Machine Learning
Assignment 4
Prasad Pande (psp150030)

Following is the program output for following text categories.

```
> summary(analytics)
ENSEMBLE SUMMARY

       n-ENSEMBLE COVERAGE n-ENSEMBLE RECALL
n >= 1                1.00              0.89
n >= 2                1.00              0.89
n >= 3                0.94              0.92
n >= 4                0.79              0.96
n >= 5                0.59              0.98


ALGORITHM PERFORMANCE

       SVM_PRECISION          SVM_RECALL          SVM_FSCORE LOGITBOOST_PRECISION
               0.868               0.870               0.866                0.780
   LOGITBOOST_RECALL    LOGITBOOST_FSCORE     GLMNET_PRECISION        GLMNET_RECALL
               0.764               0.764               0.860                0.844
       GLMNET_FSCORE       TREE_PRECISION          TREE_RECALL          TREE_FSCORE
               0.846               0.800               0.678                0.700
MAXENTROPY_PRECISION    MAXENTROPY_RECALL    MAXENTROPY_FSCORE
               0.894               0.894               0.892
> |
```

Ensemble summary refers to whether n different algorithms make the same prediction concerning the class of a particular test data event.

In our summary we tested the same for different n values using RTextTools package. Summary table consists of 2 columns Coverage which tells us that percentage of documents that matches the criteria of the recall threshold. As we can see from the summary table for n >= 3 we are getting the maximum coverage that means maximum data points are over the threshold of 0.92 for n=3.

Precision tells us how much confidence we have on relevancy of our classifier result. More precision gives us more confidence. Of all the predicted values, how much of the values actually has that predicted class label gives us precision. Here for

maximum entropy classifier we are getting maximum precision which is good for us.

Recall tells us the sensitivity of the classifier. Out of total true class labels, how much fraction of class labels we predicted correctly tells us the recall. Higher is the recall better is the classifier. For MaxEnt model, recall is high.

Therefore, based on the precision and the recall values we can say that for the given dataset, among the 6 classifiers we evaluated MaxEnt classifier gives us the best performance in terms of the precision and recall.

Value of F-score is based on both precision and recall. We used F-score because precision and recall are biased parameters. Precision and recall are more biased terms. With high precision and recall, MaxEnt classifier has maximum F-score.