PHONE SENTIMENT ANALYSIS

3/10/2017

For: Helio
By: Andy W. Campos

CONTENTS

PROJECT OVERVIEW

GOAL:
The goal of this project is to provide Helio with a report that contains an analysis of sentiment toward the short list of target devices, iPhone and Samsung Galaxy, as well as a description of the methods and processes used to arrive at the conclusions.

SUMMARY:
Helio is working with a government health agency to create a suite of smart phone medical apps for use by 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 (one of the apps, for example, enables specialists in communicable diseases to diagnose conditions by examining images and other patient data uploaded by local aid workers). The government agency requires that the app suite be bundled with one model of smart phone. Helio is in the process of evaluating potential handset models to determine which one to bundle their software with. After completing an initial investigation, Helio has created a short list of devices that are all capable of executing the app suite's functions. To help Helio narrow their list down to one device, Helio has asked Alert Analytics to examine the prevalence of positive and negative attitudes toward these devices on the web.

MY APPROACH TO THE PROJECT:
Although there are a number of ways to capture sentiment from text documents, the general approach to this project is to count words associated with sentiment toward these devices within relevant documents on the web. I then leverage this data and machine learning methods to look for patterns in the documents that enable me to label each of these documents with a value that represents the level of positive or negative sentiment toward each of these devices. I then analyze and compare the frequency and distribution of the sentiment for each of these devices.

In order to really gauge the sentiment toward these devices, I must do this on a very large scale. To that end, I use the cloud computing platform provided by Amazon Web Services (AWS) to conduct the analysis. The data sets I analyze come from Common Crawl. Common Crawl is an open repository of web crawl data (over 5 billion pages so far) that is stored on Amazon's Public Data Sets.

STEPS TAKEN TO PRODUCE THE RESULTS

1. Obtained data from the January archive of the Common Crawl
2. Counted the sentiment information using Amazon Web Services due to the large volume of data contained within the archive
3. Built predictive models to identify sentiment patterns in the Common Crawl data using R
4. Created charts and tables to express the sentiment counts and performance metrics

RESULTS OF SENTIMENT ANALYSIS

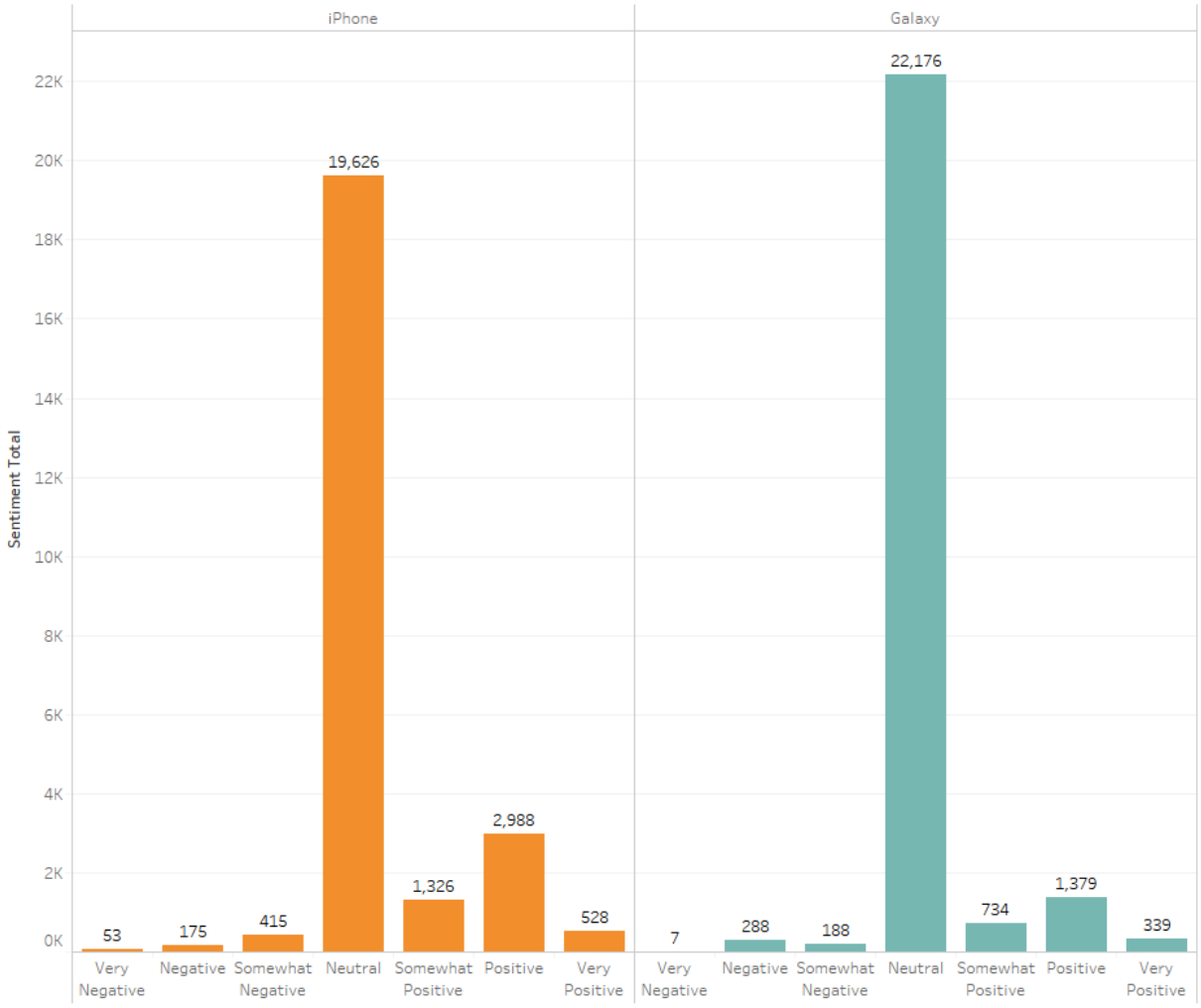

Figure A. Distribution of sentiment for mobile phones, iPhone and Samsung Galaxy containing 25,111 records per phone.

