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, roc_auc_score
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
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
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.model_selection import cross_validate
```

In [2]:

```
# using SQLite Table to read data.
con = sqlite3.connect('/Users/puravshah/Downloads/amazon-fine-food-revi
ews/database.sqlite')
```

```
# filtering only positive and negative reviews i.e.
# not taking into consideration those reviews with Score=3
# SELECT * FROM Reviews WHERE Score != 3 LIMIT 500000, will give top 50
0000 data points
# you can change the number to any other number based on your computing
 power

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

filtered_data = pd.read_sql_query(""" SELECT * FROM Reviews WHERE Score
 != 3 LIMIT 50000""", con)

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

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

```
Number of data points in our data (50000, 10)
```

Out[2]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	Helpfulnes
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	Helpfulnes
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1

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	COU

	UserId	ProductId	ProfileName	Time	Score	Text	COU
0	#oc-R115TNMSPFT9I7	B007Y59HVM	Breyton	1331510400	2	Overall its just OK when considering the price...	2
1	#oc-R11D9D7SHXIJB9	B005HG9ET0	Louis E. Emory "hoppy"	1342396800	5	My wife has recurring extreme muscle spasms, u...	3
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

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

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	Helpfuln

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	Helpfuln
0	78445	B000HDL1RQ	AR5J8UI46CURR	Geetha Krishnan	2	2
1	138317	B000HDOPYC	AR5J8UI46CURR	Geetha Krishnan	2	2
2	138277	B000HDOPYM	AR5J8UI46CURR	Geetha Krishnan	2	2
3	73791	B000HDOPZG	AR5J8UI46CURR	Geetha Krishnan	2	2
4	155049	B000PAQ75C	AR5J8UI46CURR	Geetha Krishnan	2	2

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

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

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

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

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

In [8]:
```
#Sorting data according to ProductId in ascending order
sorted_data=filtered_data.sort_values('ProductId', axis=0, ascending=True, inplace=False, kind='quicksort', na_position='last')
```

In [9]:
```
#Deduplication of entries
final=sorted_data.drop_duplicates(subset={"UserId","ProfileName","Time","Text"}, keep='first', inplace=False)
final.shape
```

Out[9]:
```
(46072, 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]:
```
92.144
```

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	Helpfuln
0	64422	B000MIDROQ	A161DK06JJMCYF	J. E. Stephens "Jeanne"	3	1
1	44737	B001EQ55RW	A2V0I904FH7ABY	Ram	3	2

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()
```

```
(46071, 10)
```

Out[13]:
```
1    38479
0     7592
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)
```

```python
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 hard to find any chicken products made in
the USA but they are out there, but this one isnt.  Its too bad too bec
ause its a good product but I wont take any chances till they know what
is going on with the china imports.
==================================================
this is yummy, easy and unusual. it makes a quick, delicous pie, crisp
or cobbler. home made is better, but a heck of a lot more work. this is
great to have on hand for last minute dessert needs where you really wa
nt to impress wih your creativity in cooking! recommended.
==================================================
Great flavor, low in calories, high in nutrients, high in protein! Usua
lly protein powders are high priced and high in calories, this one is a
great bargain and tastes great, I highly recommend for the lady gym rat
s, probably not "macho" enough for guys since it is soy based...
==================================================
For those of you wanting a high-quality, yet affordable green tea, you
should definitely give this one a try. Let me first start by saying tha
t everyone is looking for something different for their ideal tea, and
I will attempt to briefly highlight what makes this tea attractive to a
wide range of tea drinkers (whether you are a beginner or long-time tea
enthusiast).  I have gone through over 12 boxes of this tea myself, and
highly recommend it for the following reasons:<br /><br />-Quality:  Fi
rst, this tea offers a smooth quality without any harsh or bitter after
tones, which often turns people off from many green teas.  I've found m
y ideal brewing time to be between 3-5 minutes, giving you a light but
flavorful cup of tea.  However, if you get distracted or forget about y
```

```
our tea and leave it brewing for 20+ minutes like I sometimes do, the q
uality of this tea is such that you still get a smooth but deeper flavo
r without the bad after taste.  The leaves themselves are whole leaves
(not powdered stems, branches, etc commonly found in other brands), and
the high-quality nylon bags also include chunks of tropical fruit and o
ther discernible ingredients.  This isn't your standard cheap paper bag
with a mix of unknown ingredients that have been ground down to a fine
powder, leaving you to wonder what it is you are actually drinking.<br
/><br />-Taste:  This tea offers notes of real pineapple and other hint
s of tropical fruits, yet isn't sweet or artificially flavored.  You ha
ve the foundation of a high-quality young hyson green tea for those tru
e "tea flavor" lovers, yet the subtle hints of fruit make this a truly
unique tea that I believe most will enjoy.  If you want it sweet, you c
an add sugar, splenda, etc but this really is not necessary as this tea
offers an inherent warmth of flavor through it's ingredients.<br /><br
/>-Price:  This tea offers an excellent product at an exceptional price
(especially when purchased at the prices Amazon offers).  Compared to o
ther brands which I believe to be of similar quality (Mighty Leaf, Rish
i, Two Leaves, etc.), Revolution offers a superior product at an outsta
nding price.  I have been purchasing this through Amazon for less per b
ox than I would be paying at my local grocery store for Lipton, etc.<br
/><br />Overall, this is a wonderful tea that is comparable, and even b
etter than, other teas that are priced much higher.  It offers a well-b
alanced cup of green tea that I believe many will enjoy.  In terms of t
aste, quality, and price, I would argue you won't find a better combina
tion that that offered by Revolution's Tropical Green Tea.
==================================================
```

In [15]:

```
# remove urls from text python: https://stackoverflow.com/a/40823105/40
84039
sent_0 = re.sub(r"http\S+", "", sent_0)
sent_1000 = re.sub(r"http\S+", "", sent_1000)
sent_150 = re.sub(r"http\S+", "", sent_1500)
sent_4900 = re.sub(r"http\S+", "", sent_4900)

print(sent_0)
```

```
My dogs loves this chicken but its a product from China, so we wont be
buying it anymore.  Its very hard to find any chicken products made in
```

the USA but they are out there, but this one isnt. Its too bad too bec
ause its a good product but I wont take any chances till they know what
is going on with the china imports.

