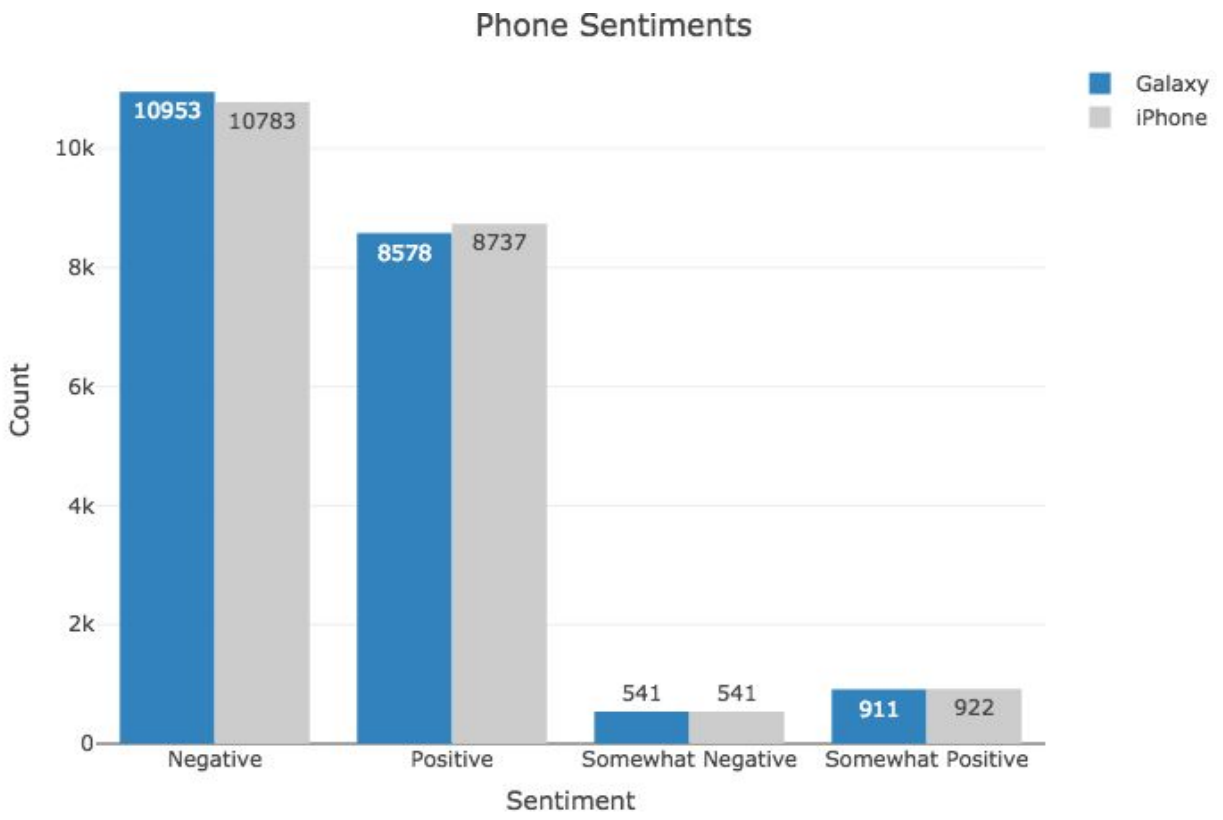
Project: Predicting Sentiment of Iphones and Samsung Galaxy
Client: Helio
Company: Alert Analytics
Analyst: Tasneem Dawoodjee

