SENTIMENT ANALYSIS for HELIO

**Performed by Alert! Analytics, 2019**

**OVERVIEW**

This project was commissioned by Helio, the company that works for the government and is aimed to develop a suite of smartphone medical apps for use by aid workers in developing countries. The government agency requires that the app suite be bundled with one model of smartphone. To help Helio select the best device, they have asked us to examine the prevalence of positive and negative attitudes toward different devices on the web. The goal of this project is to provide our client with a report that contains an analysis of sentiment toward the target devices, as well as a description of the methods and processes we used to arrive at our conclusions.

**FINDINGS**

To conduct comparison analysis, we used the Common Crawl open repository of web crawl data that is stored on Amazon’s Public Data Sets. Moreover, we also used Amazon Web Services to go through each web document and map the patterns we are interested in and then process it as we want. The AWS service that allowed us easily and cost-effectively process vast amounts of data and compound all instances into one dataset was Elastic MapReduce. Eventually, we got a Large Matrix comprised of 21000 instances, where each instance is a web article.

Initially, we have 6 categories to classify each instance (in our case – web article) but later on the experiment we reduced the number of categories to 4: negative, somewhat negative, somewhat positive, positive. This allowed us to significantly increase the accuracy and confidence level in our further inferences. After applying the developed prediction model to the Large Matrix, we got the following sentiment distribution:

Figure 1. Samsung Galaxy Sentiments Distribution

Figure 2. iPhone Sentiments Distribution

Figure 3. Numeric table.

|  |  |  |  |
| --- | --- | --- | --- |
| **Sentiment** | **iPhone** | **Samsung Galaxy** | **Diff** |
| **Negative** | 9,326 | 9,092 | 2.5% |
| **Somewhat negative** | 580 | 590 | -1.7% |
| **Somewhat positive** | 1,334 | 1,353 | -1.4% |
| **Positive** | 10,568 | 10,773 | -1.9% |

What we can see here that not only there is no clear answer which brand is better in terms of sentiments, but also there is no unambiguous answer on the customers’ attitude towards the devices. The distributions of the sentiments are equal for both brands. The number of negative reviews is just slightly less than positive and we cannot certainly state whether the device is likable by users or not.

**CONFIDENCE**

To evaluate the confidence of the prediction model we used 2 metrics – Accuracy and Kappa. The second one supported the confidence of the first one and told us how well the accuracy works taking into account that the right prediction could be acquired by chance. All experiments and model development were conducted on the small matrixes (13k raws). Here is a comparison table for all considered models:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Model** | **Feature Engineering** | **Accuracy** | **Kappa** |
| **SAMSUNG GALAXY** | C5.0 (Decision Tree) | - | 0.764 | 0.526 |
| SVM | - | 0.701 | 0.361 |
| kkNN | - | 0.689 | 0.438 |
| Random Forest | - | 0.768 | 0.532 |
| Random Forest | without highly correlated features | 0.736 | 0.467 |
| Random Forest | recursive feature elimination | 0.757 | 0.514 |
| Random Forest | recode target variable (6 -> 4) | 0.847 | 0.600 |
| Random Forest | PCA | 0.757 | 0.515 |
| Random Forest | Recode + PCA | 0.834 | 0.573 |
|  |  |  |  |  |
| **IPHONE** | C5.0 (Decision Tree) | - | 0.770 | 0.553 |
| SVM | - | 0.712 | 0.417 |
| kkNN | - | 0.307 | 0.152 |
| Random Forest | - | 0.774 | 0.562 |
| Random Forest | without highly correlated features | 0.736 | 0.492 |
| Random Forest | recursive feature elimination | 0.776 | 0.566 |
| Random Forest | recode target variable (6 -> 4) | 0.851 | 0.630 |
| Random Forest | PCA | 0.763 | 0.545 |
| Random Forest | Recode + PCA | 0.842 | 0.607 |

As we see, the maximum accuracy that we managed to get is around 85% for both datasets. The kappa scores are close to each other and equal 0.6 for Galaxy and 0.63 for iPhone. These values of kappa can be evaluated as a middle value for Moderate and Good agreement between model's predictions and true values, so we can be quite confident that the given accuracy is plausible and it could be used to evaluate the model applied to the Large Matrix.

**IMPLICATIONS**

According to the results acquired, the final goal of the client is not achieved if we state the goal the way it was stated before the research. As we can see, Sentiments Distribution tables for both brands do not give an advantage to any of them. Moreover, the distribution of negative and positive reviews is so close (I’d say identical), that we cannot give preference to any of brands. However, these inferences let us suggest using both brands because there is no difference between them in the matter of customers' attitude. Thus, Helio can develop its apps suite for any device they want, which is the most convenient for them in terms of code writing or other factors.

**METHODOLOGY**

The methodology for this research is partially declared at the beginning of this report, however, here is a concise summary points of it:

1. Over 21000 web articles were inspected and compounded into one dataset, so-called “Large Matrix”. Each part of the device was labeled either a negative or positive way by the script.

2. Small Matrixes for both brands were manually gathered and labeled by people.

3. Prediction models were developed and evaluated using Small Matrixes.

4. The best model for each brand was applied to the Large Matrix.

5. Sentiment distributions were evaluated and used for answering the project question.

**LESSONS LEARNED**

The best classifier for both small datasets was Random Forest. Other models were either very close (C5.0) or extremely far (kkNN) in terms of accuracy and kappa score.

During the cross-validation stage, the features suggested by the Recursive Feature Elimination algorithm showed the best performance, however, after applying these models to the testing datasets, the whole best result gave a random forest model that exploited all available variables. Due to this, I decided to keep all variables in place. Also, I did not drop any close-to-zero variance variables since it never gave an improvement in my practice. There were also no variables with zero variance.

