# **Sentiment Analysis Project Report**

**PREPARED BY LIN ZHAO FROM ALERT! ANALYTICS**



### Overall Impression

We are Alert! Analytics team, who is specialized in applying machine learning techniques to help make business forecasts for our clients in order to identity new business opportunities.

Instead of manually reading through each webpage and assigned a sentiment rating, our client, Helio would like us to investigate and build predictive models that understand the patterns in the two small matrices (iphone\_smallmatrix.csv with 12973 observations and galaxy\_smallmatrix.csv with 12911 observations) and then apply the best winning model to the Large Matrix file(22461 observations) that we collected from Common Crawl to complete the analysis of overall sentiment toward both iPhone and Samsung Galaxy.

### Comparative findings for iPhone and Galaxy

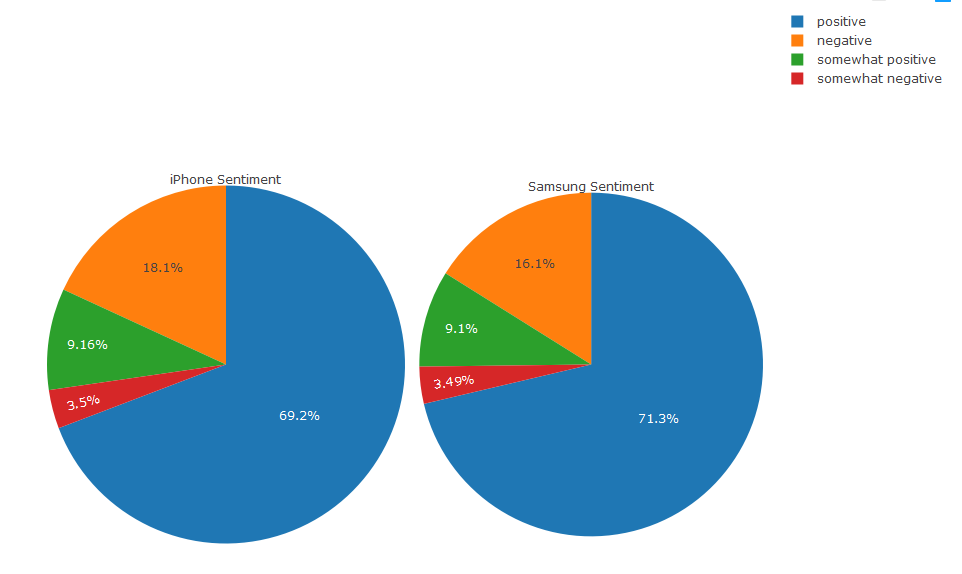
#### Sentiment Categories

Initially, we have six sentiment categories for both iPhone and Galaxy derived from the original datasets ranging from 0 very negative to 5 very positive. After a series of trials and errors on various algorithms, we found that the independent variable's factor levels had very poor Sensitivity and Balanced Accuracy. Therefore, we decided to combine some of the redundant levels and reduced them to 4 levels ranging from 1-4 (negative to positive).

#### Side by side Visualizations of the iPhone and galaxy sentiment results

There are no extremes in either matrices. Both pie charts are about the same. Although Samsung Galaxy has slightly higher positive sentiment review (71.3% Vs. 69.2%) and lower negative sentiment reviews (16.1% Vs. 18.1%) over the iPhone. We think this 2% difference is insignificant considering the error of margins the model itself and other factors bring in. Therefore, we should neglect the variance and believe that people are indifferent between iPhone and Galaxy.

Especially, when people stick to one manufacturer, i.e. iPhone, it is hard to switch to Samsung Galaxy and vice versa. Because when you get used to one OS and its ecosystem, including the apps and other digital assets purchased, it becomes “locked in” to that manufacture and it’s extremely difficult to persuade someone to switch to the other completely different ecosystem. The learning curve to adapt to a new OS system is high and people are inclined to take the easy route, upgrading to the next generation of their current phone manufacturer.



