Summary of Findings

For Helio

Nastaran Mahmoodzadeh
Winter 2020

Cover Page

Project Name
Smart Phone Sentiment Analysis (iPhone and Samsung Galaxy)

The client
Helio

Alert! Analytics
Winter 2020

Overview

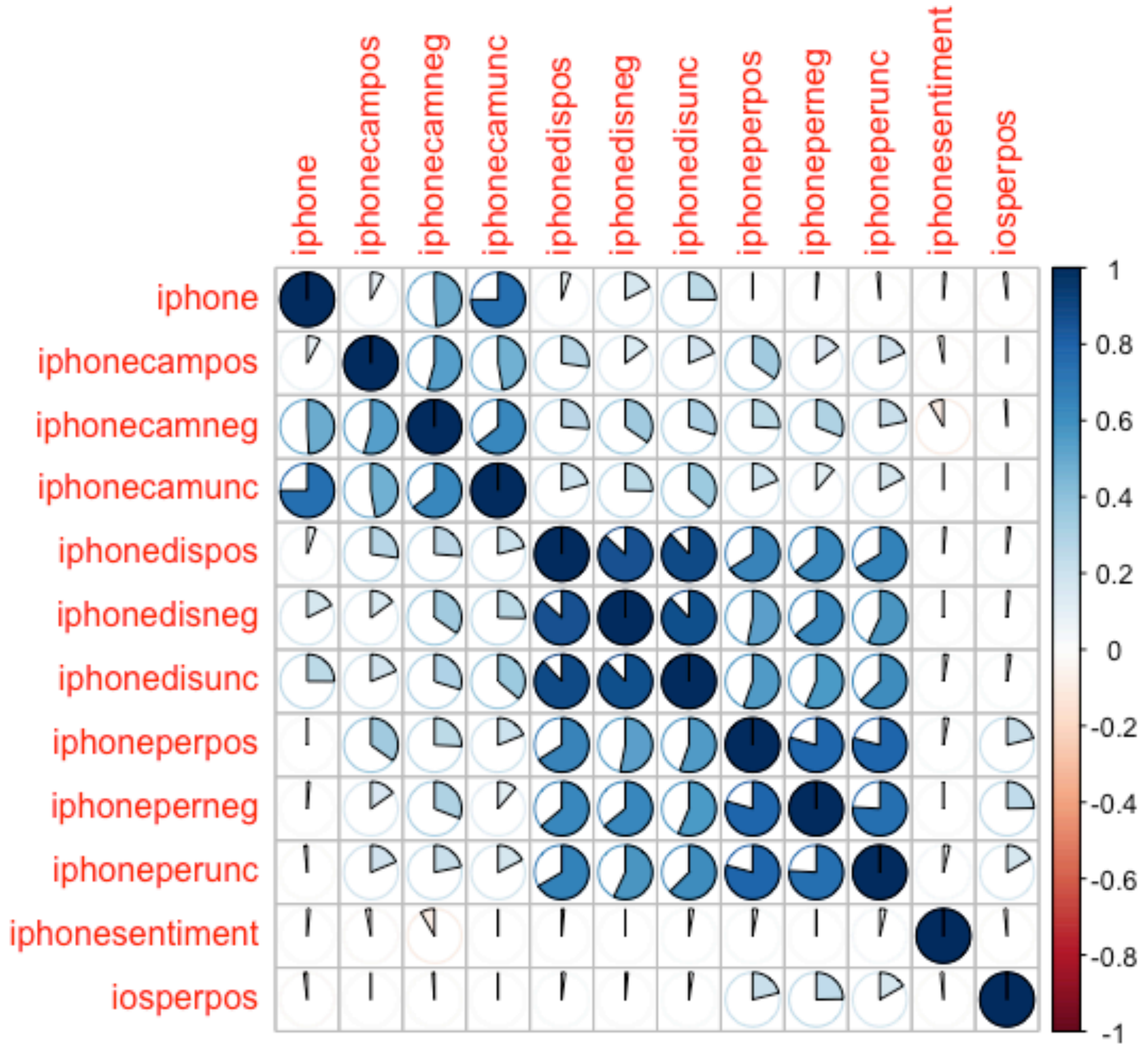
Background
We are working with a government health agency to create a suite of smartphone medical apps for aid workers in developing countries. This suite of apps will enable the aid workers to manage local health conditions by facilitating communication with medical professionals located elsewhere. The government agency requires that the app suite be bundled with one model of a smartphone. This will help them limit purchase costs and ensure uniformity when training helps workers use the device.

