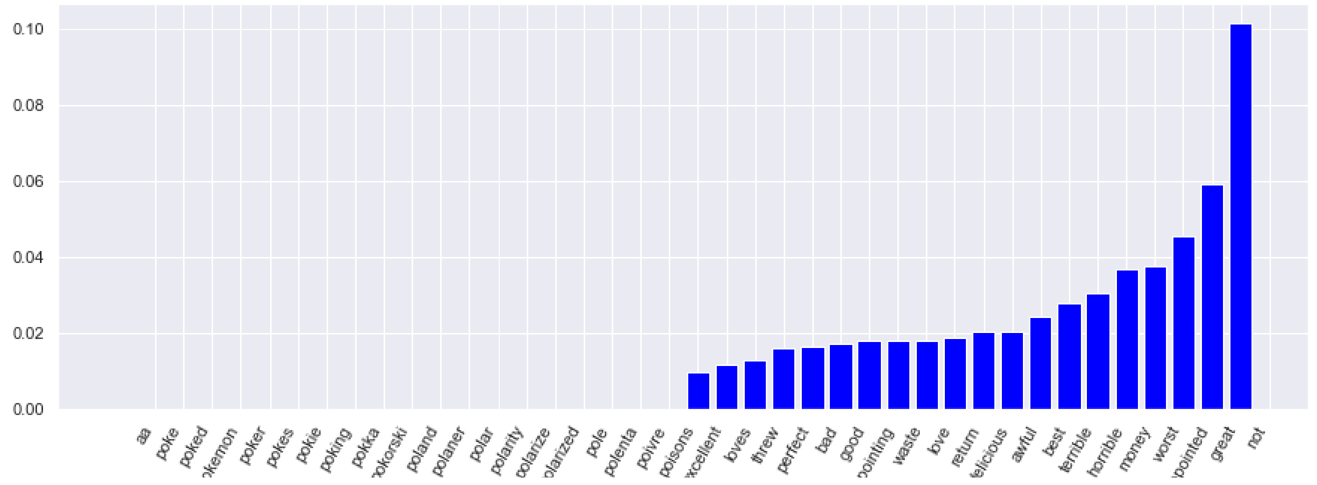




[5.1.1] Top 20 important features from SET 1

In [35]:

```
# Please write all the code with proper documentation
#Finding the top 20 features in BOW:
import matplotlib.pyplot as plt
from sklearn.tree import DecisionTreeClassifier
from sklearn import tree
def plot_coefficients(classifier, feature_names, top_features=20):
    coef = classifier.feature_importances_.ravel();
    top_positive_coefficients = np.argsort(coef)[-top_features:]
    top_negative_coefficients = np.argsort(coef)[:top_features]
    top_coefficients = np.hstack([top_negative_coefficients, top_positive_coefficients])
 # create plot
    plt.figure(figsize=(15, 5))
    colors = ['red' if c < 0 else 'blue' for c in coef[top_coefficients]]
    plt.bar(np.arange(2 * top_features), coef[top_coefficients], color=colors)
    feature_names = np.array(feature_names)
    plt.xticks(np.arange(1, 1 + 2 * top_features), feature_names[top_coefficients], rotation=60, ha='ri
ght')
    plt.show()
    print(feature_names[top_positive_coefficients])
    print(feature_names[top_negative_coefficients])

plot_coefficients(clf, model.get_feature_names(), top_features=20)
```



```
['excellent' 'loves' 'threw' 'perfect' 'bad' 'good' 'disappointing'
 'waste' 'love' 'return' 'delicious' 'awful' 'best' 'terrible' 'horrible'
 'money' 'worst' 'disappointed' 'great' 'not']
['aa' 'poke' 'poked' 'pokemon' 'poker' 'pokes' 'pokie' 'poking' 'pokka'
 'pokorski' 'poland' 'polaner' 'polar' 'polarity' 'polarize' 'polarized'
 'pole' 'polenta' 'poivre' 'poisons']
```

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

In [167]:

```python
# Please write all the code with proper documentation
from sklearn.tree import DecisionTreeClassifier
from sklearn import tree
import graphviz
from graphviz import Source
from IPython.display import Image
#import pydot
import pydotplus
import os
#os.environ['PATH'].split(os.pathsep)
#os.environ["PATH"] += os.pathsep + 'C:\Program Files (x86)\Graphviz2.38\bin'
#import pydot
#dot_data = tree.export_graphviz(clf,max_depth=2,out_file=None,feature_names=model.get_feature_names())
#graph = pydotplus.graph_from_dot_data(dot_data)
#Image(graph.create_png())
dot_data = tree.export_graphviz(clf,max_depth=2,out_file="D:/dtreebow.dot",feature_names=model.get_feat
ure_names())
#graph = pydotplus.graph_from_dot_data(dot_data.getvalue())
```

[5.2] Applying Decision Trees on TFIDF, SET 2

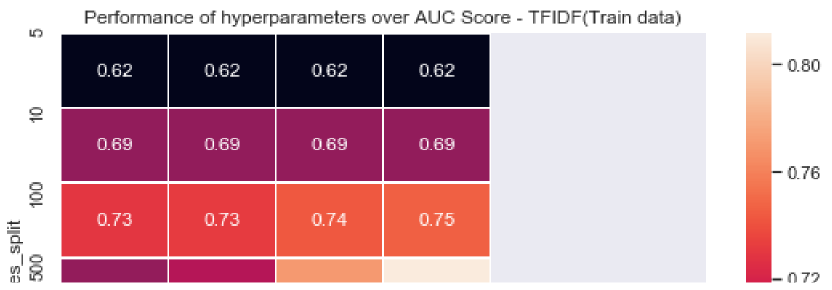
In [36]:

