Smartphone Sentiment Analysis

08.15.2021

—

Prepared By:
Alert! Analytics
Kirsten Bjornson

Prepared For:
Helio
Michael Ortiz

Overview

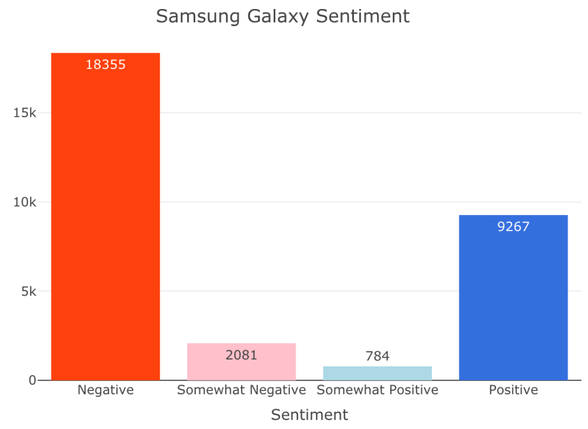
Helio is a mobile app development company that is currently working with a U.S. government health agency to develop several smartphone apps for use by medical workers providing aid in developing countries. They would like a single phone model to be chosen for the sake of streamlining the training process and ease of technical support services. They would also like to choose a model that will be popular among the workers.

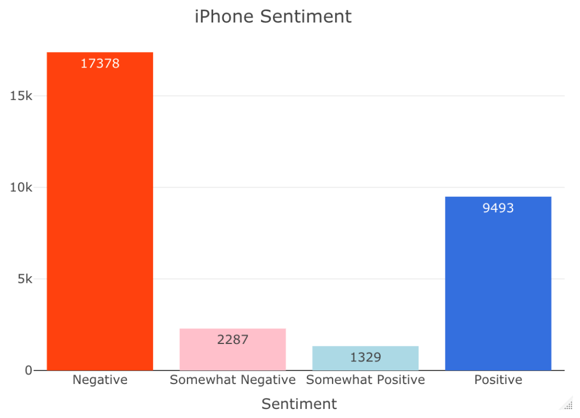
Alert! Analytic's role in this project is to conduct a broad-based web sentiment analysis of two popular smartphone models -- the iPhone and Samsung Galaxy. Our goal is to understand the overall attitude towards each of these devices, so that Helio can choose their smartphone model with confidence.

Findings

Four different sentiment categories were used when describing each phone. These categories are Negative, Somewhat Negative, Somewhat Positive, and Positive.

A comparison of the sentiment towards each phone is shown below:





Confidence

A C5.0 model was ultimately chosen for each phone. The model was chosen over others for its higher level of Accuracy and Kappa while at the same time being quick and efficient to run.

The table below illustrates the accuracy and kappa values for each model on their respective training set:

	Accuracy	Kappa
iPhone	0.8447301	0.6123915
Samsung Galaxy	0.8425723	0.5900179

Implications

The distribution of sentiment towards each phone is quite similar, with negative sentiment being the overwhelming majority for each phone. It should be taken into consideration that although negative sentiment is high for each, that people who have a bad experience are more likely to leave a negative review than people who have a good experience are to leave a positive review. That being said, based on the above results, Helio could choose either iPhone or Samsung Galaxy knowing that either will be successful with their employees.

However, it may be useful to do further analysis with the following considerations. First, the distribution of the training data was highly skewed, with there being 3.5 times more instances of 5's (positive) than 0's (negative), and even fewer of the in-between values. A more evenly distributed training dataset would make for more accurate predictions of the in-between values, whereas now the training set is very one-sided.

Additionally, it may be useful to compare these results to other phone types, or even to survey employees to find out their sentiment towards each model of phone.

Methods

Data Collection

Data was collected from the Common Crawl, an open repository of archived data from the web, consisting of billions of webpages to date. Using Amazon Web Services (AWS) and Elastic Map Reduce (EMR), we were able to build a large data matrix consisting of over 20,000 instances of pages that were relevant to our analysis. In order to assess whether a document was relevant, words indicating that the page was about the device and presented a meaningful analysis of the device were searched for. A document was only used if it contained at least one mention of the device and contained at least one of the following terms: review, critique, looks at, in depth, analysis, evaluate, evaluation, assess.

Once relevant documents were found, information was collected about the sentiment towards features of each phone, including the display, camera, performance, and operating system. The script searched for and counted positive, negative, and neutral words that were within 5 words of a feature word (such as camera or display).

Additionally, two small data matrices were created using this same method for the purpose of model training, in which the overall sentiment towards each phone was determined manually by our team. Using a trained model, the goal was to be able to predict overall sentiment towards each phone in the large data matrix that was collected.

Feature Selection

The data matrices consisted of 58 features, columns A-BF. Columns A-E contain information about the relevancy of the webpage towards each device (including iphone, samsung galaxy, and some other phone models). Columns F-G contain information on the sentiment towards the operating system of each phone, columns H-V about the sentiment towards the camera, columns W-AK about the sentiment towards the display, and columns AL-BF about the sentiment towards the performance of various features of each phone. The dependent variable is the overall sentiment towards each phone.

Three different feature-selection methods were used for each dataset:

Correlation: A correlation matrix was used to determine whether features were correlated with the dependent variable or not. While some features were more correlated than others, ultimately the decision was made not to eliminate features based on this method.

Near-zero Variance: Another feature selection method eliminated models with zero variance, while keeping features with near-zero variance or higher.

Recursive Feature Elimination (RFE): Lastly, RFE was used with random forest in order to create a dataset that only includes the most relevant features. RFE starts with all of the features in the dataset and removes features until an optimal feature-set is created.

Model Building and Evaluation

Initially, four different models were tested: **C5.0**, **Random Forest**, **SVM**, and **KKNN**. They were tested on the out of box dataset, NZV dataset, and RFE dataset for both the iPhone and Samsung Galaxy.

At first, the dependent variable consisted of 6 sentiment classes. This proved difficult for each of the models, as the 0 (negative) and 5 (positive) classes were the most abundant and therefore more easily learned, while the other 4 classes were very difficult to interpret due to their comparatively small numbers. For this reason, the classes were condensed into 4 categories (Negative, Somewhat Negative, Somewhat Positive, and Positive), which allowed for more accurate models.

Models were evaluated using PostResample and Confusion Matrices. With the Confusion Matrix, sensitivity and specificity were very important to determining how good the model was at predicting each class (again, this further exemplified that models were better at predicting the Negative and Positive values, while not being good at detecting Somewhat Negative or Somewhat Positive values).

Ultimately, C5.0 was chosen for each dataset, using the condensed dependent variable consisting of 4 classes. C5.0 was chosen due to its high confidence levels while still being able to run very quickly. Random Forest did perform slightly higher accuracy-wise, but took much longer to run, rendering it very inefficient. The other model-types, SVM and KKNN could not compare when it came to accuracy.

Conclusion

Based on the current data, both the iPhone and Samsung Galaxy have similar results when it comes to sentiment. Further analysis using either more evenly distributed training data, or additional smartphone models may shed more light on whether there is a strong preference for one phone over the other.