Objective
We were given a shortlist of devices capable of executing the app suite's functions, and we were asked to examine the prevalence of positive and negative attitudes toward these devices on the web. Our goal is to narrow this list down to one device by conducting a broad-based web sentiment analysis to gain insight into the attitudes about the devices. For the second part of the project, we will investigate predictive models using machine learning methods. We will apply these models to the Large Matrix file to complete the analysis of overall sentiment toward both iPhone and Samsung Galaxy.

iPhone

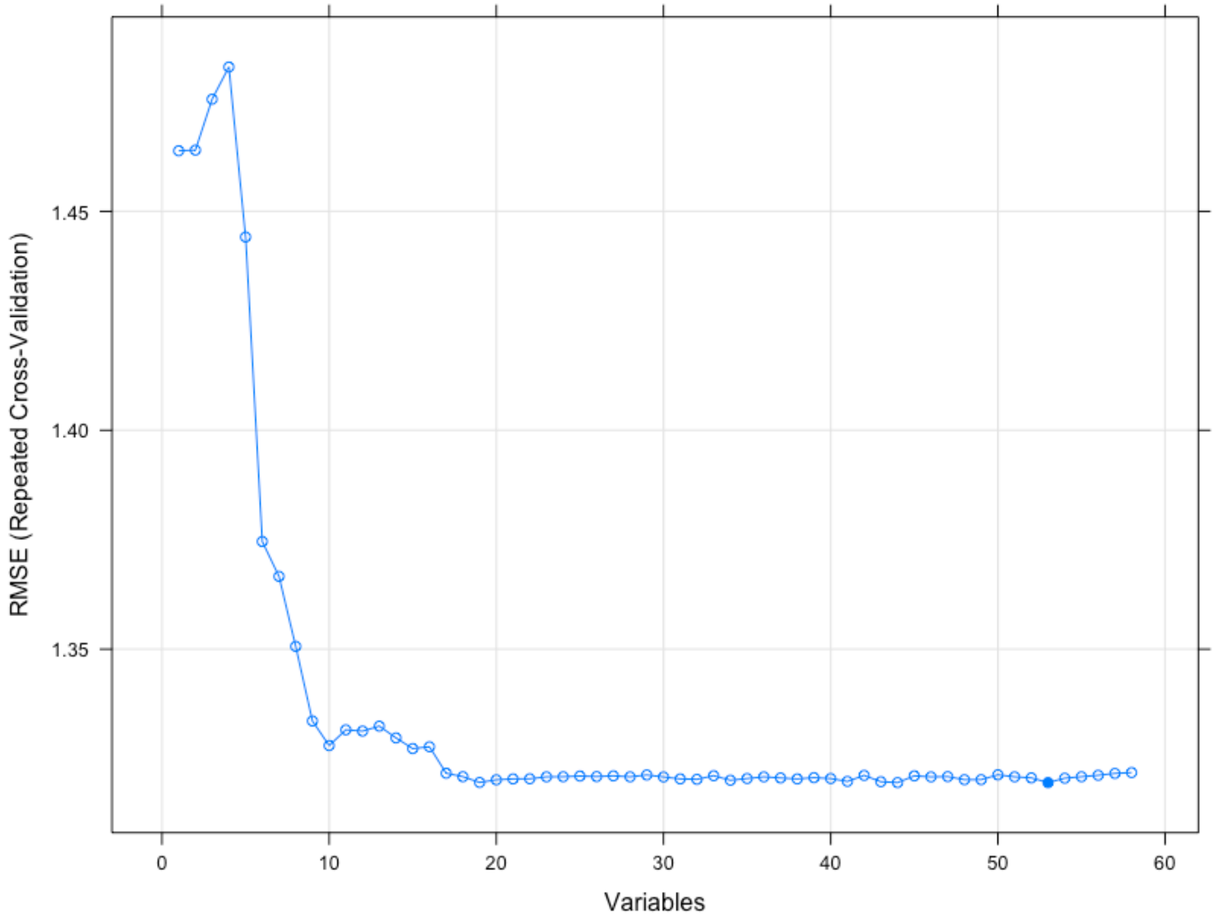
Correlation with No Collinearity

We looked at the correlation and we saw collinearity amongst “iOSPerNeg”, “iOSPerUnc”, and “iOS”, so we removed them, and plotted the correlation for iPhone features another time.



Recursive Feature Elimination

