**iPhone and Galaxy Sentiment Analysis Report.**

**Introduction**

In this project I was working to analyze sentiment on the web towards smart phone devices for Helio, Inc. Helio is a smart phone and tablet app developer and is working with a government health agency to create a suite of smart phone medical apps for use by aid workers in developing countries. The government agency requires that the app suite be bundled with one model of smart phone. After doing some research, Helio narrowed down that the app suite can be bundled with two devices: iPhone or Samsung Galaxy.

Alert! Analytics was asked to perform an extensive sentiment-based analysis for both models.  The task was to determine what model, iPhone or Samsung Galaxy, is preferred more by users, so that Helio can develop an app suite for the model with the most positive reviews.

**Methodology:**

Alert! Analytics first collected data from Common Crawl, which provides free and open crawl data in the form of datasets and archives to the public, after collecting it from the web. The small matrix data was collected and subset of documents were manually labeled with a sentiment rating. This data was put into a matrix that contains approximately 12,000 entries (small data matrix). Later, our manager wrote Python mapper, reducer, and output aggregator programs to efficiently collect and compile this type of data across the billions of documents on Common Crawl. We used this program to collect Large Matrix Dataset.

 Large Matrix Dataset was collected using Common Crawl. We took advantage of AWS “mapper” script that examines and counts data from portions of the Common Crawl data; the “reducer” script that accumulates the analysis from the individual mapper jobs; and the “aggregation” script that helps stitch together the raw output from the multiple job flows you need to initiate the analysis of all the necessary data. Using python script, we collected the dataset with 27,000 observations (Large Matrix) that we used to make the prediction, based on the models we built from the smaller matrix.

Our Small Matrix Dataset consisted of 12,000 observations. The dataset was clean and had 59 attributes. **A-F** attributes included information about the relevancy of the webpage toward each device. **H-V** attributes were about the sentiment towards a phone’s camera. **W-AK** had the information about the sentiment towards a phone’s display. **AL-AZ** contained the sentiment towards a phone’s performance(hardware). **BA-BF** columns covered information about the sentiment towards a phone’s performance (operating system). The dependent variable in our models was the iPhonesentiment for iPhone model and galaxysentiment for Galaxy model.

The scale of the sentiment was rated as follows: 0- unclear, 1- very negative, 2- somewhat negative, 3: somewhat neutral, 4-somewhat positive and 5-very positive.

Taking into consideration that our dataset is very large and RStudio only uses one core from the computer’s processor, we set up parallel processing, which means that we used additional processor cores to speed up the processing time.

We explored our small matrices and the sentiment for both devices and it was heavily distributed towards the 5 – very positive sentiment, with at about 7,500 observations; second place was 0-unclear sentiment with a bit less than 2000 observations. To analyze the dataset, classification models were built using 3 different methods: Out Of The Box, Near Zero Variance and Recursive Feature Elimination method.

Out Of The Box method consisted of keeping all the original attributes while building classification models.  The decision to keep all the attributes was made after checking for collinearity using correlation matrix.

Near Zero Variance method allowed us to get rid of the features with no variance since they hold little to no information for model building purposes.

Recursive Feature Elimination or RFE method is a form of automated feature selection. This method will try every combination of feature subsets and return a final list of recommended features.

**Results**

After setting up 3 different data frames based on the methods mentioned above, we ran 4 classification models: C5.0, Random Forest, SVM and KKNN.

Running 4 classification models revealed that the small matrix SVM models for both iPhone and Galaxy datasets, had low accuracies and were eliminated based on the complexity of the algorithm. KKNN model had the lowest results for the iPhone matrix, ranging from .31 to .33 and therefore was not used for further predictions. While KKNN models for Galaxy matrix were much higher standing at .74, this model was not for consideration since the results for other models, like C5.0 and RF were higher.

To make a predictions for the large matrix dataset of the sentiment analysis, C5.0 model, the “Out Of The Box” method was chosen for both phone models. The C5.0 model had the highest accuracy of  0.772709 for iPhone and 0.768433 for Galaxy. It is important to mention that in order to see how well the models performed, Post Resample accuracies were also taken into consideration. Making predictions on the large matrix, we found that sentiment results for iPhone and Galaxy models were very similar, with most of the sentiment falling on “Neutral” and the second highest sentiment being ‘Very Positive”. It is important to mention that “very negative,” and “somewhat positive” sentiments were omitted from the graphs because the count of both of them were close to zero. Please see the predictions results in the Graphs below:

Graph1.

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

**Recommendations**

Looking at the results above, both phone models have remarkably similar preference among users. Taking into consideration that our data was very skewed towards the “Unclear” and “Very Positive” sentiments, more data collections would be helpful to make better predictions. Based on the results above, the app suite can be developed for both models of the phones.