Alert Analytics

Helio Project

Sentiment Analysis

Floriana Trama

Data analysis department

AGENDA

- **General Overview**
- **Small matrices: Pre-process, feature selection and engineering, model development**
- **Large matrix: Dataset development and Sentiment prediction**
- **Conclusion & Recommendations**
- **Appendix**

GENERAL OVERVIEW

Description of the dataset
Three different datasets are used for the prediction of the sentiment toward two different mobile devices, namely iPhone and Galaxy:

- **"iPhone Small matrix" and " Galaxy Small matrix"** containing respectively **12.973** and **12.911 observations** representing the webpages from the Common Crawl related to the analysis
- **"Large matrix"** containing **25.207 observations** representing a set of relevant web documents from the Common Crawl.

In all the datasets, each data point comprises **59 attributes**, in particular:

- attributes containing information about the relevancy of the webpages toward each device (e.g. iPhone, SamsungGalaxy)
- attributes collecting information about the sentiment related to the operating system used on the phone (iOS, Google Android)
- attributes collecting information about the sentiment related to the phone's camera (positive, negative, uncertain; e.g. iPhone Camera Positive → iphonecampos, Samsung Camera Negative → samsungcamneg)
- attributes recording information about the sentiment toward the phone's display (positive, negative, uncertain; e.g. iPhone Display Negative → iphonedisneg, Samsung Display Positive → samsungdispos)
- attributes recording information about the sentiment toward the phone's performance (positive, negative, uncertain; e.g. iPhone Performance Uncertain→ iphoneperunc, Samsung Performance Uncertain → samsungperunc)

Although the datasets don't present missing information, they **need some pre-process and feature selection activities** to make them ready for the analysis.

In particular, in order to predict the sentiment toward iPhone and Samsung Galaxy mobile devices, a classification analysis has been performed by building two models based on the "small matrices" data. Then, these models have been applied to a "large matrix" obtained from the Amazon Web Service through the Common Crawl, an open repository of web crawl data stored on Amazon's public data sets.

Objective of the analysis: perform a classification analysis to estimate the sentiment toward both iPhone and Samsung Galaxy and advise Helio on which device is best suited for developing the bundle with its app

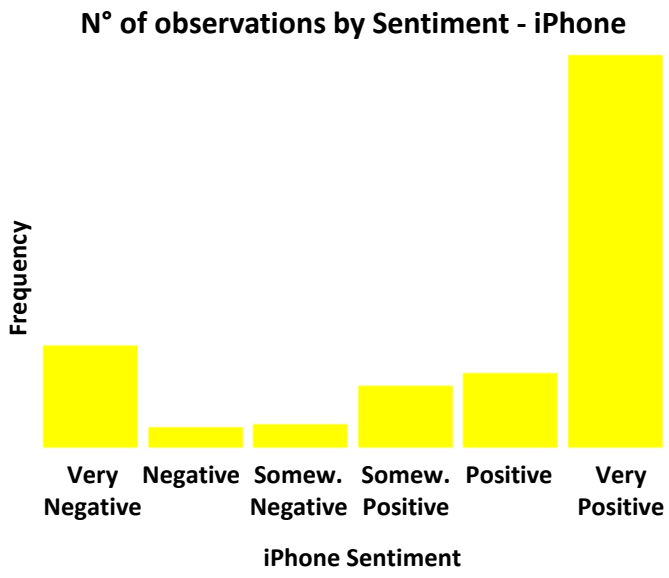
SMALL MATRICES: Pre-process, feature selection and engineering, model development