The resulting table and plot display each subset and its accuracy and kappa. An asterisk denotes the the number of features that is judged the most optimal from RFE.



Methodology

We used the small matrix dataset to run four models: C5.0, Rf, SVM, and KKNN. The Accuracy and Kappa were not ideal on SVM and KKNN, and were almost the same in C5.0 and RF. Hence, we ran the C5.0 model on NZV (Near Zero Variance) dataset and RFE (Recursive Feature Elimination) dataset, and the we still did not get ideal Accuracy and Kappa.

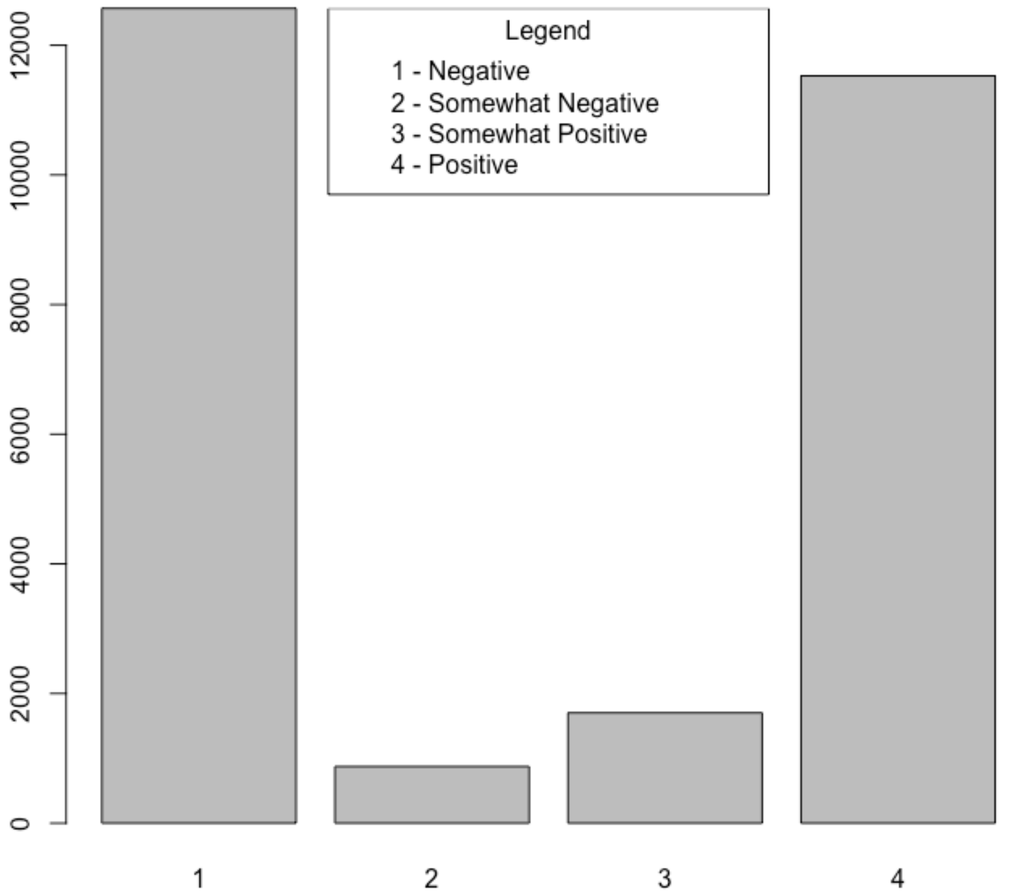
We have tried a variety of tuned algorithms on a variety of data sets that have been preprocessed and feature selected. The accuracy is not very desirable and it's not improving anymore. So, we decided to re-code the dependent variable, "iPhone sentiment", to 4 levels, instead of 5 levels, ran the C5.0 model on the new dataset. This way, we got the highest Accuracy of 0.85 and Kappa of 0.64.