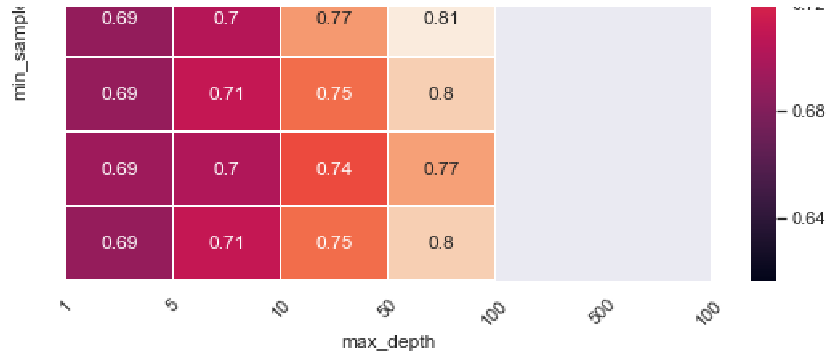
```python
# Please write all the code with proper documentation
# Please write all the code with proper documentation
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import GridSearchCV
from sklearn import linear_model# Please write all the code with proper documentation
from sklearn.tree import DecisionTreeClassifier
from sklearn import tree
param_grid = {"max_depth": [1, 5, 10, 50, 100, 500, 100], "min_samples_split":[5, 10, 100, 500]}
scoring = {'AUC': 'roc_auc', 'Accuracy': 'accuracy', 'Precision' : 'precision', 'Recall' : 'recall'}
clfA=DecisionTreeClassifier(random_state=0)
grid = GridSearchCV(clfA,param_grid=param_grid,scoring = scoring, refit = 'AUC')
grid.fit(train_tf_idf, y_tr)
print(grid)
# summarize the results of the grid search
(grid.best_score_)
print(grid.best_estimator_.max_depth)
print(grid.best_estimator_.min_samples_split)
results_tr_tfidf = grid.cv_results_
#print(results)

grid.fit(cv_tf_idf, y_cv)
print(grid)
# summarize the results of the grid search
(grid.best_score_)
print(grid.best_estimator_.max_depth)
print(grid.best_estimator_.min_samples_split)
results_cv_tfidf = grid.cv_results_
#print(results)
```

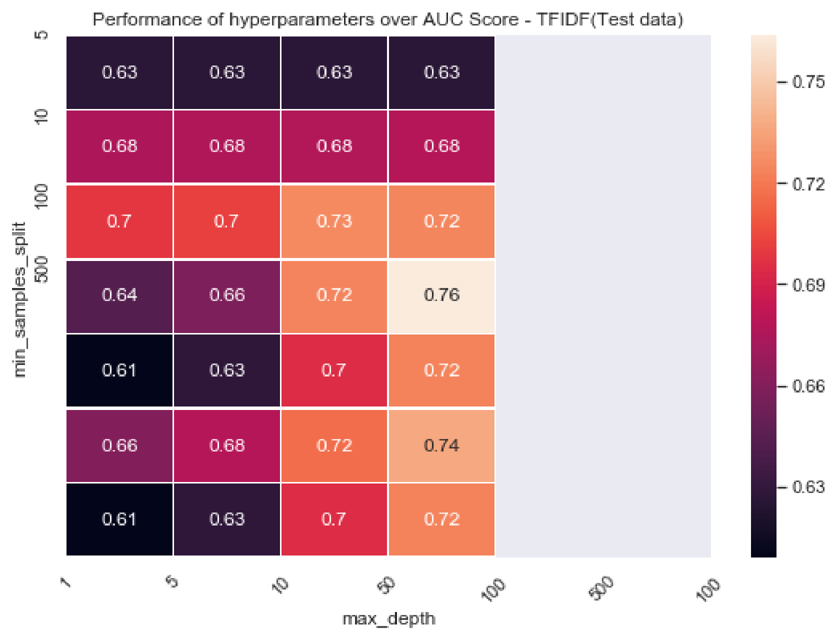
```
GridSearchCV(cv='warn', error_score='raise-deprecating',
       estimator=DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=0,
            splitter='best'),
       fit_params=None, iid='warn', n_jobs=None,
       param_grid={'max_depth': [1, 5, 10, 50, 100, 500, 100], 'min_samples_split': [5, 10, 100, 500]},
       pre_dispatch='2*n_jobs', refit='AUC', return_train_score='warn',
       scoring={'AUC': 'roc_auc', 'Accuracy': 'accuracy', 'Precision': 'precision', 'Recall': 'recall'}
,
       verbose=0)
50
500
GridSearchCV(cv='warn', error_score='raise-deprecating',
       estimator=DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=0,
            splitter='best'),
       fit_params=None, iid='warn', n_jobs=None,
       param_grid={'max_depth': [1, 5, 10, 50, 100, 500, 100], 'min_samples_split': [5, 10, 100, 500]},
       pre_dispatch='2*n_jobs', refit='AUC', return_train_score='warn',
       scoring={'AUC': 'roc_auc', 'Accuracy': 'accuracy', 'Precision': 'precision', 'Recall': 'recall'}
,
       verbose=0)
50
500
```

In [37]:

```
import seaborn as sns
max_depth = [1, 5, 10, 50, 100, 500, 100]
min_samples_split = [5, 10, 100, 500]
scores = results_tr_tfidf['mean_test_AUC'].reshape(len(max_depth),len(min_samples_split))
## Plotting Function
#plt.figure(figsize=(8, 6))
sns.set()
#plt.subplots_adjust(left=.2, right=0.95, bottom=0.15, top=0.95)
f, ax = plt.subplots(figsize=(9, 6))
sns.heatmap(scores, annot=True, linewidths=.5, ax=ax)
#plt.imshow(scores, interpolation='nearest', cmap=plt.cm.hot)
plt.xlabel('max_depth')
plt.ylabel('min_samples_split')
#plt.colorbar()
plt.xticks(np.arange(len(max_depth)), max_depth, rotation=45)
plt.yticks(np.arange(len(min_samples_split)), min_samples_split)
plt.title('Performance of hyperparameters over AUC Score - TFIDF(Train data)')
plt.show()



scores = results_cv_tfidf['mean_test_AUC'].reshape(len(max_depth),len(min_samples_split))
## Plotting Function
#plt.figure(figsize=(8, 6))
sns.set()
#plt.subplots_adjust(left=.2, right=0.95, bottom=0.15, top=0.95)
f, ax = plt.subplots(figsize=(9, 6))
sns.heatmap(scores, annot=True, linewidths=.5, ax=ax)
#plt.imshow(scores, interpolation='nearest', cmap=plt.cm.hot)
plt.xlabel('max_depth')
plt.ylabel('min_samples_split')
#plt.colorbar()
plt.xticks(np.arange(len(max_depth)), max_depth, rotation=45)
plt.yticks(np.arange(len(min_samples_split)), min_samples_split)
plt.title('Performance of hyperparameters over AUC Score - TFIDF(Test data)')
plt.show()
```