In [16]:

```python
# 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)
```

```
My dogs loves this chicken but its a product from China, so we wont be
buying it anymore.  Its very hard to find any chicken products made in
the USA but they are out there, but this one isnt.  Its too bad too bec
ause its a good product but I wont take any chances till they know what
is going on with the china imports.
==================================================
this is yummy, easy and unusual. it makes a quick, delicous pie, crisp
or cobbler. home made is better, but a heck of a lot more work. this is
great to have on hand for last minute dessert needs where you really wa
nt to impress wih your creativity in cooking! recommended.
==================================================
Great flavor, low in calories, high in nutrients, high in protein! Usua
lly protein powders are high priced and high in calories, this one is a
```

great bargain and tastes great, I highly recommend for the lady gym rat
s, probably not "macho" enough for guys since it is soy based...
==
For those of you wanting a high-quality, yet affordable green tea, you
should definitely give this one a try. Let me first start by saying tha
t everyone is looking for something different for their ideal tea, and
I will attempt to briefly highlight what makes this tea attractive to a
wide range of tea drinkers (whether you are a beginner or long-time tea
enthusiast). I have gone through over 12 boxes of this tea myself, and
highly recommend it for the following reasons:-Quality: First, this te
a offers a smooth quality without any harsh or bitter after tones, whic
h often turns people off from many green teas. I've found my ideal bre
wing time to be between 3-5 minutes, giving you a light but flavorful c
up of tea. However, if you get distracted or forget about your tea and
leave it brewing for 20+ minutes like I sometimes do, the quality of th
is tea is such that you still get a smooth but deeper flavor without th
e bad after taste. The leaves themselves are whole leaves (not powdere
d stems, branches, etc commonly found in other brands), and the high-qu
ality nylon bags also include chunks of tropical fruit and other discer
nible ingredients. This isn't your standard cheap paper bag with a mix
of unknown ingredients that have been ground down to a fine powder, lea
ving you to wonder what it is you are actually drinking.-Taste: This t
ea offers notes of real pineapple and other hints of tropical fruits, y
et isn't sweet or artificially flavored. You have the foundation of a
high-quality young hyson green tea for those true "tea flavor" lovers,
yet the subtle hints of fruit make this a truly unique tea that I belie
ve most will enjoy. If you want it sweet, you can add sugar, splenda,
etc but this really is not necessary as this tea offers an inherent war
mth of flavor through it's ingredients.-Price: This tea offers an exce
llent product at an exceptional price (especially when purchased at the
prices Amazon offers). Compared to other brands which I believe to be
of similar quality (Mighty Leaf, Rishi, Two Leaves, etc.), Revolution o
ffers a superior product at an outstanding price. I have been purchasi
ng this through Amazon for less per box than I would be paying at my lo
cal grocery store for Lipton, etc.Overall, this is a wonderful tea that
is comparable, and even better than, other teas that are priced much hi
gher. It offers a well-balanced cup of green tea that I believe many w
ill enjoy. In terms of taste, quality, and price, I would argue you wo

n't find a better combination that that offered by Revolution's Tropica l Green Tea.

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 flavor, low in calories, high in nutrients, high in protein! Usua
lly protein powders are high priced and high in calories, this one is a
great bargain and tastes great, I highly recommend for the lady gym rat
s, probably not "macho" enough for guys since it is soy based...
==================================================
```

In [19]:
```
#remove words with numbers python: https://stackoverflow.com/a/1808237
0/4084039
sent_0 = re.sub("\S*\d\S*", "", sent_0).strip()
print(sent_0)
```

```
My dogs loves this chicken but its a product from China, so we wont be
buying it anymore.  Its very hard to find any chicken products made in
the USA but they are out there, but this one isnt.  Its too bad too bec
```

```
ause its a good product but I wont take any chances till they know what
is going on with the china 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)
```

```
Great flavor low in calories high in nutrients high in protein Usually
protein powders are high priced and high in calories this one is a grea
t bargain and tastes great I highly recommend for the lady gym rats pro
bably not macho enough for guys since it is soy based
```

In [21]:
```
# https://gist.github.com/sebleier/554280
# we are removing the words from the stop words list: 'no', 'nor', 'no
t'
# <br /><br /> ==> after the above steps, we are getting "br br"
# we are including them into stop words list
# instead of <br /> if we have <br/> these tags would have revmoved in
 the 1st step

stopwords= set(['br', 'the', 'i', 'me', 'my', 'myself', 'we', 'our', 'o
urs', 'ourselves', 'you', "you're", "you've",\
            "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselve
s', 'he', 'him', 'his', 'himself', \
            'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'it
s', 'itself', 'they', 'them', 'their',\
            'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'th
is', 'that', "that'll", 'these', 'those', \
            'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'h
ave', 'has', 'had', 'having', 'do', 'does', \
            'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or',
 'because', 'as', 'until', 'while', 'of', \
            'at', 'by', 'for', 'with', 'about', 'against', 'between',
'into', 'through', 'during', 'before', 'after',\
            'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out',
'on', 'off', 'over', 'under', 'again', 'further',\
            'then', 'once', 'here', 'there', 'when', 'where', 'why', 'h
ow', 'all', 'any', 'both', 'each', 'few', 'more',\
```

```
            'most', 'other', 'some', 'such', 'only', 'own', 'same', 's
o', 'than', 'too', 'very', \
            's', 't', 'can', 'will', 'just', 'don', "don't", 'should',
"should've", 'now', 'd', 'll', 'm', 'o', 're', \
            've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't",
'didn', "didn't", 'doesn', "doesn't", 'hadn',\
            "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "is
n't", 'ma', 'mightn', "mightn't", 'mustn',\
            "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn',
 "shouldn't", 'wasn', "wasn't", 'weren', "weren't", \
            'won', "won't", 'wouldn', "wouldn't"])
```

In [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%|██████████| 46071/46071 [00:28<00:00, 1644.87it/s]
```

In [23]:

```
preprocessed_reviews[1500]
```

Out[23]:

```
'great flavor low calories high nutrients high protein usually protein
powders high priced high calories one great bargain tastes great highly
recommend lady gym rats probably not macho enough guys since soy based'
```

[3.2] Preprocessing Review Summary

In [24]:
```
## Similartly you can do preprocessing for review summary also.
```

[4] Featurization

[4.1] BAG OF WORDS

In [25]:
```
#BoW
from sklearn.preprocessing import StandardScaler
scaler=StandardScaler(with_mean=False)
X_1, X_test, y_1, y_test = train_test_split(preprocessed_reviews, final
['Score'], test_size=0.3, random_state=0)
X_tr, X_cv, y_tr, y_cv = train_test_split(X_1, y_1, test_size=0.3)
count_vect = CountVectorizer() #in scikit-learn
count_vect.fit(X_tr)
#d=count_vect.fit(X_cv)
#e=count_vect.fit(X_test)
print("some feature names ", count_vect.get_feature_names()[:10])
print('='*50)

final_counts = count_vect.transform(X_tr)

X_cv_bow=count_vect.transform(X_cv)

X_test_bow=count_vect.transform(X_test)

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])
print(y_test.shape)
```

```
some feature names  ['aa', 'aaa', 'aaaa', 'aaaaaaahhhhhh', 'aaaaaawwwww
wwwww', 'aadp', 'aaf', 'aafco', 'aah', 'aahhhs']
==================================================
the type of count vectorizer  <class 'scipy.sparse.csr.csr_matrix'>
the shape of out text BOW vectorizer  (22574, 28137)
```

```
the number of unique words  28137
(13822,)
```

[4.2] Bi-Grams and n-Grams.