Modeling: iPhone

Models					
Datasets (before re-coding)		C5.0	RF	SVM	KKNN
	Small Matrix (iphone_df)	Accuracy: 0.77 Kappa: 0.55	Accuracy: 0.77 Kappa: 0.55	Accuracy: 0.71 Kappa: 0.42	Accuracy: 0.33 Kappa: 0.16
	iphone_NZV	Accuracy: 0.77 Kappa: 0.55			
	iphone_RFE	Accuracy: 0.78 Kappa: 0.57			
Dataset (after re-coding)	iphone_RC	Accuracy: 0.85 Kappa: 0.64			

Findings on iPhone

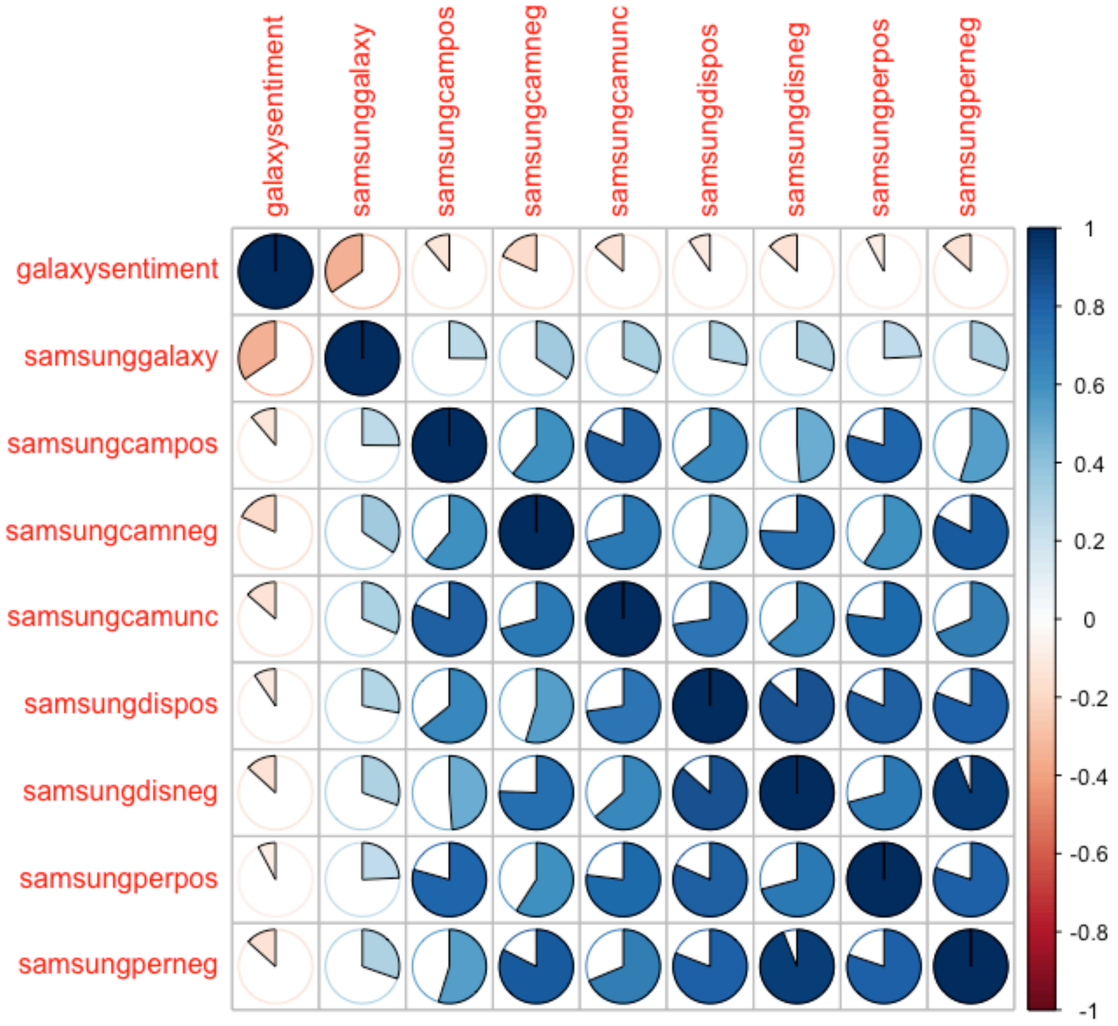
As we can see in this plot, people either love iPhone or hate it! Small amount of counts are towards neutral opinions, and the Negative sentiment counts are a bit more than Positive sentiment counts.