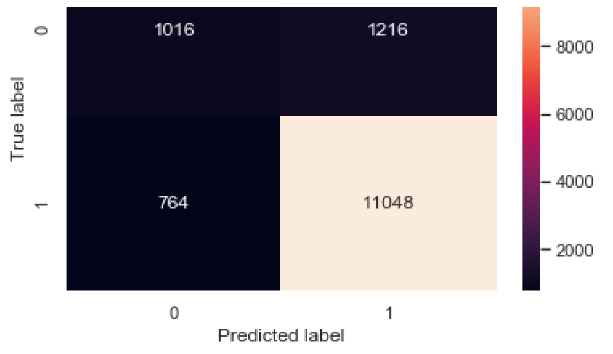
In [38]:

```
#Applying Decision tree for CV dataset
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import GridSearchCV
from sklearn import linear_model
clf = DecisionTreeClassifier(max_depth =50, min_samples_split=500)
scoring = {'AUC': 'roc_auc'}
cclf = clf.fit(train_tf_idf,y_tr).predict(cv_tf_idf)
#Caliberate the classifier.
#clf_calibrated=CalibratedClassifierCV(clf, cv='prefit', method='isotonic')
#clf_calibrated.fit(cv_bow, y_cv)
pred_cv = clf.predict_proba(cv_tf_idf)[:,1]
fpr, tpr, thresholds = roc_curve(y_cv,pred_cv)
roc_auc_cv = auc(fpr, tpr)
print('Area under the ROC curve : %f', + roc_auc_cv)

#Plotting confusion matrix
import seaborn as sns
conf_mat = confusion_matrix(y_cv, cclf)
print(conf_mat)
#conf_normalized = conf_mat.astype('int') / conf_mat.sum(axis=1)[:, np.newaxis]
sns.heatmap(conf_mat, annot=True, fmt ='g')
plt.ylabel('True label')
plt.xlabel('Predicted label')
```

```
Area under the ROC curve : %f 0.8253183916605069
[[ 1016  1216]
 [  764 11048]]
```

Out[38]:

```
Text(0.5, 12.5, 'Predicted label')
```



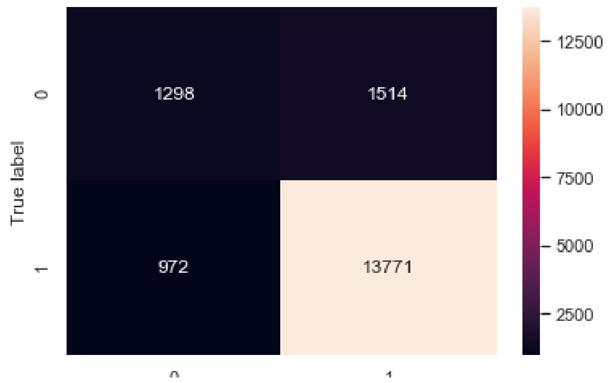
In [39]:

```
#Applying Decision tree for test dataset
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import GridSearchCV
from sklearn import linear_model
clf = DecisionTreeClassifier(max_depth =50, min_samples_split=500)
scoring = {'AUC': 'roc_auc', 'Accuracy': 'accuracy', 'Precision' : 'precision', 'Recall' : 'recall'}
cclf = clf.fit(train_tf_idf,y_tr).predict(test_tf_idf)
pred_test = clf.predict_proba(test_tf_idf)[:,1]
fpr, tpr, thresholds = roc_curve(y_test,pred_test)
roc_auc_test = auc(fpr, tpr)
print('Area under the ROC curve : %f', + roc_auc_test)

#Plotting confusion matrix
import seaborn as sns
conf_mat = confusion_matrix(y_test, cclf)
print(conf_mat)
#conf_normalized = conf_mat.astype('int') / conf_mat.sum(axis=1)[:, np.newaxis]
sns.heatmap(conf_mat, annot=True, fmt ='g')
plt.ylabel('True label')
plt.xlabel('Predicted label')

#Plot ROC Curve
# calculate the fpr and tpr for all thresholds of the classification
#https://stackoverflow.com/questions/25009284/how-to-plot-roc-curve-in-python
plt.figure(0).clf()
fpr, tpr, thresholds = roc_curve(y_test,pred_test)
roc_auc_test_tfidf = auc(fpr, tpr)
plt.plot(fpr,tpr,label="Test Data, auc="+str(roc_auc_test))

fpr, tpr, thresh = roc_curve(y_cv, pred_cv)
roc_auc_cv_tfidf = auc(fpr, tpr)
plt.plot(fpr,tpr,label="Train Data, auc="+str(roc_auc_cv))
plt.title('ROC curve for Train and Test Dataset - TFIDF')
plt.xlabel('True Positive Rate')
plt.ylabel('False Positive Rate')
plt.legend(loc=0)
```

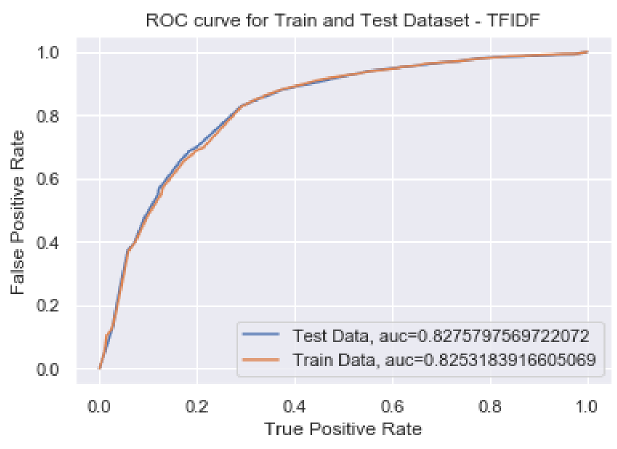
```
Area under the ROC curve : %f 0.8275797569722072
[[ 1298  1514]
 [  972 13771]]
```