In [26]:
```
#bi-gram, tri-gram and n-gram

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

count_vect_gram = CountVectorizer(ngram_range=(1,2), min_df=10, max_fea
tures=5000)
count_vect_gram.fit(X_tr)
final_bigram_counts =count_vect_gram.transform(X_tr)

X_cv_ngram = count_vect_gram.transform(X_cv)

X_test_ngram=count_vect_gram.transform(X_test)

print("the type of count vectorizer ",type(final_bigram_counts))
print("the shape of out text BOW vectorizer ",final_bigram_counts.get_s
hape())
print("the number of unique words including both unigrams and bigrams "
, final_bigram_counts.get_shape()[1])
```

```
the type of count vectorizer  <class 'scipy.sparse.csr.csr_matrix'>
the shape of out text BOW vectorizer  (22574, 5000)
the number of unique words including both unigrams and bigrams  5000
```

[4.3] TF-IDF

In [27]:
```
tf_idf_vect = TfidfVectorizer(ngram_range=(1,2), min_df=10)
```

```
tf_idf_vect.fit(X_tr)
#g=tf_idf_vect.fit(X_cv)
#h=tf_idf_vect.fit(X_test)
print("some sample features(unique words in the corpus)",tf_idf_vect.ge
t_feature_names()[0:10])
print('='*50)

final_tf_idf = tf_idf_vect.transform(X_tr)

X_cv_tfidf=tf_idf_vect.transform(X_cv)

X_test_tfidf=tf_idf_vect.transform(X_test)

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

```
some sample features(unique words in the corpus) ['ability', 'able', 'a
ble buy', 'able drink', 'able eat', 'able enjoy', 'able find', 'able ge
t', 'able give', 'able make']
==================================================
the type of count vectorizer  <class 'scipy.sparse.csr.csr_matrix'>
the shape of out text TFIDF vectorizer  (22574, 13418)
the number of unique words including both unigrams and bigrams  13418
```

[4.4] Word2Vec

In [28]:

```
i=0
list_of_sentance_train=[]
for sentance in X_tr:
    list_of_sentance_train.append(sentance.split())
```

In [29]:

```
i=0
list_of_sentance_cv=[]
```

```
for sentance in X_cv:
    list_of_sentance_cv.append(sentance.split())
```

In [30]:
```
i=0
list_of_sentance_test=[]
for sentance in X_test:
    list_of_sentance_test.append(sentance.split())
```

In [31]:
```
# Using Google News Word2Vectors

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


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

is_your_ram_gt_16g=False
want_to_use_google_w2v = False
want_to_train_w2v = True

if want_to_train_w2v:
    # min_count = 5 considers only words that occured atleast 5 times
    w2v_model_train=Word2Vec(list_of_sentance_train,min_count=5,size=50
, workers=4)
    print(w2v_model_train.wv.most_similar('great'))
    print('='*50)
```

```
    print(w2v_model_train.wv.most_similar('worst'))

elif want_to_use_google_w2v and is_your_ram_gt_16g:
    if os.path.isfile('GoogleNews-vectors-negative300.bin'):
        w2v_model=KeyedVectors.load_word2vec_format('GoogleNews-vectors
-negative300.bin', binary=True)
        print(w2v_model.wv.most_similar('great'))
        print(w2v_model.wv.most_similar('worst'))
    else:
        print("you don't have gogole's word2vec file, keep want_to_trai
n_w2v = True, to train your own w2v ")
```

```
[('awesome', 0.8476660847663879), ('good', 0.8175162076950073), ('amazi
ng', 0.8048639297485352), ('excellent', 0.7937514781951904), ('wonderfu
l', 0.7830322980880737), ('terrific', 0.7640370726585388), ('fantasti
c', 0.749491274356842), ('perfect', 0.7404717803001404), ('decent', 0.6
895331144332886), ('delicious', 0.6763255000114441)]
==================================================
[('nastiest', 0.8241625428199768), ('closest', 0.820551872253418), ('be
st', 0.7533990144729614), ('terrible', 0.7503485083580017), ('compare
s', 0.7458094358444214), ('ive', 0.7448897957801819), ('hooked', 0.7413
380742073059), ('awful', 0.7411062121391296), ('addicted', 0.7372125387
191772), ('jamaica', 0.7371932864189148)]
```

In [32]:

```
'''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_cv=Word2Vec(list_of_sentance_cv,min_count=5,size=50, work
ers=4)
    print(w2v_model_cv.wv.most_similar('great'))
    print('='*50)
    print(w2v_model_cv.wv.most_similar('worst'))

elif want_to_use_google_w2v and is_your_ram_gt_16g:
    if os.path.isfile('GoogleNews-vectors-negative300.bin'):
        w2v_model=KeyedVectors.load_word2vec_format('GoogleNews-vectors
```

```
-negative300.bin', binary=True)
        print(w2v_model.wv.most_similar('great'))
        print(w2v_model.wv.most_similar('worst'))
    else:
        print("you don't have gogole's word2vec file, keep want_to_trai
n_w2v = True, to train your own w2v ")'''
```

Out[32]:

```
'is_your_ram_gt_16g=False\nwant_to_use_google_w2v = False\nwant_to_trai
n_w2v = True\n\nif want_to_train_w2v:\n    # min_count = 5 considers on
ly words that occured atleast 5 times\n    w2v_model_cv=Word2Vec(list_o
f_sentance_cv,min_count=5,size=50, workers=4)\n    print(w2v_model_cv.w
v.most_similar(\'great\'))\n    print(\'=\'*50)\n    print(w2v_model_c
v.wv.most_similar(\'worst\'))\n    \nelif want_to_use_google_w2v and is
_your_ram_gt_16g:\n    if os.path.isfile(\'GoogleNews-vectors-negative3
00.bin\'):\n        w2v_model=KeyedVectors.load_word2vec_format(\'Googl
eNews-vectors-negative300.bin\', binary=True)\n        print(w2v_model.
wv.most_similar(\'great\'))\n        print(w2v_model.wv.most_similar
(\'worst\'))\n    else:\n        print("you don\'t have gogole\'s word2
vec file, keep want_to_train_w2v = True, to train your own w2v ")'
```

In [33]:

```
'''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_test=Word2Vec(list_of_sentance_test,min_count=5,size=50,
 workers=4)
    print(w2v_model_test.wv.most_similar('great'))
    print('='*50)
    print(w2v_model_test.wv.most_similar('worst'))

elif want_to_use_google_w2v and is_your_ram_gt_16g:
    if os.path.isfile('GoogleNews-vectors-negative300.bin'):
        w2v_model=KeyedVectors.load_word2vec_format('GoogleNews-vectors
-negative300.bin', binary=True)
        print(w2v_model.wv.most_similar('great'))
        print(w2v_model.wv.most_similar('worst'))
    else:
```

```
        print("you don't have gogole's word2vec file, keep want_to_trai
n_w2v = True, to train your own w2v ")'''
```

Out[33]:
```
'is_your_ram_gt_16g=False\nwant_to_use_google_w2v = False\nwant_to_trai
n_w2v = True\n\nif want_to_train_w2v:\n    # min_count = 5 considers on
ly words that occured atleast 5 times\n    w2v_model_test=Word2Vec(list
_of_sentance_test,min_count=5,size=50, workers=4)\n    print(w2v_model_
test.wv.most_similar(\'great\'))\n    print(\'=\'*50)\n    print(w2v_mo
del_test.wv.most_similar(\'worst\'))\n    \nelif want_to_use_google_w2v
and is_your_ram_gt_16g:\n    if os.path.isfile(\'GoogleNews-vectors-neg
ative300.bin\'):\n        w2v_model=KeyedVectors.load_word2vec_format
(\'GoogleNews-vectors-negative300.bin\', binary=True)\n        print(w2
v_model.wv.most_similar(\'great\'))\n        print(w2v_model.wv.most_si
milar(\'worst\'))\n    else:\n        print("you don\'t have gogole\'s
word2vec file, keep want_to_train_w2v = True, to train your own w2v ")'
```

In [34]:
```
w2v_words_train = list(w2v_model_train.wv.vocab)
print("number of words that occured minimum 5 times ",len(w2v_words_tra
in))
print("sample words ", w2v_words_train[0:50])
```