#### The raw sentiment counts for iPhone and Samsung Galaxy

As we can see from below table and bar charts, the positive sentiment reviews for Galaxy is a little higher than the iPhone (small matrix). The positive sentiment reviews for large iPhone matrix is slightly higher than the its negative counterpart. The neural reviews which include somewhat negative and somewhat positive reviews are about the same across all three datasets.

|  |  |  |  |
| --- | --- | --- | --- |
| Levels | iPhone Sentiment Count (small matrix) | Galaxy Sentiment Count | iPhone Sentiment Count (large matrix) |
| Negative | 2352 | 2078 | 9647 |
| Somewhat Negative | 454 | 450 | 614 |
| Somewhat Positive | 1188 | 1175 | 1407 |
| Positive | 8979 | 9208 | 10790 |
| Total Observations | 12973 | 12911 | 22458 |

### Accuracy and Kappa Values

Accuracy is self-explanatory, it refers to closeness of the measurements to a specific value. The Kappa value is a metric that compares an observed accuracy with an expected accuracy (random chance). In addition, Kappa takes into account the random chance which means it is less misleading than simply using accuracy as a metric. A Kappa value of 0.61-0.80 is considered substantial, and 0.81-1 as almost perfect. In reality, if the score is too high (close to 1), it implies risk of model overfitting, meaning the result is “too good to be true”.

#### Comparisons on key indicators before and after feature engineering recode function

In total, we have run 22 algorithms with different combinations of models and dataset selections for the entire training and testing experiment. The best algorithm was achieved by applying feature engineering recode function on random forest model.

Here we have tried to apply four different preprocessed datasets on random forest model. The average Accuracy value was between 0.75-0.77 and the average Kappa value was between 0.53-0.57 prior to implementing the feature engineering.

Later, we found that both Accuracy and Kappa values improved dramatically after applying the ultimate recode function, namely, to reduce the factor levels from 6 to 4, which eventually lead to our best result of accuracy: 0.84 and Kappa: 0.60 for iPhone small matrix and accuracy: 0.84 and Kappa: 0.59 for Samsung Galaxy small matrix.

#### Impact on the client’s goals

In response to Helio’s goal to investigate the sentiment analysis on both phone manufacturers in order to better understand customers’ demand, we believe that customers have no strong preference towards either phone, despite of the 2% marginal differences where Samsung Galaxy surpass the iPhone.

Looking into these datasets, we found that many features were mentioned like the camera, the display, the overall performance, however, there are a few features which are crucial to the customers sentiment but were not addressed in the given datasets.

1. Affordability: Samsung Galaxy offers cheaper price than iPhone with nearly all comparable features. It is a very important factor that determines the buying pattern especially for people that are price sensitive.
2. Battery life: Studies show Galaxy in general has noticeably longer battery life. It is also an essential criterion as people nowadays are addicted to smartphone, without a longstanding hour of battery, it really disappointed a lot of potential customers.

Our suggestion for future sentiment analysis is to consider adding these two features into the datasets, being said, both positive and negative variance should be bigger enough for us to make a better judgement on which brand customers have a stronger preference than the other.

### Methodology for Sentiment Analysis

Methodology consists of three steps outlined below. First, data was collected from Common Crawl. Second, the sentiment analysis was conducted using RStudio, Sublime text editor, AWS and Cyberduck to automate the data collection process and to clean the data. Third, the results of the analysis were visualized using R package namely Plotly.

#### Data Collection

In the first step, data was collected from Common Crawl by extracting reviews from sources like: websites, google, Amazon S3, Yelp, Twitter, etc. However, to manually read and scan all the reviews from various sources into a document is impracticable as it would be too time-consuming and inefficient. Hence this was automated through a python script.

#### Sentiment Analysis

After collecting the data, the second step was to analyze them. Using RStudio, we first run a number of classification models with different combinations of datasets and feature selection/engineering methods. Then the best model with the highest accuracy and kappa value was chosen.

The next step was to apply this model to the large matrix that we collected from Common Crawl. Once the data had been analyzed using the model, sentiment counts for each level were listed after running the summary function

#### Data Visualization and Results

The third step was to visualize the result using Plotly, a Rstudio package which provides interactive data visualization using graphing, analytics, and statistics tools for business intelligence. This is a very important step in the process, as data visualization can intuitively helping users quickly gain insight from a potential customer’s point of view.