Out[39]:

```
<matplotlib.legend.Legend at 0x49510cf8>
```



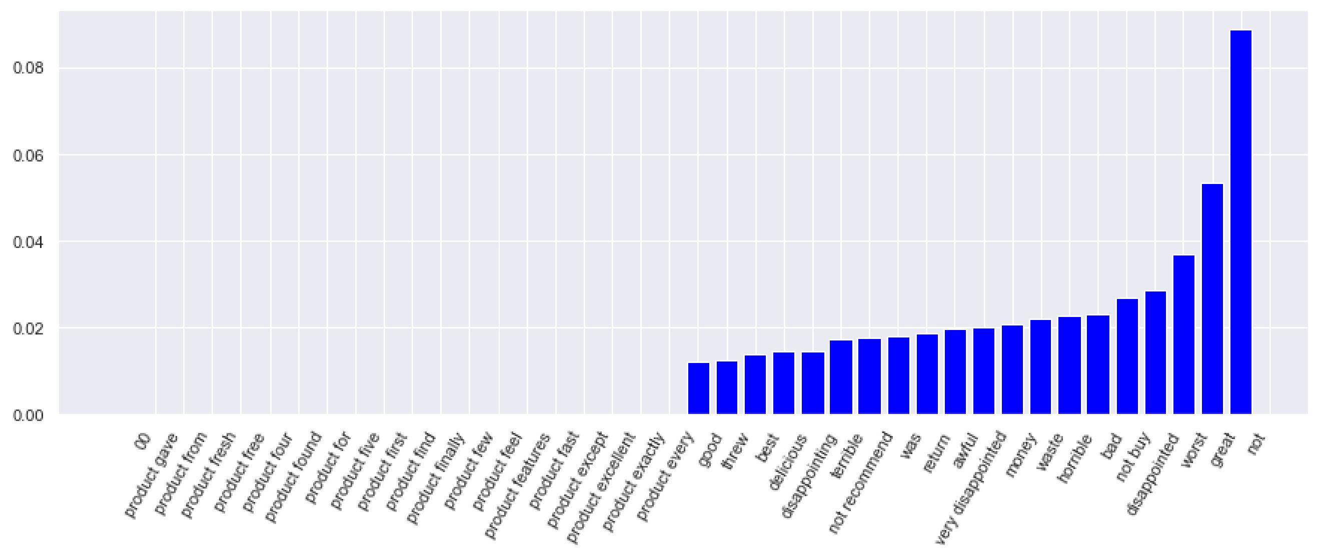
Predicted label



[5.2.1] Top 20 important features from SET 2

In [40]:

```
plot_coefficients(clf, tf_idf_vect.get_feature_names(), top_features=20)
```



```
['good' 'threw' 'best' 'delicious' 'disappointing' 'terrible'
 'not recommend' 'was' 'return' 'awful' 'very disappointed' 'money'
 'waste' 'horrible' 'bad' 'not buy' 'disappointed' 'worst' 'great' 'not']
['00' 'product gave' 'product from' 'product fresh' 'product free'
 'product four' 'product found' 'product for' 'product five'
 'product first' 'product find' 'product finally' 'product few'
 'product feel' 'product features' 'product fast' 'product except'
 'product excellent' 'product exactly' 'product every']
```

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

In [151]:

```
# Please write all the code with proper documentation
from sklearn.tree import DecisionTreeClassifier
from sklearn import tree
import graphviz
from graphviz import Source
from IPython.display import Image
#import pydot
import pydotplus
import os
#os.environ['PATH'].split(os.pathsep)
#os.environ["PATH"] += os.pathsep + 'D:\AAnaconda\Library\bin\graphviz
#dotfile = StringIO.StringIO()
```

```
#dot_data = tree.export_graphviz(clf,max_depth=2,out_file=None,feature_names=tf_idf_vect.get_feature_na
mes())
##Image(graph.create_png())
tree.export_graphviz(clf,max_depth=2,out_file="D:\dtreetf.dot",feature_names=tf_idf_vect.get_feature_na
mes())
graph = pydotplus.graph_from_dot_data(dot_data.getvalue())
```

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

In [41]:

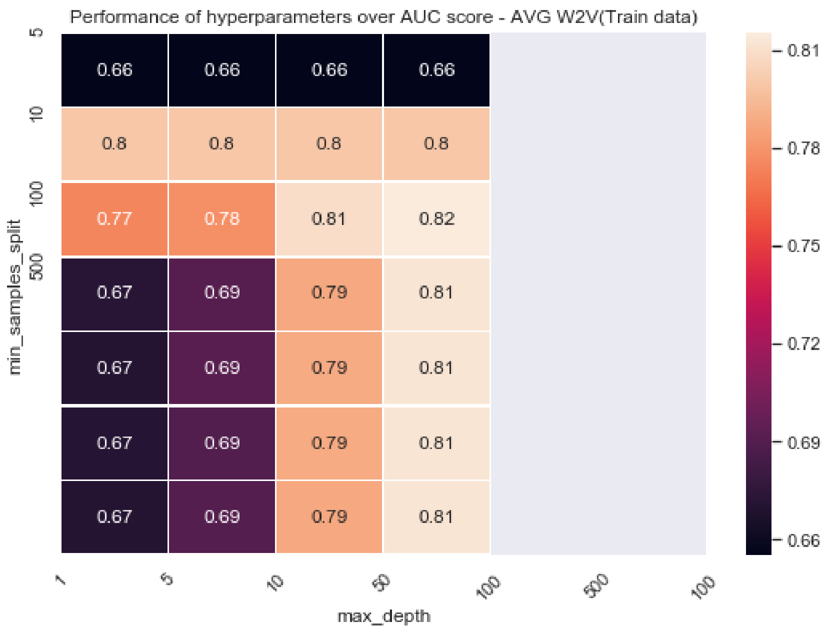
```
# Please write all the code with proper documentation
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import GridSearchCV
from sklearn import linear_model
param_grid = {"max_depth": [1, 5, 10, 50, 100, 500, 100], "min_samples_split":[5, 10, 100, 500]}
scoring = {'AUC': 'roc_auc', 'Accuracy': 'accuracy', 'Precision' : 'precision', 'Recall' : 'recall'}
clfC=DecisionTreeClassifier(random_state=0)
grid = GridSearchCV(clf,param_grid=param_grid,scoring = scoring, refit = 'AUC')
grid.fit(sent_vectors, y_tr)
#print(grid)
# summarize the results of the grid search
(grid.best_score_)
print(grid.best_estimator_.max_depth)
print(grid.best_estimator_.min_samples_split)
results_tr_avgw2v = grid.cv_results_
#print(results)


grid.fit(sent_vectors_cv, y_cv)
#print(grid)
# summarize the results of the grid search
(grid.best_score_)
print(grid.best_estimator_.max_depth)
print(grid.best_estimator_.min_samples_split)
results_cv_avgw2v = grid.cv_results_
#print(results)
```

