Aline Barbosa Alves 2019-11-19

Predict sentiment report

Smart phone preference

Introduction

Alert Analytics was asked to do a sentiment analysis for a client to compare the sentiment between 5 types of smart phones focusing in iPhone and Galaxy. To do that, we mapped around 800 websites from Common Crawl using AWS and then created a matrix with 58 features. These features contains the number of times the 'positive', 'neutral' or 'negative' words were found with some characteristics from these smart phones.

This report covers this analysis and its objective is to help our client to decide which type of app should be better to focus on.

General Insights

In order to predict the sentiment from the result of our work in AWS, we received two data sets where the sentiment was labeled manually by our colleagues from Alert Analytics. These data sets are in the same format as the result from AWS. The iPhone data contains results from 12973 websites while the Galaxy data contains 12911. The sentiment is categorized from 0 (very negative) to 5 (very positive). We can see the distribution at the figures 1 and 2 below:

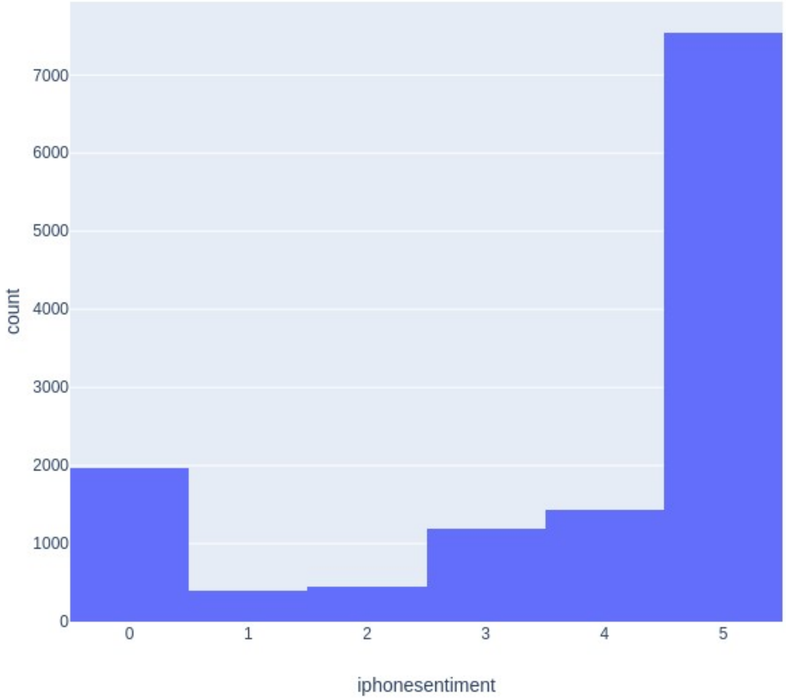

Figure 1: Distribution of the sentiment categories from iPhone manually labeled data.