```
number of words that occured minimum 5 times  9111
sample words  ['bought', 'bottle', 'hot', 'sauce', 'friend', 'shipped',
'east', 'coast', 'order', 'said', 'would', 'take', 'business', 'days',
'ended', 'taking', 'weeks', 'really', 'like', 'eggnog', 'especially',
'brandy', 'ones', 'already', 'good', 'wanted', 'try', 'product', 'usual
ly', 'season', 'experience', 'adding', 'real', 'not', 'create', 'tast
e', 'variety', 'improve', 'upon', 'using', 'milk', 'frother', 'poured',
'mixture', 'cup', 'ounce', 'found', 'needed', 'twice', 'much']
```

In [35]:
```
'''w2v_words_cv = list(w2v_model_cv.wv.vocab)
print("number of words that occured minimum 5 times ",len(w2v_words_c
v))
print("sample words ", w2v_words_cv[0:50])'''
```

Out[35]:
```
'w2v_words_cv = list(w2v_model_cv.wv.vocab)\nprint("number of words tha
t occured minimum 5 times ",len(w2v_words_cv))\nprint("sample words ",
w2v_words_cv[0:50])'
```

```
In [36]: '''w2v_words_test = list(w2v_model_test.wv.vocab)
print("number of words that occured minimum 5 times ",len(w2v_words_tes
t))
print("sample words ", w2v_words_test[0:50])'''
```

```
Out[36]: 'w2v_words_test = list(w2v_model_test.wv.vocab)\nprint("number of words
that occured minimum 5 times ",len(w2v_words_test))\nprint("sample word
s ", w2v_words_test[0:50])'
```

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

[4.4.1.1] Avg W2v

```
In [37]: sent_vectors_train = []; # the avg-w2v for each sentence/review is stor
ed in this list
for sent in tqdm(list_of_sentance_train): # for each review/sentence
    sent_vec = np.zeros(50) # as word vectors are of zero length 50, yo
u might need to change this to 300 if you use google's w2v
    cnt_words =0; # num of words with a valid vector in the sentence/re
view
    for word in sent: # for each word in a review/sentence
        if word in w2v_words_train:
            vec = w2v_model_train.wv[word]
            sent_vec += vec
            cnt_words += 1
    if cnt_words != 0:
        sent_vec /= cnt_words
    sent_vectors_train.append(sent_vec)
print(len(sent_vectors_train))
print(len(sent_vectors_train[0]))
#sent_vectors_train=scaler.fit_transform(sent_vectors_train)
```

```
100%|██████████| 22574/22574 [00:53<00:00, 418.78it/s]
```

```
22574
50
```

In [38]:
```
sent_vectors_cv = []; # the avg-w2v for each sentence/review is stored
 in this list
for sent in tqdm(list_of_sentance_cv): # for each review/sentence
    sent_vec = np.zeros(50) # as word vectors are of zero length 50, yo
u might need to change this to 300 if you use google's w2v
    cnt_words =0; # num of words with a valid vector in the sentence/re
view
    for word in sent: # for each word in a review/sentence
        if word in w2v_words_train:
            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]))
#sent_vectors_cv=scaler.fit_transform(sent_vectors_cv)
```

```
100%|██████████| 9675/9675 [00:24<00:00, 392.68it/s]
```

```
9675
50
```

In [39]:
```
sent_vectors_test = []; # the avg-w2v for each sentence/review is store
d in this list
for sent in tqdm(list_of_sentance_test): # for each review/sentence
    sent_vec = np.zeros(50) # as word vectors are of zero length 50, yo
u might need to change this to 300 if you use google's w2v
    cnt_words =0; # num of words with a valid vector in the sentence/re
view
    for word in sent: # for each word in a review/sentence
        if word in w2v_words_train:
            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]))
#sent_vectors_test=scaler.fit_transform(sent_vectors_test)
```

```
100%|██████████| 13822/13822 [00:30<00:00, 450.65it/s]
```

```
13822
50
```

[4.4.1.2] TFIDF weighted W2v

In [40]:

```
# S = ["abc def pqr", "def def def abc", "pqr pqr def"]
model_train= TfidfVectorizer()
model_train.fit(X_tr)
tf_idf_matrix_train =model_train.transform(X_tr)
# we are converting a dictionary with word as a key, and the idf as a v
alue
dictionary_train = dict(zip(model_train.get_feature_names(), list(model
_train.idf_)))
```

In [41]:

```
'''# S = ["abc def pqr", "def def def abc", "pqr pqr def"]
#model_cv= TfidfVectorizer()
tf_idf_matrix_cv = model_train.transform(X_cv)
# we are converting a dictionary with word as a key, and the idf as a v
alue
dictionary_cv= dict(zip(model_train.get_feature_names(), list(model_tra
in.idf_)))'''
```

Out[41]:

```
'# S = ["abc def pqr", "def def def abc", "pqr pqr def"]\n#model_cv= Tf
idfVectorizer()\ntf_idf_matrix_cv = model_train.transform(X_cv)\n# we a
re converting a dictionary with word as a key, and the idf as a value\n
dictionary_cv= dict(zip(model_train.get_feature_names(), list(model_tra
in.idf_)))'
```

In [42]:

```
'''# S = ["abc def pqr", "def def def abc", "pqr pqr def"]
```

```
#model_test= TfidfVectorizer()
tf_idf_matrix_test = model_train.transform(X_test)
# we are converting a dictionary with word as a key, and the idf as a v
alue
dictionary_test= dict(zip(model_train.get_feature_names(), list(model_t
rain.idf_)))'''
```

Out[42]:

```
'# S = ["abc def pqr", "def def def abc", "pqr pqr def"]\n#model_test=
TfidfVectorizer()\ntf_idf_matrix_test = model_train.transform(X_test)\n
# we are converting a dictionary with word as a key, and the idf as a v
alue\ndictionary_test= dict(zip(model_train.get_feature_names(), list(m
odel_train.idf_)))'
```

In [43]:

```
# TF-IDF weighted Word2Vec
tfidf_feat_train = model_train.get_feature_names() # tfidf words/col-na
mes
# final_tf_idf is the sparse matrix with row= sentence, col=word and ce
ll_val = tfidf

tfidf_sent_vectors_train = []; # the tfidf-w2v for each sentence/review
 is stored in this list
row=0;
for sent in tqdm(list_of_sentance_train): # for each review/sentence
    sent_vec_train = np.zeros(50) # as word vectors are of zero length
    weight_sum_train =0; # num of words with a valid vector in the sent
ence/review
    for word in sent: # for each word in a review/sentence
        if word in w2v_words_train and word in tfidf_feat_train:
            vec_train = 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_train[word]*(sent.count(word)/len(sent))
            sent_vec_train += (vec * tf_idf)
            weight_sum_train += tf_idf
    if weight_sum_train != 0:
        sent_vec_train /= weight_sum_train
    tfidf_sent_vectors_train.append(sent_vec_train)
```

```
    row += 1
#tfidf_sent_vectors_train =scaler.fit_transform(tfidf_sent_vectors_trai
n)
```