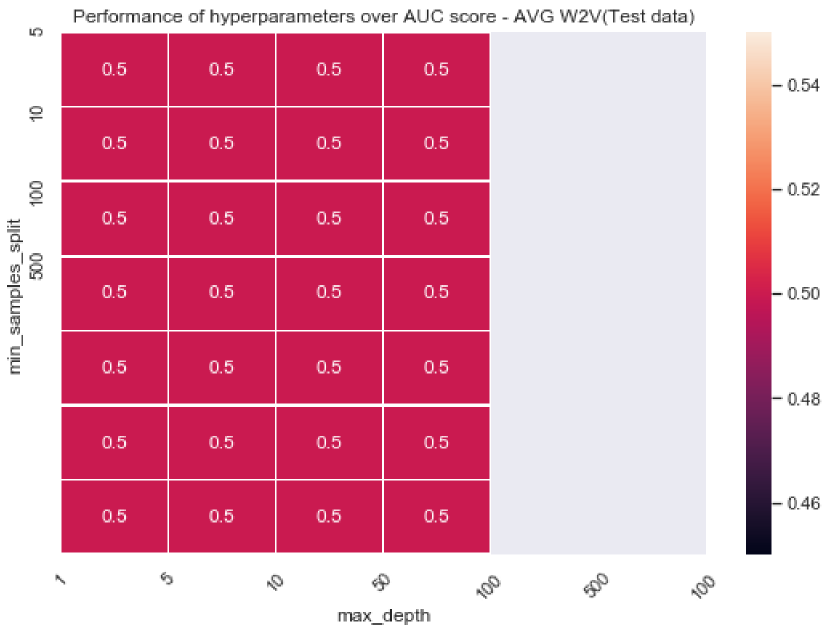
```
10
500
1
5
```

In [42]:

```
import seaborn as sns
max_depth = [1, 5, 10, 50, 100, 500, 100]
min_samples_split = [5, 10, 100, 500]
scores = results_tr_avgw2v['mean_test_AUC'].reshape(len(max_depth),len(min_samples_split))
## Plotting Function
#plt.figure(figsize=(8, 6))
sns.set()
#plt.subplots_adjust(left=.2, right=0.95, bottom=0.15, top=0.95)
f, ax = plt.subplots(figsize=(9, 6))
sns.heatmap(scores, annot=True, linewidths=.5, ax=ax)
#plt.imshow(scores, interpolation='nearest', cmap=plt.cm.hot)
plt.xlabel('max_depth')
plt.ylabel('min_samples_split')
#plt.colorbar()
plt.xticks(np.arange(len(max_depth)), max_depth, rotation=45)
plt.yticks(np.arange(len(min_samples_split)), min_samples_split)
plt.title('Performance of hyperparameters over AUC score - AVG W2V(Train data)')
plt.show()


scores = results_cv_avgw2v['mean_test_AUC'].reshape(len(max_depth),len(min_samples_split))
## Plotting Function
#plt.figure(figsize=(8, 6))
sns.set()
#plt.subplots_adjust(left=.2, right=0.95, bottom=0.15, top=0.95)
f, ax = plt.subplots(figsize=(9, 6))
sns.heatmap(scores, annot=True, linewidths=.5, ax=ax)
```

```
#plt.imshow(scores, interpolation='nearest', cmap=plt.cm.hot)
plt.xlabel('max_depth')
plt.ylabel('min_samples_split')
#plt.colorbar()
plt.xticks(np.arange(len(max_depth)), max_depth, rotation=45)
plt.yticks(np.arange(len(min_samples_split)), min_samples_split)
plt.title('Performance of hyperparameters over AUC score - AVG W2V(Test data)')
plt.show()
```





In [43]:

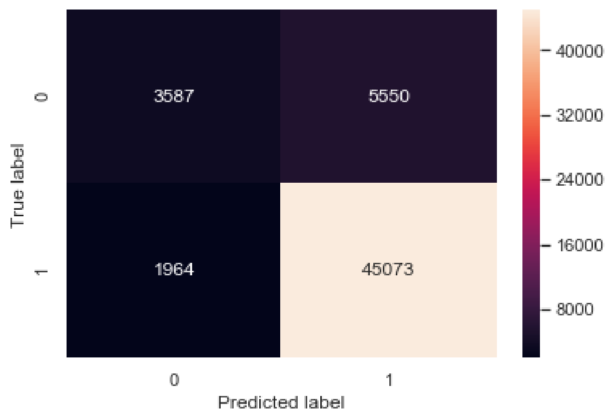
```
#Applying Decision Tree on CV dataset
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import GridSearchCV
from sklearn import linear_model
clf = DecisionTreeClassifier(max_depth =10, min_samples_split=500)
scoring = {'AUC': 'roc_auc'}
cclf = clf.fit(sent_vectors,y_tr).predict(sent_vectors)
#Caliberate the classifier.
#clf_calibrated=CalibratedClassifierCV(clf, cv='prefit', method='isotonic')
#clf_calibrated.fit(cv_bow, y_cv)
pred_cv = clf.predict_proba(sent_vectors)[:,1]
fpr, tpr, thresholds = roc_curve(y_tr,pred_cv)
roc_auc_cv = auc(fpr, tpr)
print('Area under the ROC curve : %f', + roc_auc_cv)

#Plotting confusion matrix
import seaborn as sns
```

```
conf_mat = confusion_matrix(y_tr, cclf)
print(conf_mat)
#conf_normalized = conf_mat.astype('int') / conf_mat.sum(axis=1)[:, np.newaxis]
sns.heatmap(conf_mat, annot=True, fmt ='g')
plt.ylabel('True label')
plt.xlabel('Predicted label')
```

```
Area under the ROC curve : %f 0.8627391460478316
[[ 3587  5550]
 [ 1964 45073]]
```

