Assignment 2 Report
COL774

Shivanshu Verma
(2015CH70186)

Question 1 : Text Classification

a) Test data set accuracy = 0.6064404193900597
Training data set accuracy = 0.6590118757384944

b) Test accuracy by random guessing = 0.2008031828175713
Test accuracy by majority prediction = 0.43885452968186783
i.e the probability of most occurred class
Improvement over randomly guessing accuracy = 202%
Improvement over majority prediction accuracy = 38%

c) Confusion matrix for test data set is :

```
[[14985  2800  1190   708   486]
 [ 3195  2435  3008  1697   503]
 [ 1540  1143  4055  6664  1129]
 [ 1136   392  1541 18504  7785]
 [ 3006   168   367 14168 41113]]
```

5 star has the highest value of diagonal entry which means predictions for 5 star is most accurate than that of other ratings.
Confusion matrix is the matrix between expected and predicted values. Diagonal value must be higher for correct prediction and other values within a tuple must be zero for 100% accuracy. Here we can find out percent accuracy of each class by finding the ratio of diagonal value vs sum of row.

d) i) Removing stopwords only:
Accuracy over training data set = 0.6650899654496777
Accuracy over test data set = 0.6076519241986867
ii) Removing stopwords and Stemming words:
Accuracy over training data set = 0.6546257796257796
Accuracy over test data set = 0.6027311207167322

From the above observations, it was found that stemming led to decrease in accuracy whereas stopwords removal increased the accuracy. On test dataset, the accuracy was decreased for stemming by 0.37% and increased by 0.12% for stopwords removal using a part accuracy as the basis.

e) In this part, I have used bigrams and trigrams as the two alternative features for feature engineering. The top model obtained in part d was only stopwords removal.Therefore, I would stick to that for both bigrams and trigrams.
Results for bigram are as follows :
Test dataset accuracy = 0.6340507635471664
Results for trigram are as follows :
Test dataset accuracy = 0.4947651026787717
Trigram led to decrease in accuracy upto 11% where bigram led to increase in accuracy upto 3.5 %.
Therefore, bigrams helped in improving overall accuracy.

f) F1 Score for each class is : [0.74733144 0.1620865 0.22235191 0.48623454 0.79533437]
F1 Score on macro average basis is : 0.48266775263019096
For this kind of dataset, F1 score is more useful than accuracy as this score takes both false positives and false negatives into account.

g) Best performing model : bigram with removed stopwords
Test accuracy = 0.8056806114360071
F1 score = [0.84045742 0.62415505 0.65695408 0.71805254 0.89422037]
Macro avg = 0.746767892
Observation was that, more no of points, helped in improving accuracy.

'Confusion matrix for test data set : '
[[14985 2800 1190   708   486]
 [3195 2435 3008 1697   503]
 [1540 1143 4055 6664 1129]
 [1136   392 1541 18504 7785]
 [3006   168   367 14168 41113]]
accuracy over test dataset is 0.6064404193900597

Confusion matrix for training data set :
[[61309   9851   4638   2762   1731]
 [12076 13611   9516   6644   1696]
 [6219   3215 23714 21255   4310]
 [4623   1219   3054 81546 27152]
 [10645   467   1329 49983 172307]]
accuracy over training dataset is 0.6590118757384944

accuracy over test dataset for random guess is 0.2008031828175713

Stemmed and stopwords:
 train.json
test.json
d
accuracy over test dataset is 0.6027311207167322
Time taken : 2133.1685252189636

Stopwords:
train.json
test.json
d
Enter 1 for stemmed and stopwords removed vocabulary
Enter 2 for only stopwords removed vocabulary>? 2
here
accuracy over test dataset is 0.6076519241986867
Time taken : 125.91505813598633

train.json
test.json
d
Enter 1 for stemmed and stopwords removed vocabulary
Enter 2 for only stopwords removed vocabulary>? 2
accuracy over training dataset is 0.6650899654496777
Time taken : 572.5375080108643

train.json
test.json
d
Enter 1 for stemmed and stopwords removed vocabulary
Enter 2 for only stopwords removed vocabulary>? 1
accuracy over training dataset is 0.6546257796257796
Time taken : 8715.907351970673

train.json
test.json
e
Accuracy over test dataset using bigram is 0.6340507635471664
Time taken : 546.430860042572

train.json
test.json
e
Enter 1 for bigrams and stopwords removed vocabulary

Enter 2 for trigrams and stopwords removed vocabulary>? 2
Accuracy over test dataset using trigram is 0.4947651026787717
Time taken : 562.3820190429688

runfile('/Users/shivanshu/PycharmProjects/COL774/Assignment2/A2Q1_g.py',
wdir='/Users/shivanshu/PycharmProjects/COL774/Assignment2')
train_full.json
test.json
g
6722.342654943466
Accuracy over test dataset using trigram is 0.8056806114360071
Time taken : 473.85395193099976