```
100%|██████████| 22574/22574 [04:25<00:00, 85.15it/s]
```

In [44]:

```
tfidf_feat_cv = model_train.get_feature_names() # tfidf words/col-names
# final_tf_idf is the sparse matrix with row= sentence, col=word and ce
ll_val = tfidf

tfidf_sent_vectors_cv = []; # the tfidf-w2v for each sentence/review is
 stored in this list
row=0;
for sent in tqdm(list_of_sentance_cv): # for each review/sentence
    sent_vec_cv = np.zeros(50) # as word vectors are of zero length
    weight_sum_cv =0; # num of words with a valid vector in the sentenc
e/review
    for word in sent: # for each word in a review/sentence
        if word in w2v_words_train and word in tfidf_feat_cv:
            vec_cv = 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_train[word]*(sent.count(word)/len(sent
))
            sent_vec_cv += (vec * tf_idf)
            weight_sum_cv += tf_idf
    if weight_sum_cv != 0:
        sent_vec_cv /= weight_sum_cv
    tfidf_sent_vectors_cv.append(sent_vec_cv)
    row += 1
#tfidf_sent_vectors_cv =scaler.fit_transform(tfidf_sent_vectors_cv)
```

```
100%|██████████| 9675/9675 [01:55<00:00, 83.43it/s]
```

In [45]:

```
tfidf_feat_test = model_train.get_feature_names() # tfidf words/col-nam
es
# final_tf_idf is the sparse matrix with row= sentence, col=word and ce
```

```
ll_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_sentance_test): # for each review/sentence
    sent_vec_test = np.zeros(50) # as word vectors are of zero length
    weight_sum_test =0; # num of words with a valid vector in the sente
nce/review
    for word in sent: # for each word in a review/sentence
        if word in w2v_words_train and word in tfidf_feat_test:
            vec_test = 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_train[word]*(sent.count(word)/len(sent
))
            sent_vec_test += (vec * tf_idf)
            weight_sum_test += tf_idf
    if weight_sum_test != 0:
        sent_vec_test /= weight_sum_test
    tfidf_sent_vectors_test.append(sent_vec_test)
    row += 1
```

```
100%|██████████| 13822/13822 [02:29<00:00, 92.68it/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.](#)

Applying Decision Trees

[5.1] Applying Decision Trees on BOW, SET 1

In [46]:
```
# Please write all the code with proper documentation
```

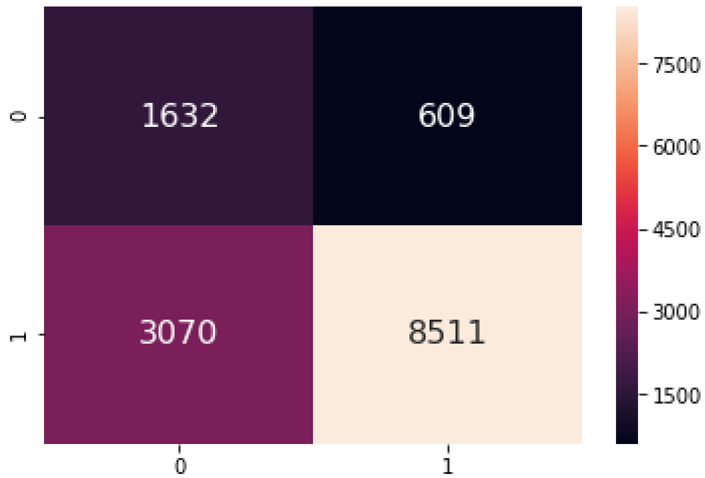
```
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import GridSearchCV
tuned_parameters=[{'max_depth':[1,5,10,50,100,500,1000],'min_samples_sp
lit':[5,10,100,500]}]
model=GridSearchCV(DecisionTreeClassifier(),tuned_parameters,scoring='r
oc_auc',cv=10)
model.fit(final_counts,y_tr)
print(model.best_estimator_)
optimal_depth=model.best_estimator_.max_depth
optimal_samples_split=model.best_estimator_.min_samples_split
print(model.score(X_cv_bow,y_cv))
```

```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=5
0,
                       max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=No
ne,
                       min_samples_leaf=1, min_samples_split=500,
                       min_weight_fraction_leaf=0.0, presort=False,
                       random_state=None, splitter='best')
0.8027993424730238
```

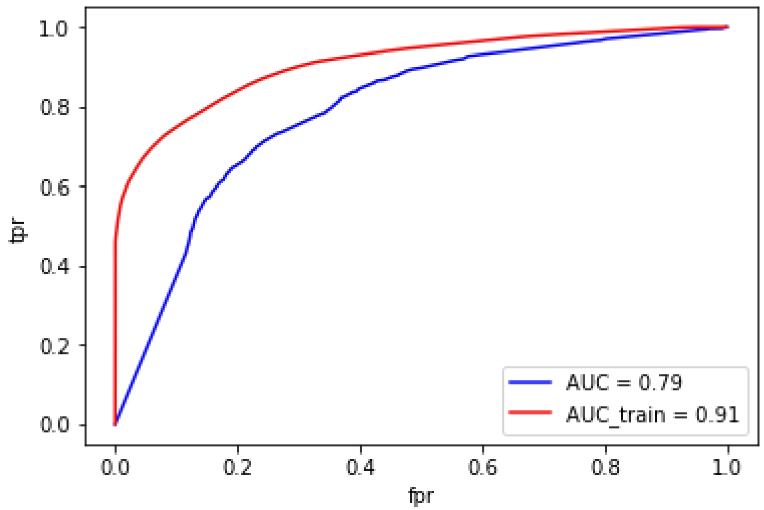
In [47]:

```
model_new=DecisionTreeClassifier(max_depth=optimal_depth,min_samples_sp
lit=optimal_samples_split,class_weight='balanced')
model_new.fit(final_counts,y_tr)
pred=model_new.predict(X_test_bow)
probab=model_new.predict_proba(X_test_bow)[:,1]
fpr,tpr,thresholds=roc_curve(y_test,probab)
auc=roc_auc_score(y_test,probab)
acc=accuracy_score(y_test,pred,normalize=True)*float(100)
pred2=model_new.predict(final_counts)
probab2=model_new.predict_proba(final_counts)[:,1]
fpr_train,tpr_train,thresholds=roc_curve(y_tr,probab2)
auc2=roc_auc_score(y_tr,probab2)
df_cm = pd.DataFrame(confusion_matrix(y_test, pred), range(2),range(2))
    #heatman for visualization of matrix
sns.heatmap(df_cm, annot=True,annot_kws={"size": 16}, fmt='g')
plt.show()
plt.plot(fpr,tpr,'b', label = 'AUC = %0.2f' %auc)
```

```
plt.plot(fpr_train,tpr_train,'r', label = 'AUC_train = %0.2f' %auc2)
plt.legend(loc='lower right')
plt.xlabel('fpr')
plt.ylabel('tpr')
```



Out[47]: Text(0,0.5,'tpr')