Out[43]:

```
Text(0.5, 12.5, 'Predicted label')
```



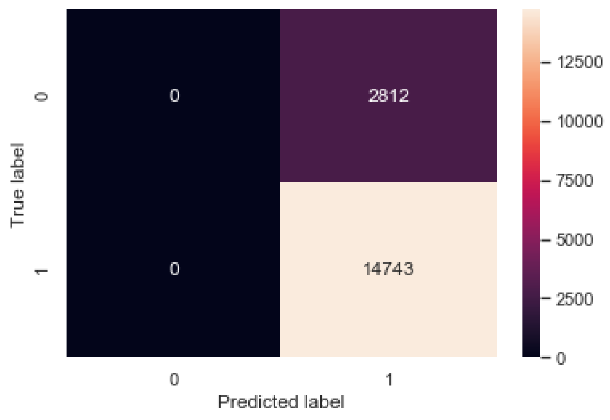
In [44]:

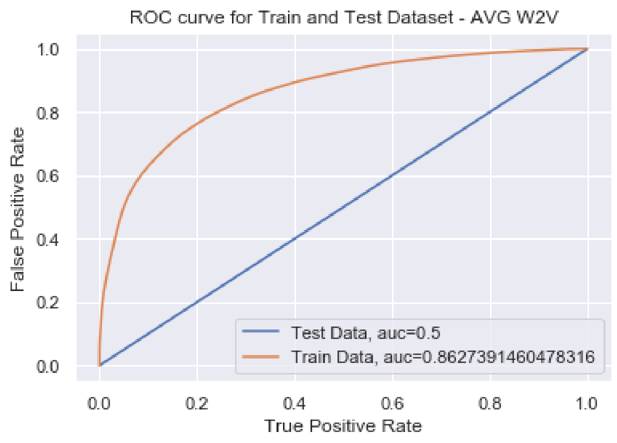
```
#Applying Decision Tree on test dataset
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import GridSearchCV
from sklearn import linear_model
clf = DecisionTreeClassifier(max_depth =10, min_samples_split=500)
scoring = {'AUC': 'roc_auc', 'Accuracy': 'accuracy', 'Precision' : 'precision', 'Recall' : 'recall'}
cclf = clf.fit(sent_vectors,y_tr).predict(sent_vectors_test)
pred_test = clf.predict_proba(sent_vectors_test)[:,1]
fpr, tpr, thresholds = roc_curve(y_test,pred_test)
roc_auc_test = auc(fpr, tpr)
print('Area under the ROC curve : %f', + roc_auc_test)

#Plotting confusion matrix
import seaborn as sns
conf_mat = confusion_matrix(y_test, cclf)
print(conf_mat)
#conf_normalized = conf_mat.astype('int') / conf_mat.sum(axis=1)[:, np.newaxis]
sns.heatmap(conf_mat, annot=True, fmt ='g')
plt.ylabel('True label')
plt.xlabel('Predicted label')

#Plot ROC Curve
# calculate the fpr and tpr for all thresholds of the classification
#https://stackoverflow.com/questions/25009284/how-to-plot-roc-curve-in-python
plt.figure(0).clf()
fpr, tpr, thresholds = roc_curve(y_test,pred_test)
roc_auc_test_tfidf = auc(fpr, tpr)
plt.plot(fpr,tpr,label="Test Data, auc="+str(roc_auc_test))

fpr, tpr, thresh = roc_curve(y_tr, pred_cv)
roc_auc_cv_tfidf = auc(fpr, tpr)
plt.plot(fpr,tpr,label="Train Data, auc="+str(roc_auc_cv))
plt.title('ROC curve for Train and Test Dataset - AVG W2V')
plt.xlabel('True Positive Rate')
plt.ylabel('False Positive Rate')
plt.legend(loc=0)
```

```
Area under the ROC curve : %f 0.5
[[    0  2812]
 [    0 14743]]
```

Out[44]:

```
<matplotlib.legend.Legend at 0x4a1cc518>
```





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

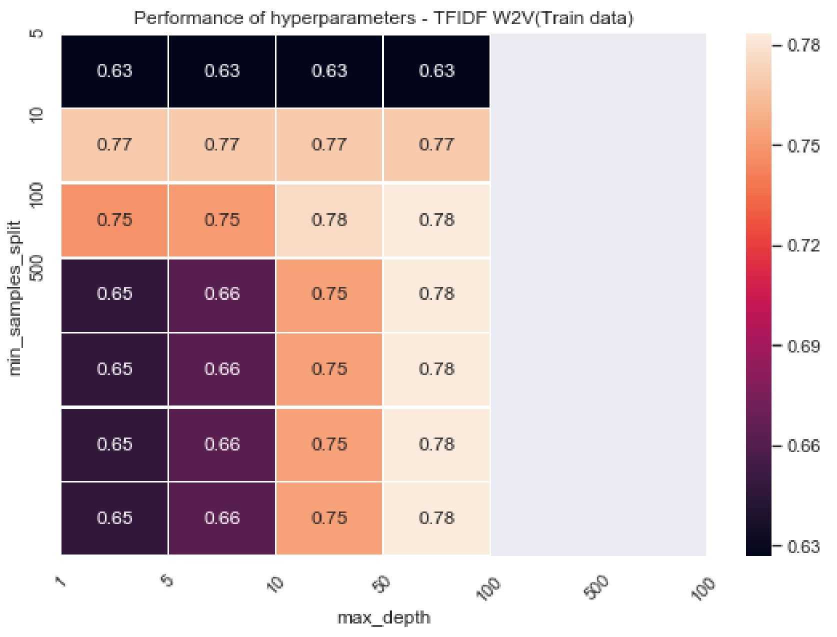
In [45]:

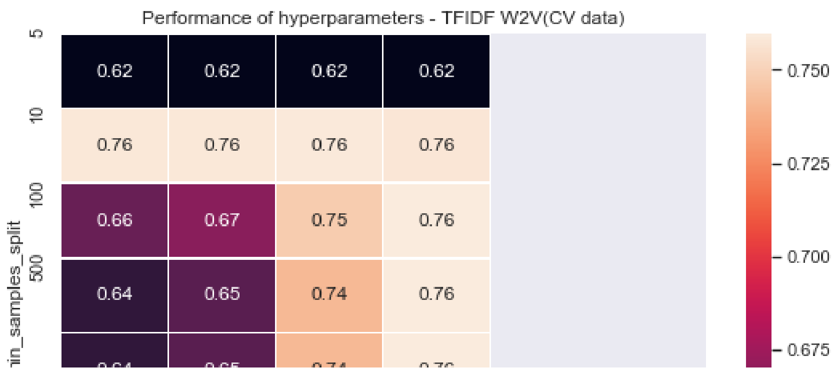
```python
# Please write all the code with proper documentation
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import GridSearchCV
from sklearn import linear_model
param_grid = {"max_depth": [1, 5, 10, 50, 100, 500, 100], "min_samples_split":[5, 10, 100, 500]}
scoring = {'AUC': 'roc_auc', 'Accuracy': 'accuracy', 'Precision' : 'precision', 'Recall' : 'recall'}
clfD=DecisionTreeClassifier(random_state=0)
grid = GridSearchCV(clfD,param_grid=param_grid,scoring = scoring, refit = 'AUC')
grid.fit(tfidf_sent_vectors, y_tr)
#print(grid)
# summarize the results of the grid search
(grid.best_score_)
print(grid.best_estimator_.max_depth)
print(grid.best_estimator_.min_samples_split)
results_tr_tfidfw2v = grid.cv_results_
#print(results)

grid.fit(tfidf_sent_vectors_cv, y_cv)
#print(grid)
# summarize the results of the grid search
(grid.best_score_)
print(grid.best_estimator_.max_depth)
print(grid.best_estimator_.min_samples_split)
results_cv_tfidfw2v = grid.cv_results_
#print(results)
```

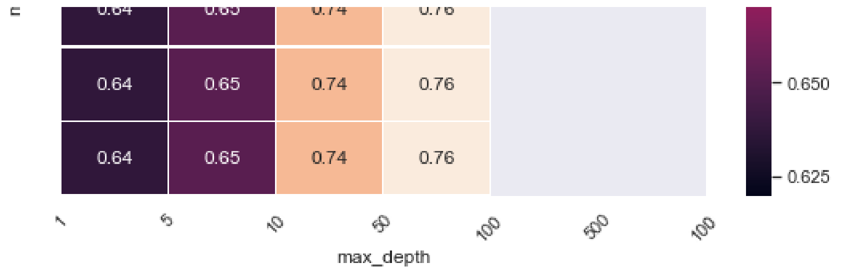
```
10
500
50
500
```

In [46]:

```
import seaborn as sns
max_depth = [1, 5, 10, 50, 100, 500, 100]
min_samples_split = [5, 10, 100, 500]
scores = results_tr_tfidfw2v['mean_test_AUC'].reshape(len(max_depth),len(min_samples_split))
## Plotting Function
#plt.figure(figsize=(8, 6))
sns.set()
#plt.subplots_adjust(left=.2, right=0.95, bottom=0.15, top=0.95)
f, ax = plt.subplots(figsize=(9, 6))
sns.heatmap(scores, annot=True, linewidths=.5, ax=ax)
#plt.imshow(scores, interpolation='nearest', cmap=plt.cm.hot)
plt.xlabel('max_depth')
plt.ylabel('min_samples_split')
#plt.colorbar()
plt.xticks(np.arange(len(max_depth)), max_depth, rotation=45)
plt.yticks(np.arange(len(min_samples_split)), min_samples_split)
plt.title('Performance of hyperparameters - TFIDF W2V(Train data)')
plt.show()


scores = results_cv_tfidfw2v['mean_test_AUC'].reshape(len(max_depth),len(min_samples_split))
## Plotting Function
#plt.figure(figsize=(8, 6))
sns.set()
#plt.subplots_adjust(left=.2, right=0.95, bottom=0.15, top=0.95)
f, ax = plt.subplots(figsize=(9, 6))
sns.heatmap(scores, annot=True, linewidths=.5, ax=ax)
#plt.imshow(scores, interpolation='nearest', cmap=plt.cm.hot)
plt.xlabel('max_depth')
plt.ylabel('min_samples_split')
#plt.colorbar()
plt.xticks(np.arange(len(max_depth)), max_depth, rotation=45)
plt.yticks(np.arange(len(min_samples_split)), min_samples_split)
plt.title('Performance of hyperparameters - TFIDF W2V(CV data)')
plt.show()
```







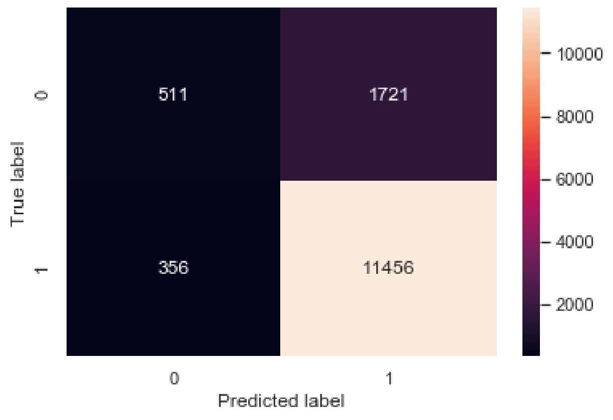
In [49]:

```
#Applying Decision Tree on CV dataset
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import GridSearchCV
from sklearn import linear_model
clf = DecisionTreeClassifier(max_depth =10, min_samples_split=500)
scoring = {'AUC': 'roc_auc'}
cclf = clf.fit(tfidf_sent_vectors,y_tr).predict(tfidf_sent_vectors_cv)
#Caliberate the classifier.
#clf_calibrated=CalibratedClassifierCV(clf, cv='prefit', method='isotonic')
#clf_calibrated.fit(cv_bow, y_cv)
pred_cv = clf.predict_proba(tfidf_sent_vectors_cv)[:,1]
fpr, tpr, thresholds = roc_curve(y_cv,pred_cv)
roc_auc_cv = auc(fpr, tpr)
print('Area under the ROC curve : %f', + roc_auc_cv)

#Plotting confusion matrix
import seaborn as sns
conf_mat = confusion_matrix(y_cv, cclf)
print(conf_mat)
#conf_normalized = conf_mat.astype('int') / conf_mat.sum(axis=1)[:, np.newaxis]
sns.heatmap(conf_mat, annot=True, fmt ='g')
plt.ylabel('True label')
plt.xlabel('Predicted label')
```

```
Area under the ROC curve : %f 0.7983169263503367
[[  511  1721]
 [  356 11456]]
```