Samsung Galaxy

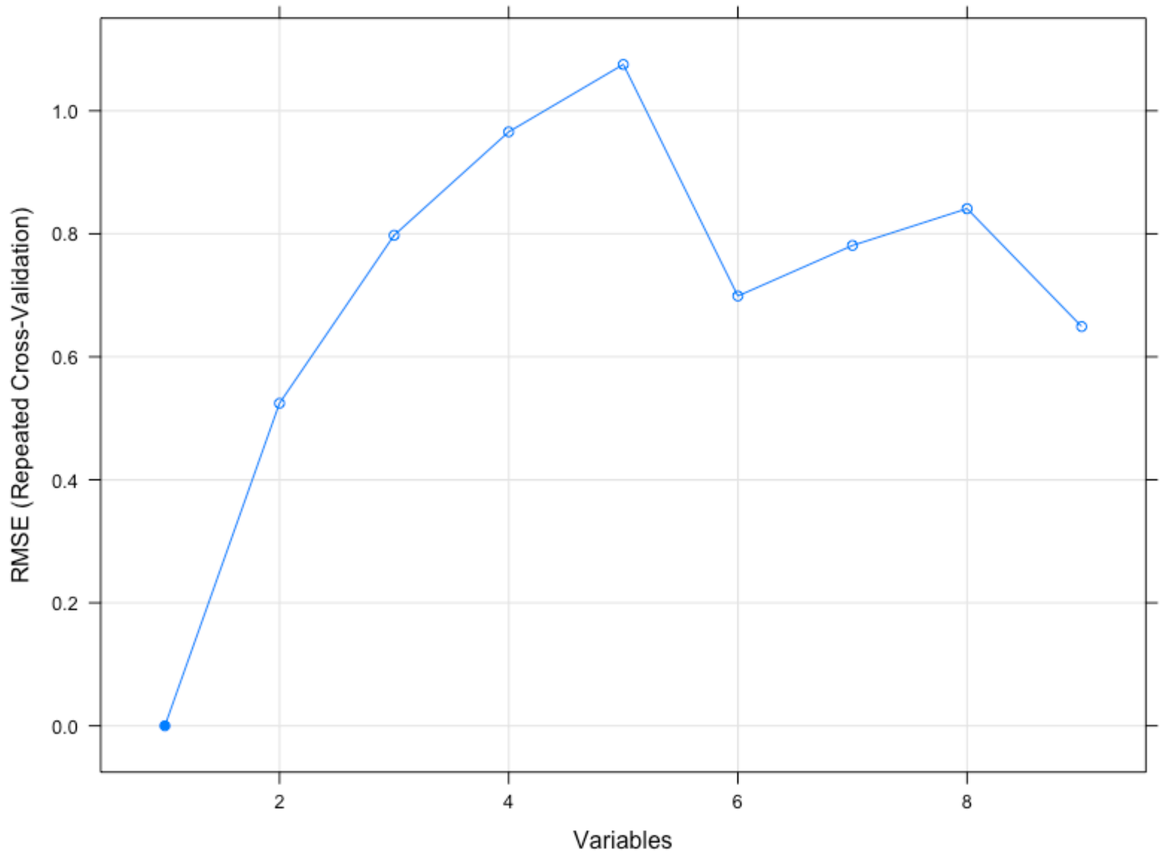
Correlation with No Collinearity

We looked at the correlation and we saw collinearity amongst “samsungdisunc” and “samsungperunc”, so we removed them, and plotted the correlation for Samsung Galaxy features another time.



Recursive Feature Elimination

The resulting table and plot display each subset and its accuracy and kappa. An asterisk denotes the the number of features that is judged the most optimal from RFE.

