Phone Sentiment Analysis

Predictive Analysis of iPhone and Galaxy Market Sentiment

Helio/Alert! Analytics

Kevin Burr

Helio is working with a government health care agency to develop a suite of apps to connect medical professionals in developing countries. The government agency requested Helio develop the suite of apps on one smart phone platform. IOT Analytics is contracted to analyze web page data and create a sentiment analysis model used to recommend the platform for Helio's development.

The sentiment categories are:

0: very negative

1: negative

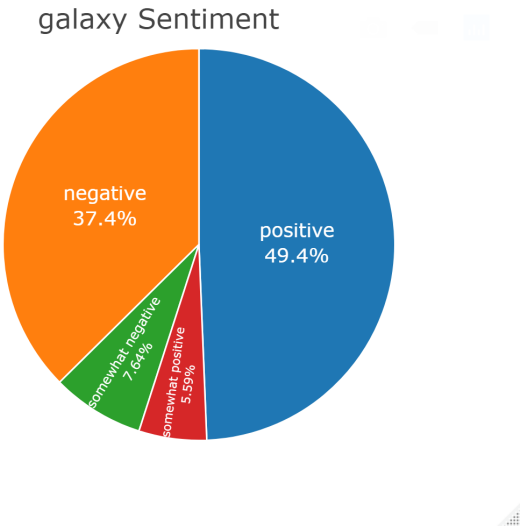
2: somewhat negative

3: somewhat positive

4: positive

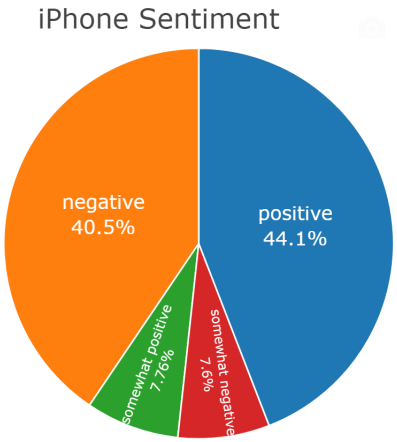
5: very positive

Out of the box distribution on the large data set set for Galaxy is interesting. The distribution of sentiment results suggest a generally favorable opinion of the Galaxy phones with far more (50% positive) versus not a favorable opinion (37% negative). Since the results are skewed so far to the left or right, we can conclude the sentiments 2-3, while important are low in magnitude, relatively speaking – so on the whole, people really like the Samsung products. For the final predictive results, I found through analysis I could achieve a far better accuracy if I combined the 0 & 1 ratings and 4 & 5 ratings. This one step improved predictive accuracy by almost 20%. The following table demonstrates the distribution of results by percentage.