What definitely boosted the accuracy was recoding the target variable and reducing the number of categories from 6 to 4. For this task, it was acceptable because we do not need so many levels of sentiments.

It was interesting to apply PCA here, but unfortunately, some information lost after this and the accuracy and kappa dropped compared to other models. Probably, this dataset is not suitable for this kind of transformation.

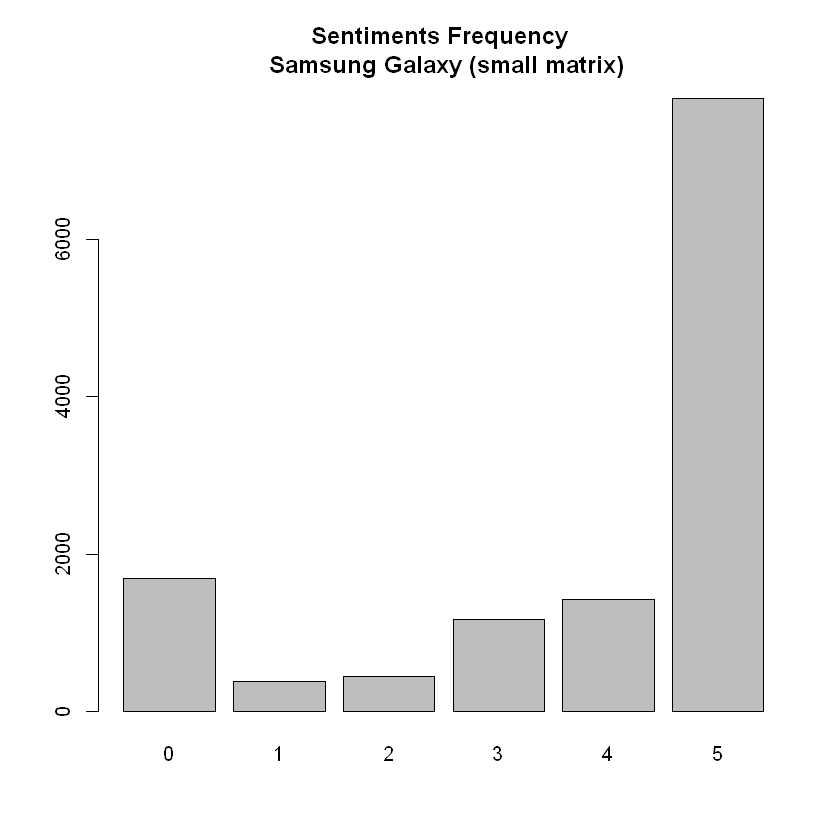
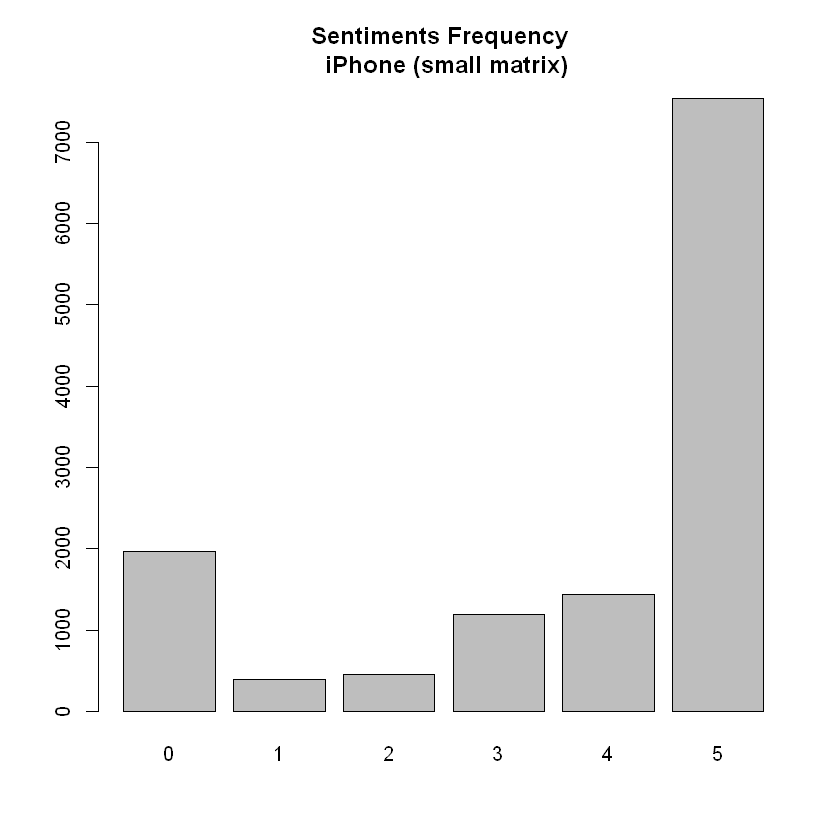
Additional plots and all experiments conducted can be found on my GitHub repository:

<https://github.com/ms888ekb/sentiment_r>

Please, feel free to explore.

**APPENDIX**

I have no reason to not to believe those people who manually gathered and labeled data for the small matrixes, but one fact says they probably did their job incorrectly. Here are two sentiment distribution plots for small matrixes of both brands:

Without any special analysis, it is clear that both brands have an equal distribution percentage of each sentiment category. For me, this is extremely suspicious, because this never happens in the real world. Thus, it makes sense to double-check this input data. Also, this anomaly led to almost identical results on the Large Matrix and hindered us from giving a precise answer and give select one brand.