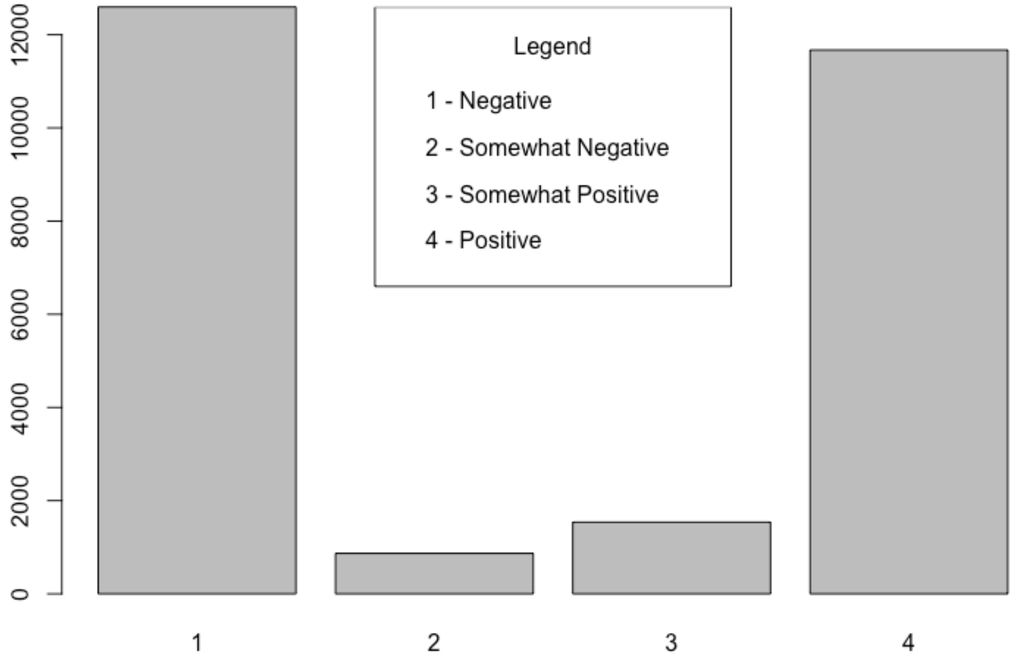


Methodology

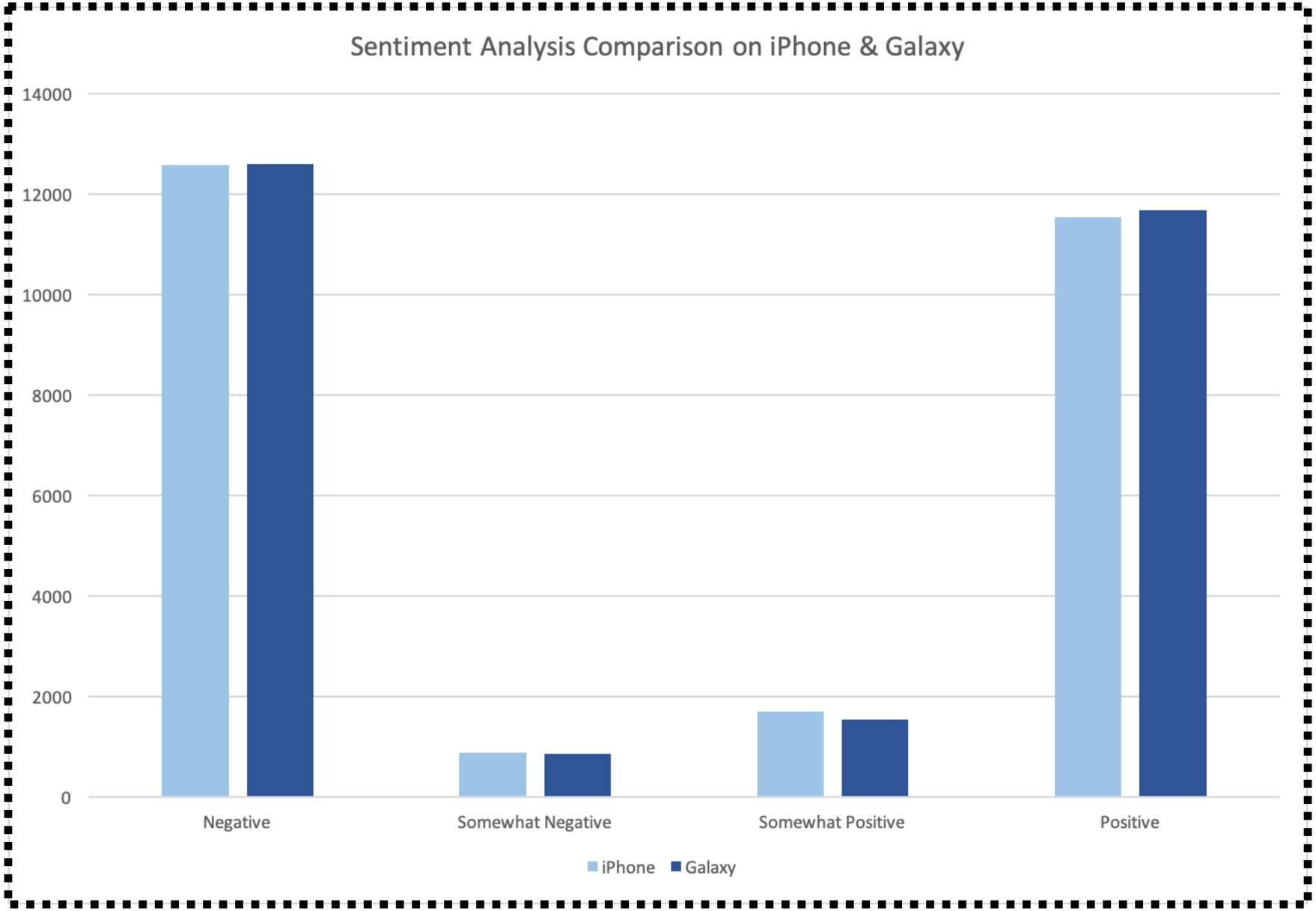
We repeated the same method on Galaxy labeled dataset. The same issue of low Accuracy and Kappa also happened with this dataset so we decided to try re-coding method on the “Galaxy sentiment” as well, to 4 levels, instead of 5 levels. The C5.0 model worked the best again here, with Accuracy of 0.84 and Kappa of 0.58.

Findings on Galaxy

Again, we can see that sentiment count prediction shows that people either love Galaxy or hate it! Small amount of counts are towards neutral opinions, while “Somewhat positive” is slightly more than “Somewhat Negative”.



Sentiment Count Comparison



Implications

- Ratings mostly distributed in Negative and Positive. The ratings represents Neutral sentiments (somewhat negative and somewhat positive) only made up about 10% of all ratings we analyzed.

- In this analysis, there are little differences of the sentiment rating distributions for iPhone and Galaxy. Both devices have slightly higher negative ratings than positive.

- If we take price and convenience into consideration, Galaxy can be the better choice to be used in developing countries since it is cheaper than iPhone and is more convenient with Android system.