Overview

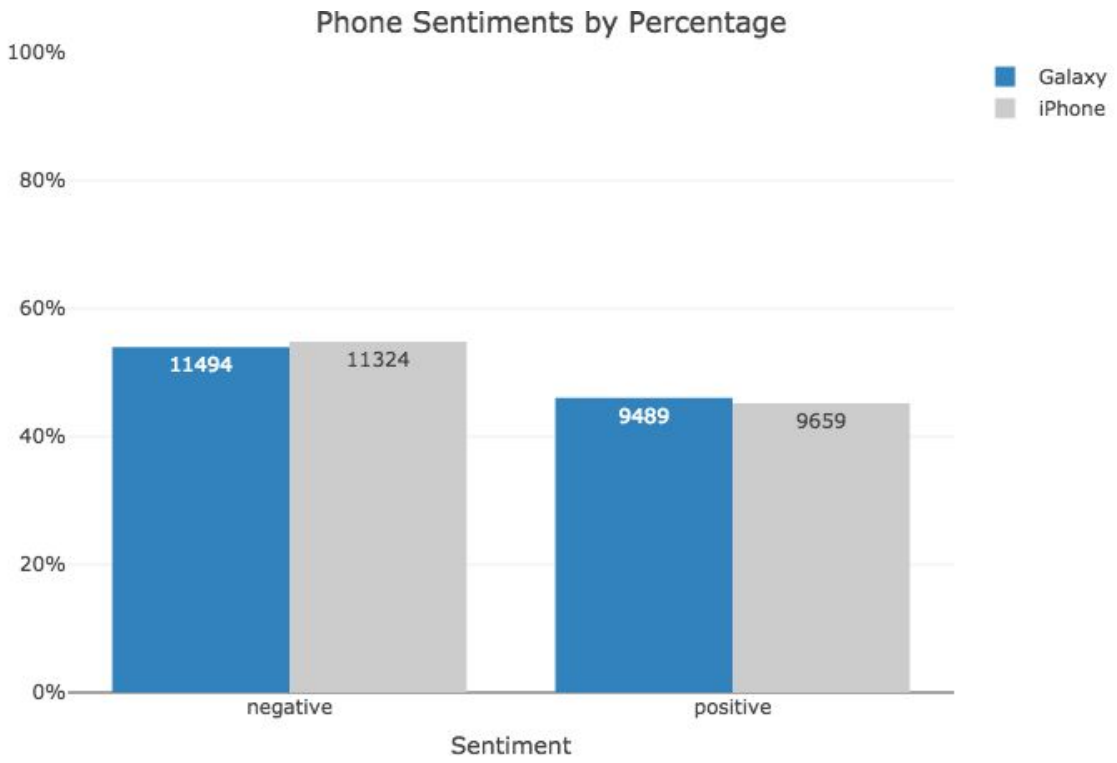
Alert Analytics is a data analytics consulting firm. Our client, Helio, is a smart phone and app developer. Helio is building a smart phone app for medical workers of a government agency. Helio provided us with a short list of phones that are suitable for their project. However, after discussions with Apple and Samsung, Helio requested that Alert Analytics examine sentiment for the iPhone and Samsung Galaxy. Our team went through over 12,900 webpages and manually identified words relevant to sentiment for each device and their attributes (like iphone camera). Based on the relevant words and their counts, our team applied sentiment ratings for each of these webpages. Using this data, we trained models to identify the sentiment for iPhones and Samsung Galaxys' on over 20,000 web pages within the last month.

Findings

We separated sentiment into four categories: negative, somewhat negative, somewhat positive, and positive. From over 20,000 webpages, reviews were primarily either negative or positive but not usually in between. From the chart below, we can see that the iPhone had more positive sentiments and less negative sentiments than the Galaxy, although not by much.



In order to address the small margin, we condensed our categories into just negative and positive, and used percentages. Percentage is defined as the total count of a particular sentiment for a particular phone divided by the total web articles with mentions of that phone. This method allows us to better assess the results since we analyzed 62 more articles mentioning iPhones than Galaxy.



From the Phone Sentiments by Percentage chart above, we can see that the Galaxy actually has less negative reviews by percentage and more positive reviews by percentage. Still, the differences are marginal.

Sentiment Analysis on 20,000 web pages for the month of April 2020				
Sentiment	Galaxy Count	Iphone Count	Galaxy %	Iphone %
1: negative	10783	10953	51.39%	52.20%
2: somewhat negative	541	541	2.58%	2.58%
3: somewhat positive	922	911	4.39%	4.34%
4: positive	8737	8578	41.64%	40.88%