Using the data collected from Amazon Web Services, Figure A illustrates a comparison of the frequency and distribution of sentiment towards the iPhone and Samsung Galaxy mobile phones. The charts show the polarity of the sentiment is neutral. The frequency of the neutral rating has over 19,000 instances from the Common Crawl data I analyzed. This accounts for the majority of the data points, approximately 78% and 88% of the data collected for the iPhone and Samsung Galaxy devices respectively.

Reviewing the details of the remaining 22% and 12% of the iPhone and Samsung Galaxy sentiments respectively, the individual next largest polarity is positive. Next to neutral, there is a positive attitude towards each of the phones. There is very minimal negative sentiment for either phone.

Comparing the results of both the iPhone and Samsung Galaxy sentiment while excluding the neutral attitudes (22% iPhone data and 12% Samsung Galaxy data), there is a preference towards iPhone. The collection of positive sentiments (Somewhat positive, positive, very positive) for the iPhone have greater frequency than the positive sentiments towards the Samsung Galaxy. However, the collection of negative sentiments (Somewhat negative, negative, very negative) within the iPhone is also greater than the negative sentiments for the Samsung Galaxy. When taking into account both of these groups for the iPhone and Samsung Galaxy, there is a minor trend with greater frequency of positive sentiment for the iPhone overall.

IMPLICATIONS OF THE REPORT TO CHOOSING A PHONE

The frequency for positive and negative sentiment account for 22% of the iPhone and 12% of the Samsung Galaxy data collected. It is likely that the overall sentiment will remain neutral as long as the current trend remains in place. Choosing either the iPhone or Samsung Galaxy device would be received with similar neutral attitudes toward the devices. The underlying minor trend towards iPhone preference may fluctuate from preferred device towards neutral when further analyzing new data points. Keeping this minor trend in mind, choosing the iPhone device may be received well by some with neutral sentiment overall and minimal negative attitudes.

PERFORMANCE METRICS AND CONFIDENCE IN RESULTS

iPhone Performance Metrics testing 4 algorithms - using a sample of the data		
Model Name	**Accuracy**	**Kappa**
C50MFit	0.9002865	0.7117759
KnnMFit	0.8867797	0.6596404
RFMFit	**0.9038559**	**0.7253988**
SvmMFit	0.8943	0.6940787
iPhone PostResample on Random Forest, best performing model with the sample data		
Model Name	**Accuracy**	**Kappa**
RFMFit	**0.9039265**	**0.7212807**
iPhone Performance Metric on best performing model, Random Forest, with 25,111 records		
Model Name	**Accuracy**	**Kappa**
RFMFit	0.9113866	0.7483105
iPhone PostResample on Random Forest, best performing model with 25,111 records		
Model Name	**Accuracy**	**Kappa**
RFMFit	**0.9110108**	**0.7463446**

Galaxy Performance Metrics testing 4 algorithms - using a sample of the data		
Model Name	**Accuracy**	**Kappa**
C50MFit	0.9612603	0.8086122
KnnMFit	0.9495606	0.7305696
RFMFit	**0.9629117**	**0.8186104**
SvmMFit	0.9618659	0.8086658
Galaxy PostResample on Random Forest, best performing model with the sample data		
Model Name	**Accuracy**	**Kappa**
RFMFit	**0.9632599**	**0.8168763**
Galaxy Performance Metric on best performing model, Random Forest, with 25,111 records		
Model Name	**Accuracy**	**Kappa**
RFMFit	0.9689043	0.8511672
Galaxy PostResample on Random Forest, best performing model with 25,111 records		
Model Name	**Accuracy**	**Kappa**
RFMFit	**0.9694555**	**0.8541208**

Legend

C50MFit	C5.0	**RFMFit**	Random Forest
KnnMFit	K Nearest Neighbor	SvmMFit	Support Vector Machines

Figure B. Performance Metrics for Models used to develop analysis

The performance metrics in Figure B summarize the results of testing four models: C5.0, K Nearest Neighbor, Random Forest, and Support Vector Machines. Each of these models used 4,000 rows of sample data to gauge performance. The best performing model for both iPhone and Samsung Galaxy sentiment, Random Forest, was then tested against known sentiment values using a post resample of the data. The Random Forest model was further tested using the full set of 25,111 records collected per phone from the Common Crawl. The final set of metrics collected were the accuracy and kappa statistics in the post resample sections of the best performing models with 25,111 records.

In gathering data for the analysis, six types of attributes were collected: relevancy of webpage toward devices, operating system, phone's camera, phone's display, hardware performance, and operating system performance. These types of attributes worked fairly well in identifying pages with relevant sentiment toward each phone. By using these attributes to capture sentiment, I am able to factor in attitudes related to the features Helio may use for app development. For example, by adding camera preferences to the sentiment counts, I can get a sense of how the feature contributes to the overall sentiment and by extension, gain an understanding of whether the camera is likely to be used by aid workers to upload images that can be examined by specialists in other locations, as mentioned in the summary section of the project overview.

However, the accuracy of the process used to obtain sentiment could be improved. The process to capture sentiment uses particular mentions of keywords to count sentiment toward each device. If the list of relevant keywords were to expand to include a wider range of synonyms more sentiment data could be collected. Additionally, if more specific relevant words are targeted towards sentiment, the accuracy in collecting sentiment would improve. A balanced approach of including a wider range of synonyms while controlling for keyword accuracy could improve the quality of sentiment measured in the next round of analysis.