[5.1.1] Top 20 important features from SET 1

In [48]:

```
# Please write all the code with proper documentation
feature_importance=model_new.feature_importances_
print(len(feature_importance))
bow=count_vect.get_feature_names()
#print(len(bow))
#print(feature_importance.shape)
df=pd.DataFrame([feature_importance],columns=bow)
df=df.T
#df=df[0].sort_values(ascending=False)

#print(df.head(10))
print('The top twenty positive features are-\n',df[0].sort_values(ascen
ding=False)[0:20])
```

```
28137
The top twenty positive features are-
 not             0.161161
great           0.079406
best            0.043125
delicious       0.040122
love            0.038607
good            0.022855
loves           0.022656
perfect         0.019593
disappointed    0.016798
nice            0.014870
favorite        0.013938
excellent       0.013549
wonderful       0.011750
awful           0.010093
would           0.009274
disappointing   0.008572
bad             0.008084
```

```
changed          0.007789
highly           0.007557
tasty            0.006656
Name: 0, dtype: float64
```

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

In [49]:

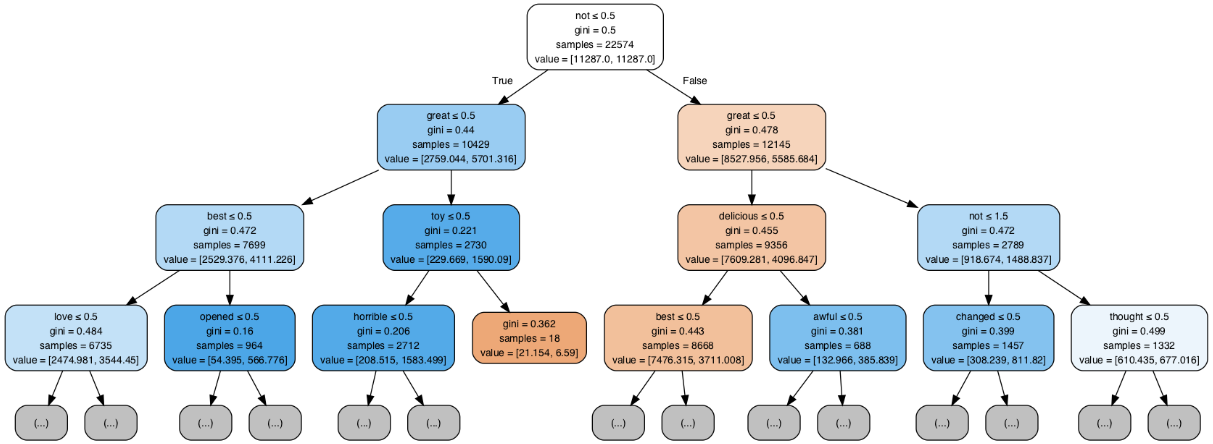
```
# Please write all the code with proper documentation
import pydotplus
from sklearn import tree
from IPython.display import Image
from IPython.display import SVG
from graphviz import Source
from IPython.display import display

#target = ['negative','positive']
# Create DOT data
data =tree.export_graphviz(model_new,out_file=None,max_depth=3,feature_names=bow,filled=True,rounded=True,special_characters=True)

# Draw graph
graph = pydotplus.graph_from_dot_data(data)
#graph = Source(data)

# Show graph
Image(graph.create_png())
```

Out[49]:



[5.2] Applying Decision Trees on TFIDF, SET 2

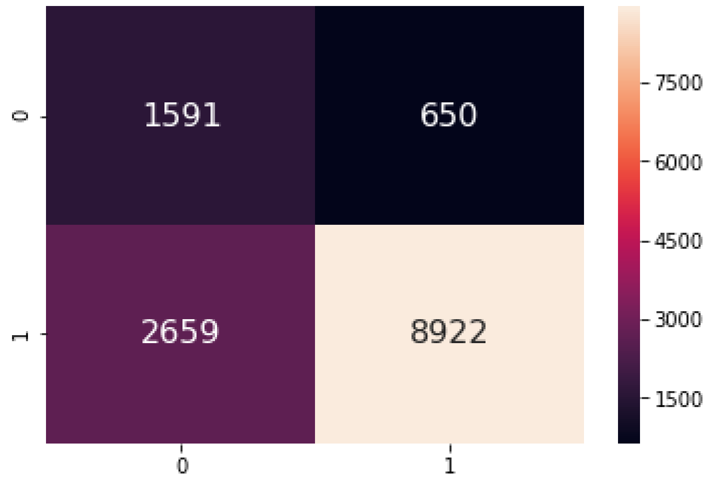
In [50]:

```
# Please write all the code with proper documentation
# Please write all the code with proper documentation
from sklearn.tree import DecisionTreeClassifier
#from sklearn.grid_search import GridSearchCV
tuned_parameters=[{'max_depth':[1,5,10,50,100,500,1000],'min_samples_sp
lit':[5,10,100,500]}]
model=GridSearchCV(DecisionTreeClassifier(),tuned_parameters,scoring='r
oc_auc',cv=10)
model.fit(final_tf_idf,y_tr)
print(model.best_estimator_)
optimal_depth=model.best_estimator_.max_depth
optimal_samples_split=model.best_estimator_.min_samples_split
print(model.score(X_cv_tfidf,y_cv))
```

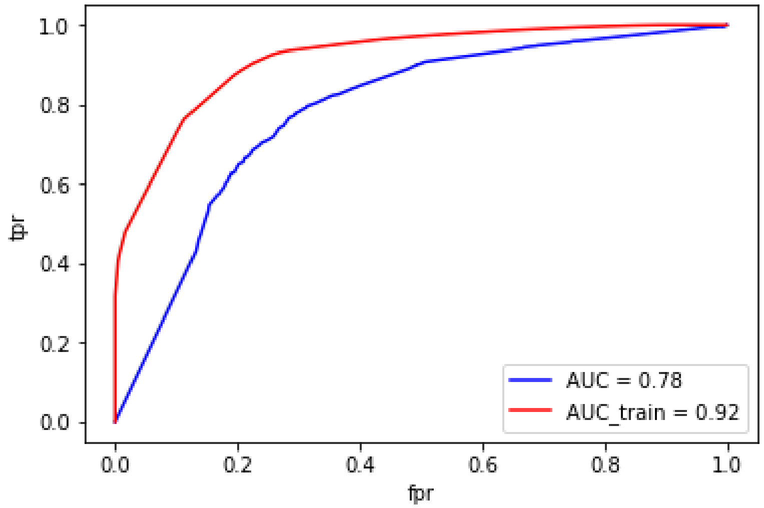
```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=5
0,
                       max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=No
ne,
                       min_samples_leaf=1, min_samples_split=500,
                       min_weight_fraction_leaf=0.0, presort=False,
                       random_state=None, splitter='best')
```

0.7992498700794511

In [51]:
```
model_new=DecisionTreeClassifier(max_depth=optimal_depth,min_samples_sp
lit=optimal_samples_split,class_weight='balanced')
model_new.fit(final_tf_idf,y_tr)
pred=model_new.predict(X_test_tfidf)
probab=model_new.predict_proba(X_test_tfidf)[:,1]
fpr,tpr,thresholds=roc_curve(y_test,probab)
auc=roc_auc_score(y_test,probab)
acc=accuracy_score(y_test,pred,normalize=True)*float(100)
pred2=model.predict(final_tf_idf)
probab2=model.predict_proba(final_tf_idf)[:,1]
fpr_train,tpr_train,thresholds=roc_curve(y_tr,probab2)
auc2=roc_auc_score(y_tr,probab2)
df_cm = pd.DataFrame(confusion_matrix(y_test, pred), range(2),range(2))
    #heatman for visualization of matrix
sns.heatmap(df_cm, annot=True,annot_kws={"size": 16}, fmt='g')
plt.show()
plt.plot(fpr,tpr,'b', label = 'AUC = %0.2f' %auc)
plt.plot(fpr_train,tpr_train,'r', label = 'AUC_train = %0.2f' %auc2)
plt.legend(loc='lower right')
plt.xlabel('fpr')
plt.ylabel('tpr')
```