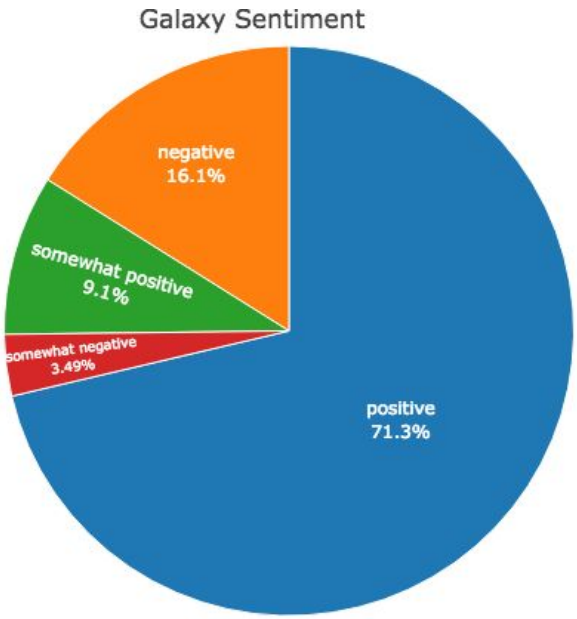
Confidence Section

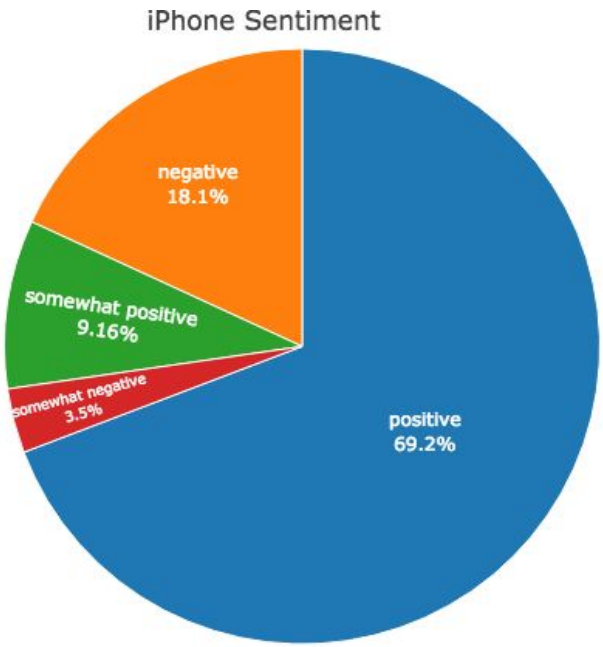
Error Metrics

Accuracy Rates of Models		
Device	Accuracy	Kappa
Galaxy	84.2%	59.0%
iPhone	85.2%	63.2%

Our Opinion and Caveats

We cannot confidently recommend one device over the other because the difference is marginal. Both the iPhone and the Galaxy have similar sentiments. Our best models were able to classify sentiment correctly by approximately 85%. If the difference in sentiments were larger, we would be more confident in recommending one device over the other. However, a variance of 15% could change the overall sentiment of a device significantly.





It is interesting that there are more negative reviews than positive reviews for both phones. Marketing research explains that If people are happy with their purchase, they will move on. It's more likely that people who are upset with their purchases will take the time to write a negative review[1]. However, the frequency distribution of sentiments on the web pages that are team manually rated did not have similar distributions as the web data from the month of February 2020. Rather, we reviewed considerably more positive webpages. To mitigate this imbalance, we recommend increasing the scale of this project. Per Helio's request, Alert Analytics shortened our analysis due to time constraints. In the future, we recommend allocating

more time to this project in order for our team to analyze web articles with more negative ratings. We expect our accuracy rate to increase after addressing class imbalance, and therefore, more certain insights on sentiments.

Implications

Overall, we learned that the Galaxy phone had more positive and less negative sentiments than the iPhone; however, not by much. We conclude that when considering which phone to continue development with, to consider other factors other than sentiment. For more certainty between devices and sentiment, we recommend pursuing this project further but with more hours allocated to manual ratings to improve pattern recognition for all sentiment categories.

Methodology

We divide this project into three parts: obtaining data, capturing sentiment, and training models.

We used RStudio, OpenOffice, Amazon Web Services' (Elastic Compute Cloud (EC2), Elastic MapReduce (EMR), and Simple Storage Service (S3)) as our platform to conduct this analysis. We obtained our data from Common Crawl, a not-for-profit open repository of web crawl data that is stored on Amazon's Public Data Sets.We wrote Python scripts to identify, obtain, and compile web text files relevant to Helio's list of devices. Our team counted words associated with positive and negative sentiment, and applied ratings to based on the word counts. We used this data and machine learning methods to look for patterns in the documents which enabled us to label each of these documents with a value that represents the level of positive or negative sentiment toward each of these devices. We explored various machine learning methods, preprocessing, and feature engineering to train the most accurate model. In total, we trained 38 different models, and selected the best performing model for each device. We applied our models to the web data from March 2020 and analyzed and compared the frequency and distribution of the sentiment for each of these devices.

Sources

1. https://www.reputationbuilder.us/angry-customers-likely-post-bad-review-happy-customers-good-one/