WEB MINING HW - 4

- Compared with the model in Q1, how is the performance improved on the test dataset?

  1) Compared to Q1, in Q2 we can see an improvement in performance. There is an increase in the AUC and PRC values which tells us there is an improvement over the model in Q1.

  2) We get better performance as we have used GridSearch CV in Q2, this technique is used to search through the best parameter values from the given set of the grid of parameters.

  3) It is basically a cross-validation method. the model and the parameters are required to be fed in. Best parameter values are extracted and then the predictions are made.

- Why do you think the new parameter values help deceptive comment classification?

  1) We have used unigrams in Q1, whereas in Q2 we use both unigrams and bigrams. The ngram range obtained is (1,2).

  2) It is quite evident that using bigrams adds more information when compared to using unigrams alone. This will eventually improve the classification.

  3) Also, higher min_df values along with GridSearch CV help eliminate typo mistakes. Min_df is used for removing terms that appear too infrequently.

- How does sample size affect each classifier's performance?

1) From the graph plotted in Question 3, we can clearly see that there is a great impact of sample size on the classifier's performance.
2) For Naive Bayes - we see a slight improvement in the F1 - macro score with an increase in the sample size. Now the improvement in Naive Bayes is quite less when compared to that of the SVM classifier.
3) We all know that the Naive Bayes classifier can quickly learn to use high-dimensional features with limited training data. This can be useful in situations where the dataset is small which is why we don't see a great improvement in the graph.
4) Hence, we observe that the Naive Bayes classifier's F1 macro score for sample size 100 is quite good as compared to SVM.

- If it is expensive to collect and label samples, can you decide an optimal sample size with model performance and the cost of samples both considered?

1) Considering it is expensive to collect and label samples, Yes - we can decide on an optimal sample size with model performance and the cost of samples both considered.
2) We will have to find a middle ground where the cost of sample collection is less and the performance of the model is not compromised.

3) As you can see in our case, we do not see any great improvement in the performance after 700 sample size. This can be considered as the middle ground where the performance of the model is not compromised.

BONUS

For this dataset, both Naive Bayes and SVM model can provide good classification accuracy. The question is, how the models conclude that a document is deceptive. What features have the discriminative power? Do your research to find the most descriminative features in a document. To illustrate, you can randomly select a few samples from the test subset to highlight these most discriminative featues.

1) One reason I can think of is common bigrams in the fake documents. There can be a similar number of frequency in the fake documents.
2) As Naive Bayes algorithm works on the Bayes theorem, the output of finding out whether the document is fake will depend on various conditional probabilities.
3) Some factors that I can think of are - number of words in the document, common unigrams, common words and similar tfidf scores. When all these factors are considered based on each other then the model concludes a document to be deceptive.
4)Both the model try and identify different patterns between the two labels. Based on these patterns the model concludes that a particular document is deceptive.