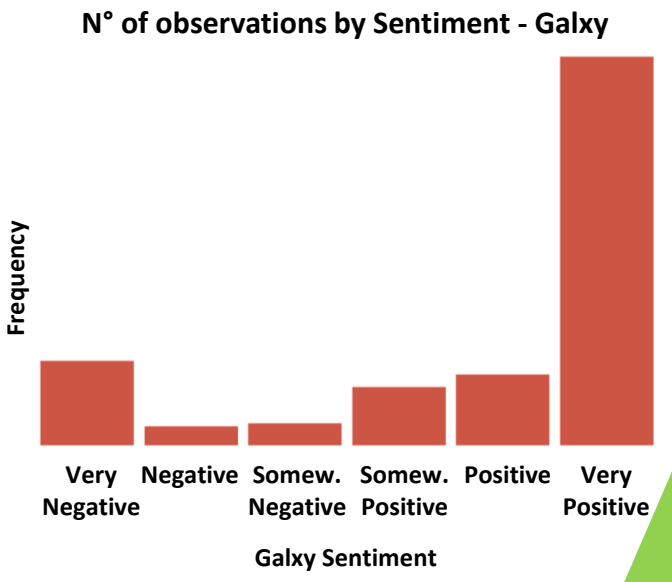
Both small matrices have been created by scanning the Common Crawl for documents that express meaningful sentiment about one of the phones and then, for each relevant document, by collecting information about the sentiment toward key features of the phone, namely: camera, display, performance, and operating system.
The sentiment variables that record the phones' general sentiment have been manually labeled by the analytics team by reading and analyzing each webpage.

The analysis starts with performing some pre-process activities in order to have a better understanding of the data (outliers, missing values, observation distribution) and proceed with the feature selection and engineering in both the small matrices:

- In the **iPhone Small Matrix**, all the variables related to iPhone have been selected with the exception those concerning the operating system (i.e. iOS) as these are too generic and could be related to other Apple devices thus creating noise in the analysis. For the same reason, all of the webpages only containing the word "iPhone" and missing any other information linked to the sentiment toward the device, have been deleted. Finally, the sentiment attribute has been reduced from 6 levels (0: very negative, 1: negative, 2: somewhat negative, 3: somewhat positive, 4: positive, 5: very positive) to just to 2 levels (0: negative, 5: positive) so as to simplify the following modeling activity;
- In the **Galaxy Small Matrix**, all the attributes related to Galaxy device have been selected apart from the attributes related the operating system (Google Android and Google Android performance) and the attribute "Samsung Galaxy"; this latter is too generic and cause noise into the dataset. Moreover, all the webpages with a number of relevant words smaller than 2 have been deleted because they don't add value to the analysis. Similarly to the iPhone matrix, the sentiment attribute has been reduced from 6 to 2 levels (0: negative, 5: positive).

Both datasets feature an unequal distribution of observations among the different sentiment classes making the databases "not balanced": reducing the sentiment classes to 2 levels and using both "oversampling" and "undersampling" methods (applied only in the Galaxy matrix) have helped in getting better balanced datasets.





After implementing these indispensable activities, **several classification models are built** by training and testing **different algorithms**, namely **KNN, KKNN, SVM, C5.0 and Random Forest** in order to **identify the top performer** in terms of Accuracy and K.
In both cases, the best performing algorithm is the Random Forest showing the following metrics:

- iPhone Small Matrix → Accuracy: 89,92% and K: 0,7433
- Galaxy Small Matrix → Accuracy: 87,32% and K: 0,7475

LARGE MATRIX: Dataset development and Sentiment prediction

The Large Matrix has been created by using the Common Crowl in the Amazon Web Service by running two clusters, obtaining more than 25K webpages with relevant words for assessing the sentiment of the two devices.

The model are then applied to the Large Matrix to get the sentiment predictions:

iPhone	
Negative Pred.	**Positive Pred.**
10,798	14,409

Galaxy	
Negative Pred.	**Positive Pred.**
161	25,046

Following the predictions above, it emerges that there is a wider positive sentiment toward Samsung Galaxy than iPhone on the Web, hence it is advisable to proceed with the development of an app optimized for Samsung to include in the bundle.

We are very confident about the results of our analysis: the approach used for collecting the observations, labelling the attributes, cleaning and making our datasets balanced, the features selection done, the metrics obtained and the data mining effort, reassure that the predictions are based on logical and clear basis.

CONCLUSIONS & RECOMMENDATIONS

The classification analysis has been conducted with the objective of understanding the sentiment toward two different mobile phones, iPhone and Samsung Galaxy, and advise our client Helio about which device to choose for its bundle activity.

Our analysis shows that there is a wider positive attitude toward Samsung Galaxy than iPhone on the Web, thus the former is the best candidate to focus on and be launched in bundle with the developed app.

APPENDIX

- Confusion Matrix

iPhone	Reference	
Prediction	Negative	Positive
Negative	287	12
Positive	123	917

Galaxy	Reference	
Prediction	Negative	Positive
Negative	30	2
Positive	7	32