weBigData&DataAnalytics

Sentiment Analysis

Mauricio Carvajal
HELIO AND ALERT! ANALYTICS

Table of contents

1. Overview Section

1.1 Project at a high level.

The overall project consists in 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). Due to this situation, they need to choose only 1 model of smart phone. Due to that our main goal consist in provide a report that contains an analysis of sentiment toward the target devices, to be able to recommend which is the best option to start the project.

1.2 Client

1.2.1 What do they do?

Helio is governmental health agency, that wants to improve the health of the development countries thru the usage of technology.

1.2.2 Goal of the project of Helio

The main goal of the Helio, is to create a medical app for aid workers, to help the local health of developing countries connecting medial professions and patients thru smart phones, to facilitate the evaluations, diagnoses.

1.3 The Task IOT Analytics

The task for IOT Analytics is to conduct a sentimental analysis of the different smart phones brands in order to choose what is the optimal model to start development the medical app. For that they want to want to

2. Findings Section

- Paragraph comparing your comparative findings for iPhone and Galaxy.

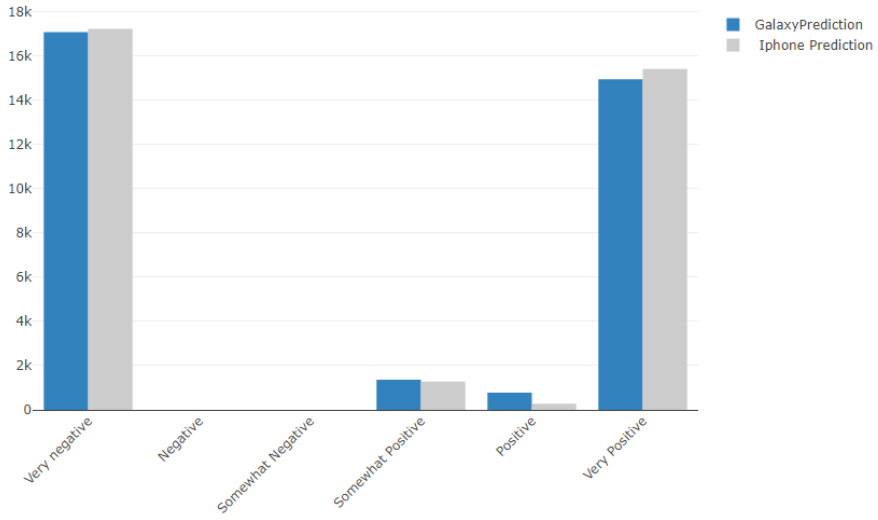
2.2 Sentiment categories

After a lot preprocessing and trying to find the best way to compare the two categorization models that had been used for the predictions the sentiments that has been used are:

- very negative
- negative
- somewhat negative
- somewhat positive
- positive
- very positive

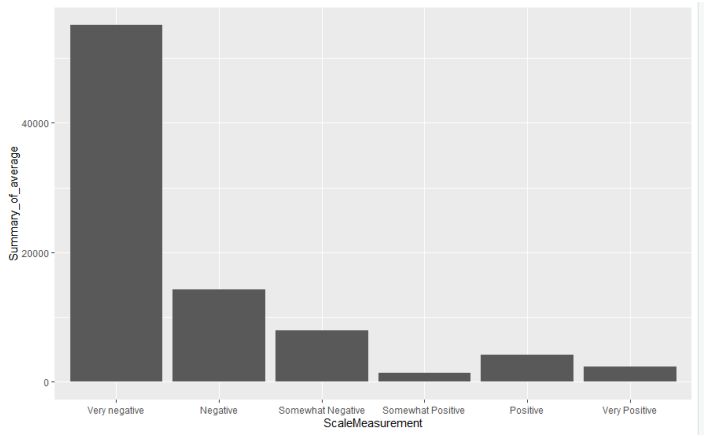
2.3 Side by Side comparation

After the preprocess that has been done, the results of each model are shown in the next figure, in which we can find that present similar behavior. But there are 2 main findings, iPhone has a tendency of being loved or hated. In the case of galaxy if has an overall more positive acceptance.



2.4 Raw sentiment count for iPhone and galaxy

The following figure show the raw data that has been used for the sentimental analysis that has been develop thru this project. The distribution of this show us overall there is a very negative tendency that can impact our analysis.



3. Confidence Section

Paragraph reporting on your confidence in your modeling results. What were the accuracies from your modeling? This should not be overly technical.

4. Implications Section

Paragraph stating your opinion regarding the impact of your results on the client's goals.

In my opinion, base on the results, the client should invest on the galaxy cellphones, due to several factos.

- Based on the results, it has a more positive reception trend than iphone.
- Also is know tha galaxy uses android as operative system, so it can be easy include the app to others brands

5. Methodology Section

The overall methodology that has been followed to complete this analysis; it has been divided in 3 main sections.

The first part is having a data set that has been labeled, this means a set of information that is going to allow us to train our models, that then those will be used for the prediction and final analysis.

Once we have this data, we need to preprocess the information, for this this, the approach that we follow, is to use 4 different preprocess methods (Recursive Feature Elimination, Correlation Elimination, Feature Variance (NZV), recoding sentiment).

After have those data sets already defined, we use the method called “"Out of the Box" Model Development”, that consist in choose 1 data set and test with different classifiers to validate which is the best model, in this case we used 4 different algorithms of machine learning:

- C5.0
- Random Forest
- SVM (from the e1071 package)
- kknn (from the kknn package)

In order to evaluate which model is the best, different characteristics has been taking in consideration, for example:

- **Kappa, Accuracy:** This tell how good the model is,
- **Test time:** This help us understand how much time takes the model to run.

For the development of this project is has a tradeoff between the how good is the model and how much time takes to complete, due to if the differences in the model in kappa, Accuracy is less 5%, but the time to finished are more 100%.

After those analsys, the large data set has been imported, and use the choosen model to make the predicionts and make conclusions base on the data.