Out[51]: Text(0,0.5,'tpr')



[5.2.1] Top 20 important features from SET 2

In [52]:
```
# Please write all the code with proper documentation
# Please write all the code with proper documentation
feature_importance=model_new.feature_importances_
tfidf=tf_idf_vect.get_feature_names()
#print(len(bow))
#print(feature_importance.shape)
df=pd.DataFrame([feature_importance],columns=tfidf)
df=df.T
#df=df[0].sort_values(ascending=False)

#print(df.head(10))
print('The top twenty positive features are-\n',df[0].sort_values(ascen
ding=False)[0:20])
```

```
The top twenty positive features are-
 not             0.156930
great           0.078032
best            0.042751
```

```
delicious       0.040500
love            0.038557
good            0.027655
perfect         0.023408
loves           0.021877
favorite        0.019109
disappointed    0.014801
nice            0.014738
excellent       0.013283
wonderful       0.009313
morning         0.008961
easy            0.008490
amazing         0.007575
product         0.007543
awful           0.007159
tasty           0.006951
taste           0.006945
Name: 0, dtype: float64
```

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

In [53]:

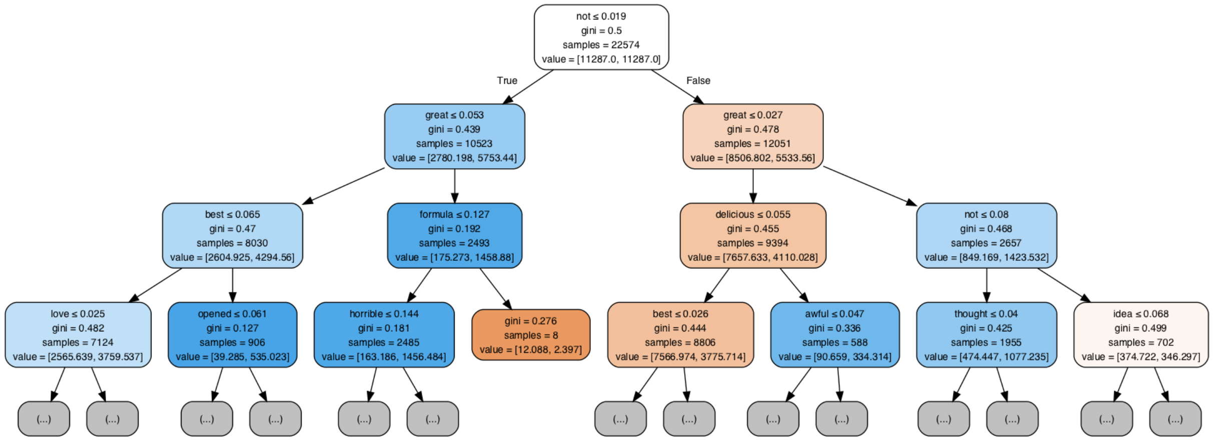
```
# Please write all the code with proper documentation
import pydotplus
from sklearn import tree
from IPython.display import Image
from IPython.display import SVG
from graphviz import Source
from IPython.display import display

target = ['negative','positive']
# Create DOT data
data =tree.export_graphviz(model_new,out_file=None,max_depth=3,feature_
names=tfidf,filled=True,rounded=True,special_characters=True)

# Draw graph
graph = pydotplus.graph_from_dot_data(data)
#graph = Source(data)
```

```
# Show graph
Image(graph.create_png())
```

Out[53]:



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

In [54]:

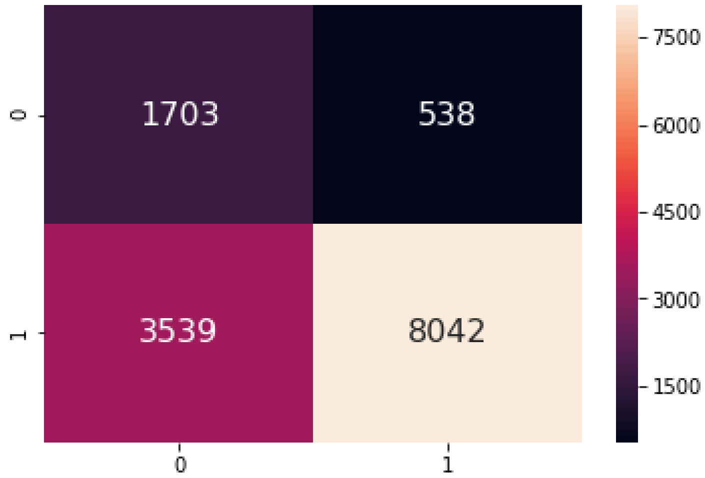
```
# Please write all the code with proper documentation# Please write all
 the code with proper documentation
# Please write all the code with proper documentation
from sklearn.tree import DecisionTreeClassifier
#from sklearn.grid_search import GridSearchCV
tuned_parameters=[{'max_depth':[1,5,10,50,100,500,1000],'min_samples_sp
lit':[5,10,100,500]}]
model=GridSearchCV(DecisionTreeClassifier(),tuned_parameters,scoring='r
oc_auc',cv=10)
model.fit(sent_vectors_train,y_tr)
print(model.best_estimator_)
optimal_depth=model.best_estimator_.max_depth
optimal_samples_split=model.best_estimator_.min_samples_split
print(model.score(sent_vectors_cv,y_cv))
```

```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=1
0,
                       max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=No
ne,
```

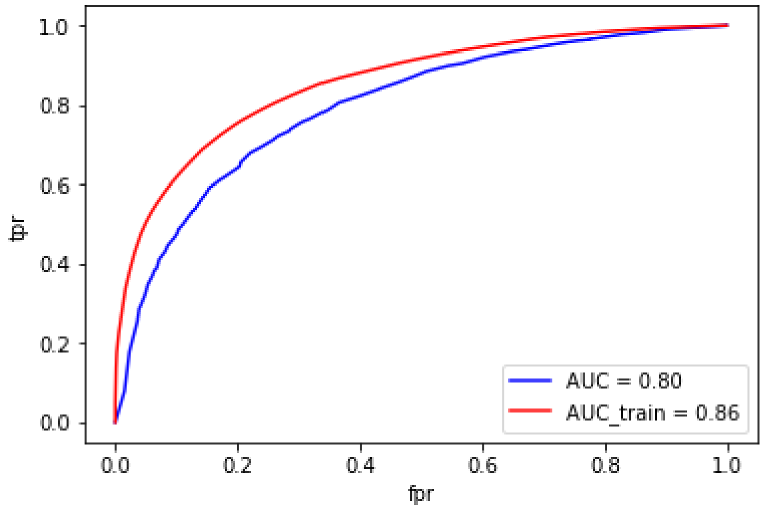
```
ne,
                              min_samples_leaf=1, min_samples_split=500,
                              min_weight_fraction_leaf=0.0, presort=False,
                              random_state=None, splitter='best')
0.8030289685966983
```