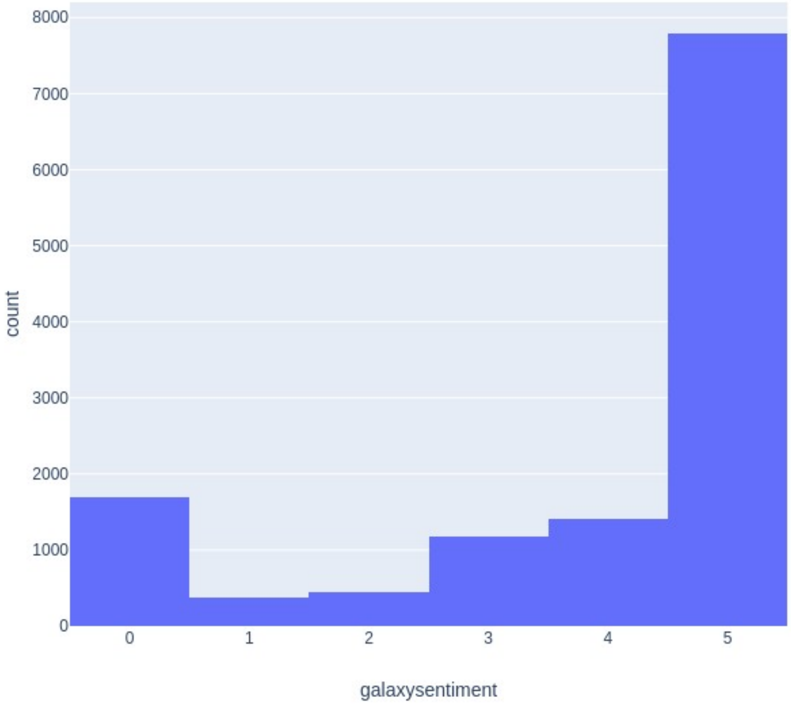

Figure 2: Distribution of the sentiment categories fromGalaxy manually labeled data.

After get to know the data, we chose to develop 3 models: Random Forest, SVM and kNN to evaluate the ‘out of the box’ data. The results are as ‘iphone’ and ‘galaxy’ in the Table 2.

Then we took off the high correlated features. To do that, we calculated the correlation between the features and chose to keep only correlations lower than 0.9. In Table 1, we can see which features were removed. The results are found in Table 2 as ‘cor’.

iPhone	Galaxy
Nokia camera unclear	Nokia camera unclear
Nokia camera positive	Nokia camera positive
Nokia performance unclear	Nokia performance unclear
Nokia performance negative	Nokia performance negative
Nokia display positive	Nokia display positive
ios	ios
ios performance unclear	ios performance unclear
ios performance negative	ios performance negative
Samsung display unclear	Samsung display unclear
Samsung display negative	Samsung display negative
Google performance negative	Google performance negative
HTC display positive	HTC display positive
	Sony display negative

Table 1: Features with correlation > 0.9 that were removed.

As the results did not improve, we tried to use recursive feature selection to keep only the most important half of the features. These results are in Table 2 as ‘rfe’.

Figures 1 and 2 shows us that there are much more positive reviews than negative. To solve this problem, we decided to try over and under sampling. So we chose the ‘cor’ data to do that because it were the best results so far. We randomly over sampling the data by duplicating rows until all the categories have the same number of ‘very positive’ (or 5) and randomly under sampling by picking random rows until all the categories have the same number of ‘negative’ (or 1). The result for random over sampling are as ‘ros’ and for random under samplig are as ‘rus’.

As the results were not getting better, we decided to use only 3 categories: negative (0 and 1), neutral (2 and 3) and positive (4 and 5). Following we can see the figures 3 and 4 with new distribution.

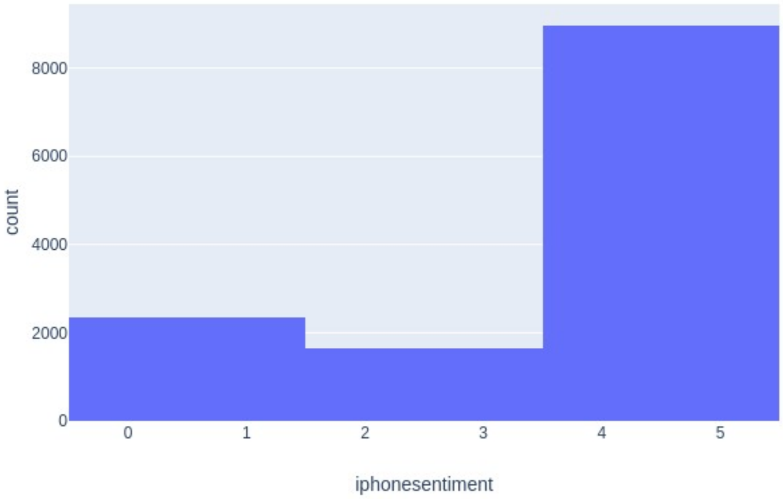

Figure 3: Distribution of the sentiment from iPhone with 3 categories.

