***Data Analytics Report***

**„ Alert Analytics “- Predicting Sentiment**

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date** | **Author** | **Comments** |
| 1.0 | Sep 12th 2019 | Steffen Adolf | Initial write |
|  |  |  |  |
|  |  |  |  |

Task: Analysing Sentiment on the web towards a number of Smart Phones

**1. Subject of Analysis**

Helio, a smart phone and tablet app developer, wants the support of Data Science to find insights about the current public sentiment towards different types of Smartphones.

It is currently working with a government health agency to create a smart phone medical app. The app will be used by aid workers in developing countries. The task of the app is to connect aid workers with medical professionals located elsewhere.

One of the governments demands is that the app is bundled to just one smartphone type. To choose the most appropriate Smartphone type for the specific tasks of this app, Helio wants the support of Data Analytics.

Data Analysis can show the general public sentiment towards different types of Smartphones and its features by examining the prevalence of positive and negative attitudes toward these devices on the web.

This report contains an analysis of sentiment toward the target devices, as well as a description of the methods and processes that has been used.

**2. Working process and used methods**

There are a number of ways to capture sentiment from text documents. For this project sentiment has been captured simply by counting words in text documents that are associated with a certain sentiment towards these Smartphone types within relevant documents on the web.

With the help of machine learning tools, patterns of words that represent a specific sentiment have been calculated, so that these algorithms can be used to label other text documents, where different combinations of these words about phones and reviews of these phones can be found.

At the end the public sentiment towards different phone types have been compared, so that Helio can use these insights for their decision-making process.

In order to really measure the sentiment toward these devices, data had to be captured on a large scale. Therefore, the cloud-computing platform of Amazon (AWS) had been used. The data set itself came from “Common Crawl”, an open repository of web crawl data.

In a first step relevant documents have been filtered out (from these over 5 billion pages collected by Common Crawl), that are relevant for the analysis.

In a second step these relevant documents have been checked for word combinations that represent information about the sentiment toward key features of the phone (camera, display, performance, and operating system).

The patterns of sentiment towards each of these features then can be used to build a model that defines a general sentiment towards the phone itself (positive, negative and in between).

**3. Modeling Process**

Over 34.000 relevant documents (webpages) have been found for August 2019 by using the cloud-computing platform AWS from Amazon for web crawling.

With the help of a manually labelled data set, that gives a summarized sentiment towards a phone type with a value between 0 (very negative) and 5 (very positive) , machine learning algorithms have been build and could then been used on this collected data to make predictions about the general sentiment of these websites too.

The following models have been trained and tested:

**1. Support Vector Classifier**

**Accuracy = 0.76 Kappa = 0.52**

**2. K-nearest-Neighbours**

**Accuracy = 0.74 Kappa = 0.51**

**3. Random Forest Classifier**

**Accuracy = 0.76 Kappa = 0.54**

**4. Neural Network**

**Accuracy = 0.76 Kappa = 0.54**

**5. Logistic Regression**

**Accuracy = 0.71 Kappa = 0.43**

Reducing the number of categories from 6 to 3 into ‘positive’, ’negative’ and ‘neutral’ has helped to improve the accuracy of the models:

**1. Support Vector Classifier**

**Accuracy = 0.84 Kappa = 0.59**

**2. K-nearest-Neighbours**

**Accuracy = 0.84 Kappa = 0.59**

**3. Random Forest Classifier**

**Accuracy = 0.84 Kappa = 0.60**

**4. Neural Network**

**Accuracy = 0.84 Kappa = 0.60**

**5. Logistic Regression**

**Accuracy = 0.79 Kappa = 0.44**

Using regression models instead of classifying the sentiment has produced worse results. For example:

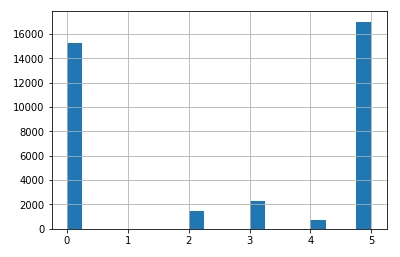
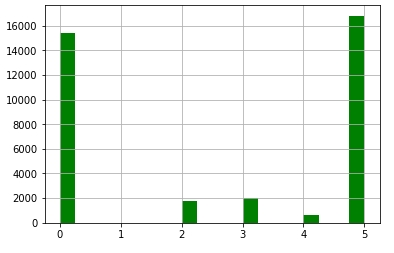
**Neural Network**

**mean absolute error: 1.16 r2\_score: 0.396**

Feature selection and the reduction of features did not improve the results either.

That’s why I have chosen to use the Neural Network model for classification with 3 and with 6 classes for making predictions about the sentiment in the new unlabelled data set.

**4. Results**

 **iPhone Samsung Galaxy**

Average sentiment towards Iphone and Samsung Galaxy for August 2019

**Iphone: 2.66 Galaxy: 2.62**

Classification with 3 classes:

**Iphone: 3.16 Galaxy: 3.18**

The results are not clear enough to give an unambiguous recommendation. Taken into account that the accuracy of the model for 6 classes was 0.76 and 0.84 for 3 classes, the results are in such a narrow margin, where errors can tip both phone types on first or second place randomly.

That’s why we can conclude that for the month of August 2019 the sentiment towards the iPhone and Galaxy Samsung are equally distributed for the data used.