Out[49]:

```
Text(0.5, 12.5, 'Predicted label')
```



In [50]:

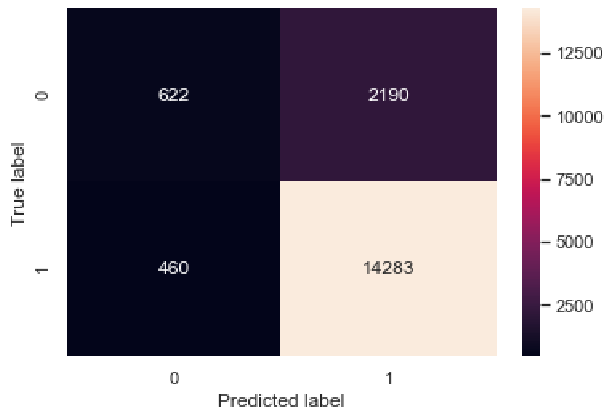
```
#Applying Decision Tree on test dataset
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import GridSearchCV
from sklearn import linear_model
clf = DecisionTreeClassifier(max_depth =10, min_samples_split=500)
scoring = {'AUC': 'roc_auc', 'Accuracy': 'accuracy', 'Precision' : 'precision', 'Recall' : 'recall'}
cclf = clf.fit(tfidf_sent_vectors,y_tr).predict(tfidf_sent_vectors_test)
pred_test = clf.predict_proba(tfidf_sent_vectors_test)[:,1]
fpr, tpr, thresholds = roc_curve(y_test,pred_test)
roc_auc_test = auc(fpr, tpr)
print('Area under the ROC curve : %f', + roc_auc_test)
```

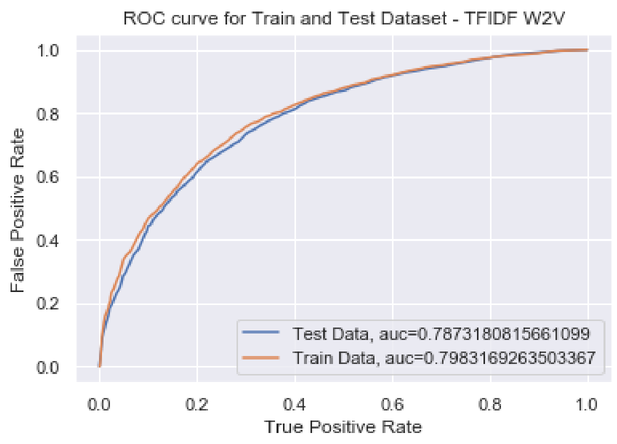
```
#Plotting confusion matrix
import seaborn as sns
conf_mat = confusion_matrix(y_test, cclf)
print(conf_mat)
#conf_normalized = conf_mat.astype('int') / conf_mat.sum(axis=1)[:, np.newaxis]
sns.heatmap(conf_mat, annot=True, fmt ='g')
plt.ylabel('True label')
plt.xlabel('Predicted label')

#Plot ROC Curve
# calculate the fpr and tpr for all thresholds of the classification
#https://stackoverflow.com/questions/25009284/how-to-plot-roc-curve-in-python
plt.figure(0).clf()
fpr, tpr, thresholds = roc_curve(y_test,pred_test)
roc_auc_test_tfidf = auc(fpr, tpr)
plt.plot(fpr,tpr,label="Test Data, auc="+str(roc_auc_test))

fpr, tpr, thresh = roc_curve(y_cv, pred_cv)
roc_auc_cv_tfidf = auc(fpr, tpr)
plt.plot(fpr,tpr,label="Train Data, auc="+str(roc_auc_cv))
plt.title('ROC curve for Train and Test Dataset - TFIDF W2V')
plt.xlabel('True Positive Rate')
plt.ylabel('False Positive Rate')
plt.legend(loc=0)
```

```
Area under the ROC curve : %f 0.7873180815661099
[[  622  2190]
 [  460 14283]]
```

Out[50]:

```
<matplotlib.legend.Legend at 0x4928b550>
```





[6] Conclusions

In [153]:

```
# Please compare all your models using Prettytable library
from prettytable import PrettyTable
```

```
table = PrettyTable(["model","max_depth","min_samples_split","ROC"])
table.add_row(["Decision tree using BoW", "50","500","0.83"])
table.add_row(["Decision tree using TFIDF", "50","500","0.79"])
table.add_row(["Decision tree using AVG W2V", "10","500","0.5"])
table.add_row(["Decision tree using TFIDF W2V", "10","500","0.78"])
print(table)
```

```
+-------------------------------+-----------+-------------------+------+
|             model             | max_depth | min_samples_split | ROC  |
+-------------------------------+-----------+-------------------+------+
|    Decision tree using BoW    |     50    |        500        | 0.83 |
|   Decision tree using TFIDF   |     50    |        500        | 0.79 |
|  Decision tree using AVG W2V  |     10    |        500        | 0.5  |
| Decision tree using TFIDF W2V |     10    |        500        | 0.78 |
+-------------------------------+-----------+-------------------+------+
```

Observation:

```
1. Decision Tree classifier works well with BOW, TFIDF and Weighted TFIDF except AVG W2V.
2. ROC score is 0.5 for AVG W2V and for the rest of the vectorization models, it is above 0.7.
3. Have chosen min_df = 5 in case of TFIDF vectorizer, as it eliminates rare words which occurs
in less than 5 reviews.
4. Though the AUC score is fair enough, false positives are higher. (negative reviews are predic
ted as positive)
5. Have converted dot file to .png format via command line: "dot -Tpng dtreetf.dot -o dtreetf.pn
g"
```