**DATA EXPLORATION & OBSERVATIONS**

1. The distributions of the ratings are clustered at the left end of the scale because the class value 0 or close to 0 is the most frequent value for most of the attributes. This means that the term doesn’t appear in most cases.
2. Since our metrics are based on adjective/attribute counting and do not take into account the whole context, there is always a likelihood that relationships between terms and qualifying adjectives may not report accurate meaning.
3. The Data Visualization indicates that regardless of the sentiment (positive, negative or uncertain) Iphone or IPhone related terms are the terms that appear with more frequency than any other mobile phone name or its related terms. This tendency can be observed in most instances.
4. Our observations support the idea that Iphone is the most popular brand.
5. Galaxy appears to be the second brand in popularity and it is more strongly related to other phones that use android operating system.
6. Possible Anomalies in the Correlations: Iphone and Iphone related terms appear at least once in most of the documents (even when the document’s main content refers to a different brand). This fact may indicate that Iphone is the main reference for any mobile phones comparisons. The more a term appears, the more it seems related to other terms, however some correlations between Iphone attributes and others brand attributes may be mostly incidental, and further correlation analysis is required to determine how strong or weak are these correlations.

**FINDINGS & SUMMARY**

In order to determine the sentiment towards Iphone and Samsung Galaxy we created a Small Data Matrix with 59 attributes. Each instance in the Small Data Matrix was a sample from a specific URL. The total number of URL collected was of 12,973. Each instance sentiment attribute was graded manually in relation to the URL document context from 0 to 5, being 5 the highest positive score. We tested several classifiers (J48, Random Forest, IBK, SMO) on the Small Data Matrix to find the best predictor for sentiment and use it as the Machine Learning Prediction Model. The selection criteria were based on number of correctly predicted instances and error rates. The best performance classifier model was to be applied to our Big Data collection. We collected information on mobile phones reviews and evaluations from 20,000 sources in Internet and created a Big Data Matrix for both brands (Iphone and Galaxy.) We ran our prediction model on the Iphone Big Data Matrix and on the Galaxy Big Data Matrix to predicted sentiment values for each. Our results were consistent with those observed in our Machine Learning tests.

The reported error metrics from the WEKA Small matrix tasks for 12,973 instances:

|  |  |  |
| --- | --- | --- |
| PREDICTION MODELS BEST PERFORMANCE METRICS | IPHONE MODEL | GALAXY MODEL |
| Correctly classified instances | 987 / 76.08% | 10135 / 78.123% |
| Incorrectly classified instances | 3102 / 23.91% | 2838 / 21.87% |
| Kappa Stat | 0.5351 | 0.5627 |
| Mean Absolute Error | 0.1269 | 0.1177 |
| Relative Mean Squared Error | 0.2571 | 0.2473 |
| Relative Absolute Error | 61.75% | 58.98% |
| Relative Root Squared Error | 80.194% | 78.299% |
| Time used to create the model | 0.91 s | 0.29 s |

1. Based on the results, most reviewers have a positive sentiment towards Iphone followed by Galaxy. However, for our training set we used human intervention to accurately grade sentiment, and it was a full context grading that we are unable to reproduce 100% applying machine learning to our data mining. Our level of confidence on getting an accurate sentiment in the data mining for Iphone sentiment was 76% and for Galaxy sentiment was 78%.

Our Data Mining Scripts selected webpages identified as “review”, “critique”, “look”, “insight/ depth”, “analysis”, “evaluation” or “assess(ment)”. Therefore, we have a high confidence on the webpage selection. The positive, negative and uncertain qualifying adjectives/attributes sets are typical of specialized and informal mobile phone literature, so we are also confident that they accurately represent relevant sentiment expressions.

Our study did not take into account or measure other features such as price, distribution, proprietary operating system apps, market tendencies, technical support, and such. Decision-making should not be solely based on sentiment analysis. The aforementioned aspects should be added to effectively qualify a product, and we recommend doing it in the next round of analysis.

Our analysis preselected models mostly based on technical qualifications. Our Sentiment analysis didn’t consider the particularities of the geographical regions where these devices are to be used. We didn’t assess features such as price, distribution, technical support, ISP services availability and cost on site nor did we identify the Internet information’s geographical source, therefore we cannot warranty the accuracy of the sentiment in specific developing countries. The analysis was solely run on English written webpages, letting outside any other language-related features or considerations.

In summary, Helio’s Mobile Phone Sentiment Analysis consisted on the creation of a statistical sentiment prediction model to be applied on a Big Data set of documents that express sentiment towards a pre-selected set of mobile phones. The sentiment predictions provided approximately a 76%-78% confidence on public preferences toward specific brands, and supported the decision-making process to create a suite of smart phone medical apps for use by aid workers in developing countries.