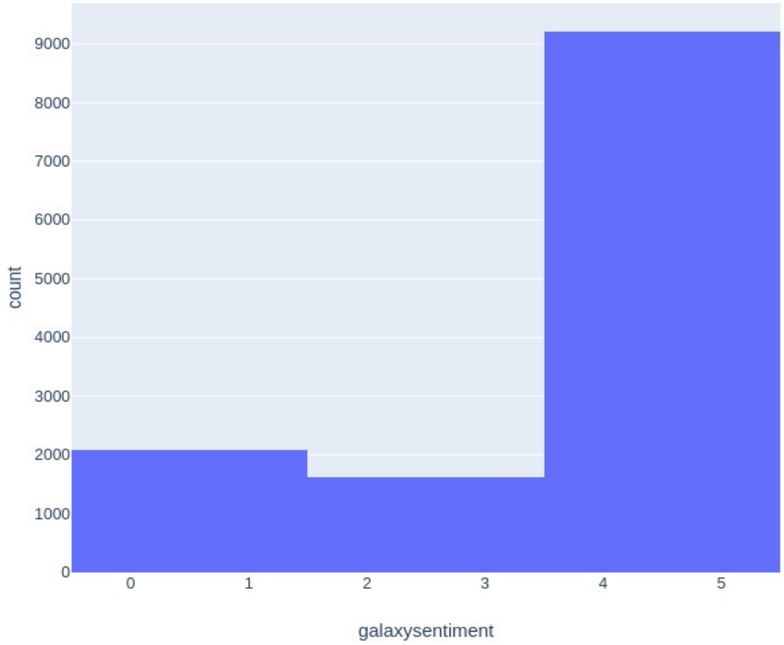

Figure 4: Distribution of the sentiment from Galaxy with 3 categories.

Comparing the results ‘ros’ and ‘rus’ with the other, we can see that got worse but over sampling was better than under sampling. So we used synthetic minority over sampling technique until we have 100000 rows per each of the 3 categories. The results are found as ‘smote’ in Table 2. We tried only random forest with these data because it were to big and we got better results with this model on the other tries.

Data	RF		SVM		kNN	
	accuracy	*kappa*	*accuracy*	*kappa*	*accuracy*	*kappa*
iphone	0.7685	0.5619	0.7364	0.4855	0.7038	0.4743
iphone_cor	0.7685	0.5620	0.7426	0.5006	0.7022	0.4730
iphone_rfe	0.6942	0.3885	0.6816	0.3433	0.3859	0.1684
iphone_ros	0.5745	0.4900	0.5019	0.4036	0.5316	0.4374
iphone_rus	0.4701	0.3671	0.4325	0.3227	0.4427	0.3340
iphone_smote	0.7552	0.6327	-	-	-	-
galaxy	0.7574	0.5207	0.7389	0.4621	0.7509	0.5128
galaxy_cor	0.7581	0.5229	0.7392	0.4627	0.7509	0.5135
galaxy_rfe	0.7017	0.3749	0.6989	0.3396	0.3398	0.1694
galaxy_ros	0.5504	0.4606	0.4713	0.3659	0.5270	0.4310
galaxy_rus	0.4363	0.3190	0.3770	0.2600	0.3909	0.2750
galaxy_smote	0.7508	0.6262	-	-	-	-

Table 2: Results to compare the all feature selection, engineering and models.

After all those tries, we used the last one to apply on the data collected from Common Crawl. Below are the figures 5 and 6 with the final results.