In [56]:

```
model_new=DecisionTreeClassifier(max_depth=optimal_depth,min_samples_split=optimal_samples_split,class_weight='balanced')
model_new.fit(sent_vectors_train,y_tr)
pred=model_new.predict(sent_vectors_test)
probab=model_new.predict_proba(sent_vectors_test)[:,1]
fpr,tpr,thresholds=roc_curve(y_test,probab)
auc=roc_auc_score(y_test,probab)
acc=accuracy_score(y_test,pred,normalize=True)*float(100)
pred2=model.predict(sent_vectors_train)
probab2=model.predict_proba(sent_vectors_train)[:,1]
fpr_train,tpr_train,thresholds=roc_curve(y_tr,probab2)
auc2=roc_auc_score(y_tr,probab2)
df_cm = pd.DataFrame(confusion_matrix(y_test, pred), range(2),range(2))
    #heatman for visualization of matrix
sns.heatmap(df_cm, annot=True,annot_kws={"size": 16}, fmt='g')
plt.show()
plt.plot(fpr,tpr,'b', label = 'AUC = %0.2f' %auc)
plt.plot(fpr_train,tpr_train,'r', label = 'AUC_train = %0.2f' %auc2)
plt.legend(loc='lower right')
plt.xlabel('fpr')
plt.ylabel('tpr')
```



Out[56]: Text(0,0.5,'tpr')



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

In [57]:
```
# Please write all the code with proper documentation
# Please write all the code with proper documentation# Please write all
```

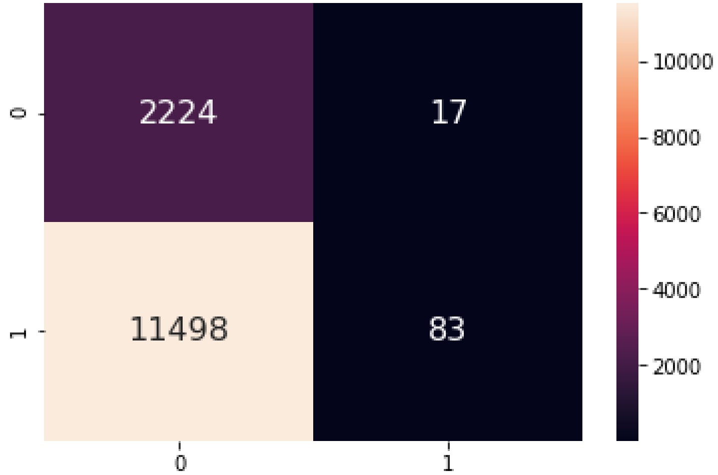
```
 the code with proper documentation
# Please write all the code with proper documentation
from sklearn.tree import DecisionTreeClassifier
#from sklearn.grid_search import GridSearchCV
tuned_parameters=[{'max_depth':[1,5,10,50,100,500,1000],'min_samples_sp
lit':[5,10,100,500]}]
model=GridSearchCV(DecisionTreeClassifier(),tuned_parameters,scoring='r
oc_auc',cv=10)
model.fit(tfidf_sent_vectors_train,y_tr)
print(model.best_estimator_)
optimal_depth=model.best_estimator_.max_depth
optimal_samples_split=model.best_estimator_.min_samples_split
print(model.score(tfidf_sent_vectors_cv,y_cv))
```

```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=1
0,
                       max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=No
ne,
                       min_samples_leaf=1, min_samples_split=100,
                       min_weight_fraction_leaf=0.0, presort=False,
                       random_state=None, splitter='best')
0.5005674809787717
```

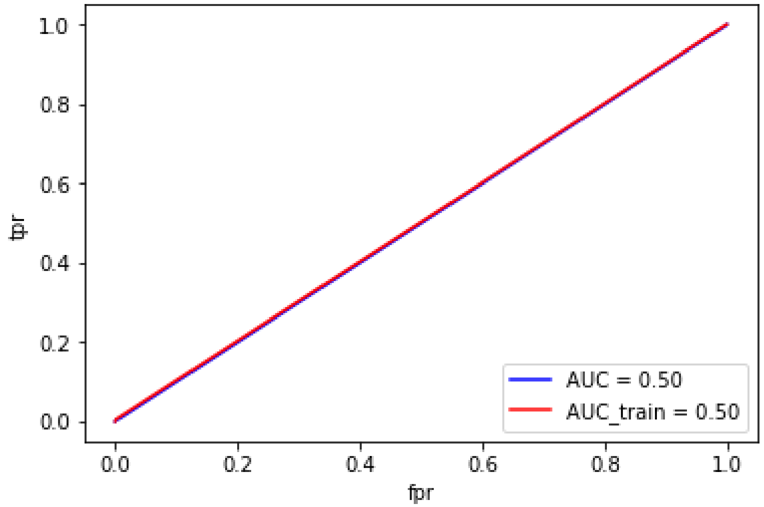
In [58]:

```
model_new=DecisionTreeClassifier(max_depth=optimal_depth,min_samples_sp
lit=optimal_samples_split,class_weight='balanced')
model_new.fit(tfidf_sent_vectors_train,y_tr)
pred=model_new.predict(tfidf_sent_vectors_test)
probab=model_new.predict_proba(tfidf_sent_vectors_test)[:,1]
fpr,tpr,thresholds=roc_curve(y_test,probab)
auc=roc_auc_score(y_test,probab)
acc=accuracy_score(y_test,pred,normalize=True)*float(100)
pred2=model.predict(tfidf_sent_vectors_train)
probab2=model.predict_proba(tfidf_sent_vectors_train)[:,1]
fpr_train,tpr_train,thresholds=roc_curve(y_tr,probab2)
auc2=roc_auc_score(y_tr,probab2)
df_cm = pd.DataFrame(confusion_matrix(y_test, pred), range(2),range(2))
    #heatman for visualization of matrix
sns.heatmap(df_cm, annot=True,annot_kws={"size": 16}, fmt='g')
```

```
plt.show()
plt.plot(fpr,tpr,'b', label = 'AUC = %0.2f' %auc)
plt.plot(fpr_train,tpr_train,'r', label = 'AUC_train = %0.2f' %auc2)
plt.legend(loc='lower right')
plt.xlabel('fpr')
plt.ylabel('tpr')
```



Out[58]: Text(0,0.5,'tpr')



[6] Conclusions

In [59]:

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

x=PrettyTable()
x.field_names=(['Vectorizer','Depth','Min samples split','AUC'])
x.add_row(['BOW',50,500,0.79])
x.add_row(['TFIDF',50,500,0.78])
x.add_row(['AVG W2V',10,500,0.80])
x.add_row(['Weighted TFIDF W2V',10,100,0.50])
print(x)
```

```
+--------------------+-------+-------------------+------+
|     Vectorizer     | Depth | Min samples split | AUC  |
+--------------------+-------+-------------------+------+
|        BOW         |   50  |        500        | 0.79 |
|       TFIDF        |   50  |        500        | 0.78 |
|      AVG W2V       |   10  |        500        | 0.8  |
| Weighted TFIDF W2V |   10  |        100        | 0.5  |
+--------------------+-------+-------------------+------+
```