Lessons Learned: Sentiment Analysis

I used accuracy as the performance metric that guided my decision process in choosing the best model for the data. For the Galaxy data I ended up using the original data set with all the features. The only thing I changed to the data were the manually labelled sentiments. Instead of having levels 0 to 5 in sentiment, I recoded the sentiment to only range from 1 to 4. The reason for this is that it increased my predictive accuracy from 77% to 84%. Below is a diagram of how my selection process went using accuracy as the guiding metric.

Figure 1: Galaxy Model Selection Process based on Accuracy

The figure above shows the different accuracies obtained as I narrowed down the Galaxy model to the best suitable one for the data being modeled. The first step was choosing the most accurate model; classifier C5.0 ended up being the most accurate out the classifiers tested with the “out of the box” data and settings. Moving forward I tested the C5.0 classifier on the data using different feature selection methods. The highest accuracy yielding dataset ended up being the original data set with no modifications. The last step was checking if recoding would help my accuracy. It ended up increasing my accuracy by 7%. Using the same methodology, the highest accuracy obtained when predicting the iPhone sentiment was 84% as well. This accuracy was reached with predicting sentiment with the C5.0 classifier, recursive feature elimination (RFE) and recoding the sentiments from 0-5 to 1-4.

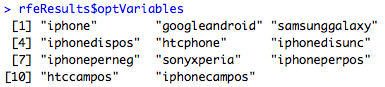


Figure 2: iPhone Optimum Variables using RFE

The figure above shows the features selected as most optimum when applying RFE to the iPhone dataset.

Choosing a performance metric to select the best model worked very well in this project, since it provided a clear path to follow when testing different model and feature combinations. Along with accuracy, creating a confusion matrix made it visually clear how the accuracy was arrived upon using the table format. Below is the confusion matrix obtained by comparing the iPhone sentiment predictions against the actual manually labeled sentiment of the data set used to test the iPhone model.

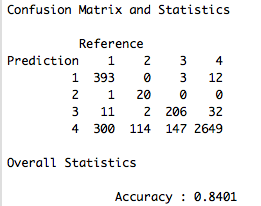


Figure 3: Best Model Confusion Matrix of iPhone Data (C5.0, RFE & Recode)

By creating the confusion matrix, I understood how accuracy was calculated; the total number of sentiment predictions matching the reference sentiment is divided by the total number of possible correct sentiment predictions. The amount of time it took to fit some of the training classifiers to the data was usually the biggest difficulty, or frustration, because of the many iterations and small adjustments I made to the models. In the future, I want to be able to implement a bigger variety of classifiers to my data, in order to see if I can achieve even better predictive performance.