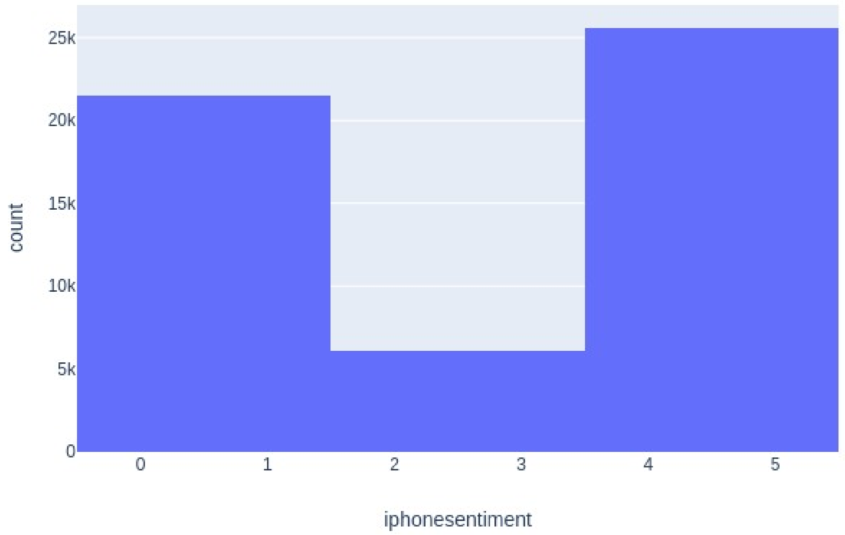

Figure 5: Distribution of the predicted sentiment from iPhone with 3 categories.

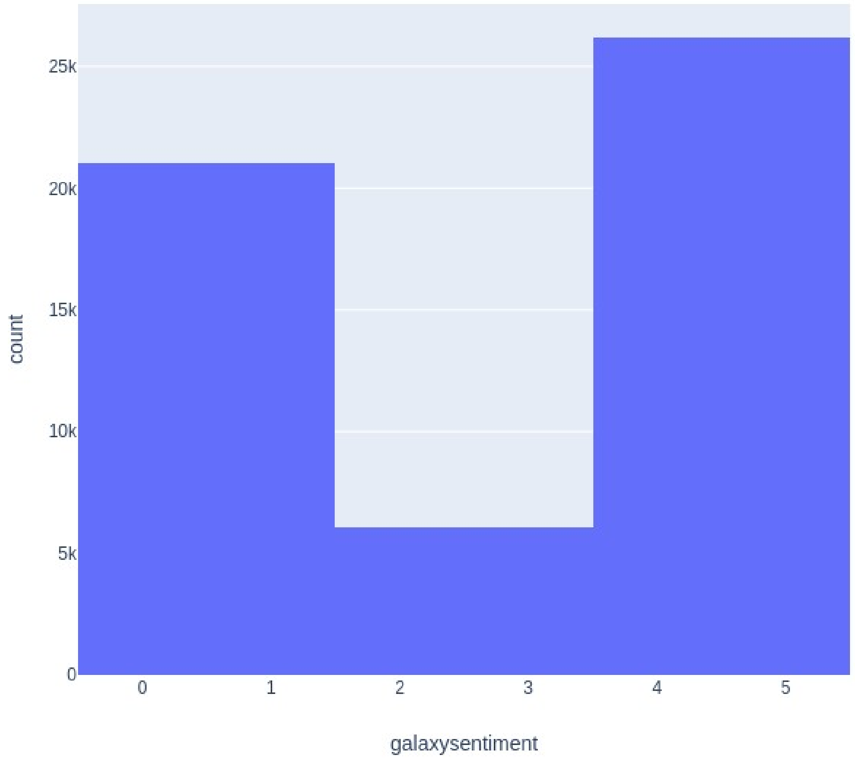

Figure 6: Distribution of the predicted sentiment from Galaxy with 3 categories.

As we can see, the results are pretty similar but galaxy has a slightly preference in the reviews.

Conclusions

After the analysis, we can conclude that the clients have almost the same perception between iPhone and Galaxy. But considering that galaxy has a little preference and it is cheaper, we recommend to focus in a development of an android app.