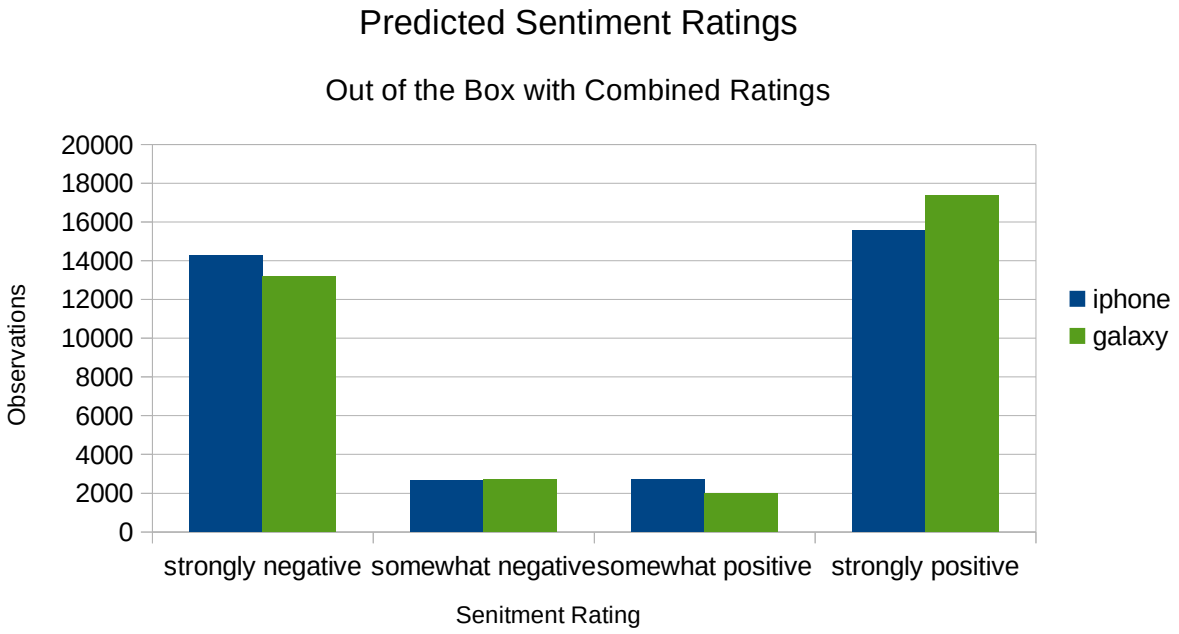


Out of the box distribution for iPhones is equally interesting. The distribution of sentiment results suggest a generally favorable opinion of the iPhone with far more (44% positive) versus

not a favorable opinion (40% negative). Since the results are skewed so far to the left or right, we can conclude the sentiments 2-3, while important are low in magnitude, relatively speaking – so on the whole, people really like the iPhone products based on the training data set.



The iPhone appears to more closely split down the middle in sentiment with almost 50% positive plus somewhat positive and 50% negative plus somewhat negative. So, compared to Galaxy, the iPhone is more closely polarized evenly as half love it and half hate it. The results for the iPhone are not wildly different than those for the Galaxy. Plus, since both experienced better predictive results when I combined negative (0 &1) and positive (4 &5) ratings, I could create a fairly similar comparison.



While the outcome of the analysis did not yield a really clear and dominant leader in the market, at least relative to sentiment analysis, based on this analysis, I recommend going with the Samsung phone for your application platform.

I ran 4 different algorithms and evaluated the results when using five different permutations of the source data from the large matrix. In the end, with the C5.0 model and a data set with combined attribute values (0 &1, 4&5), I was able to create predictions with 84% accuracy with confidence of 60% - so very good for Galaxy. For iPhone it was even better, 98% accuracy with 97% confidence.

The results of this analysis reasonably justify pursuing product suite implementation on Galaxy over iPhone.

For this analysis, I ran four different models using a variety of different sets of attributes from the small matrix of cleansed data from you team with known sentiment values assigned. I ran the four models against the small matrix for each phone out of the box as a baseline. Then I created a few different versions of that data sets by selecting a subset of the total attributes available and ran the models again. When I was satisfied with the results from the four models, I selected the best performing one, and ran it one more time by combining the 0 &1, and 4 &5, ratings. This gave me the absolute best predictive outcome – for each of the phones. With this understanding, I combined the input ratings on the large matrix, as I did in the test set. I then ran the final model against the large matrix to get the predicted sentiment values for Galaxy and iPhone.

Considerations for Future Analysis

There are possibly a few things you could consider adding into this analysis. To help drive to a clearer winner, you could research the market share of each of these phones in your target market. You could use the percentage of market share in the medical industry, for instance, as a weight to use to multiple each outcome above. If market share is very different, you could end up with one phone coming out ahead more strongly. If they are close in market share in the target market, then the results will be similar to the conclusions here.

The team might also consider expanding the vocabulary describing a positive or negative response as well as the vocabulary used to identify mentions of the products on web pages. This might provide a richer set of data, but at a cost for performance of the algorithms and possibly the complexity of the analysis.

Finally, the team could pull an even large set of observations from Common Crawl. While you may run into an economies of scale and/or diminishing returns for performance and results, increasing the observation set size might provide better insight – or better results from a predictive algorithm.