

SENTIMENT PREDICTIONS FOR APPLE IPHONE AND SAMSUNG GALAXY – LESSONS LEARNED

October 15, 2020

Data characteristics

The data attributes mainly consist of counts of positive, negative, and unclear mentions regarding the operating system, camera, display, and the performance for iPhone, Samsung Galaxy, Sony Xperia, Nokia Lumina, and HTC Phone.

Training and testing of predictive models were performed on the data sets with manually determined overall sentiments toward iPhone and Samsung Galaxy grouped as following:

“0” – very negative

“1” – negative

“2” – somewhat negative

“3” – somewhat positive

“4” – positive

“5” – very positive

After the best performing algorithm was identified, sentiment predictions for iPhone and Samsung Galaxy have been generated based on the data set from Amazon Web Services (Common Crawl files).

Model development

Multiple combinations of data pre-processing techniques have been applied. The performance of each combination was checked against testing sets with `postResample()` and `confusionMatrix()`. The goal was to achieve higher Accuracy, but, most importantly, higher Kappa score.

The first step was to condense 6 sentiment categories to 4 groups: “1” – negative, “2” – somewhat negative, “3” – somewhat positive, “4” – positive.

After condensing the categories, category “2” (somewhat negative) represented only 3.5% of the total observations for both data sets. Synthetic Minority Oversampling Technique was applied to rebalance the sets.

The following feature selections were applied:

1. Handset specific features have been selected manually - iPhone related features for iPhone set and Samsung Galaxy related features for Samsung Galaxy set.
2. `NearZeroVar()` was used to eliminate zero/near-zero variance features.
3. Recursive Feature Elimination was applied to identify important features.

Principal Component Analysis was applied to each set in order to standardize the features.

Recursive Feature Elimination along with the Principal Component Analysis delivered the best results.

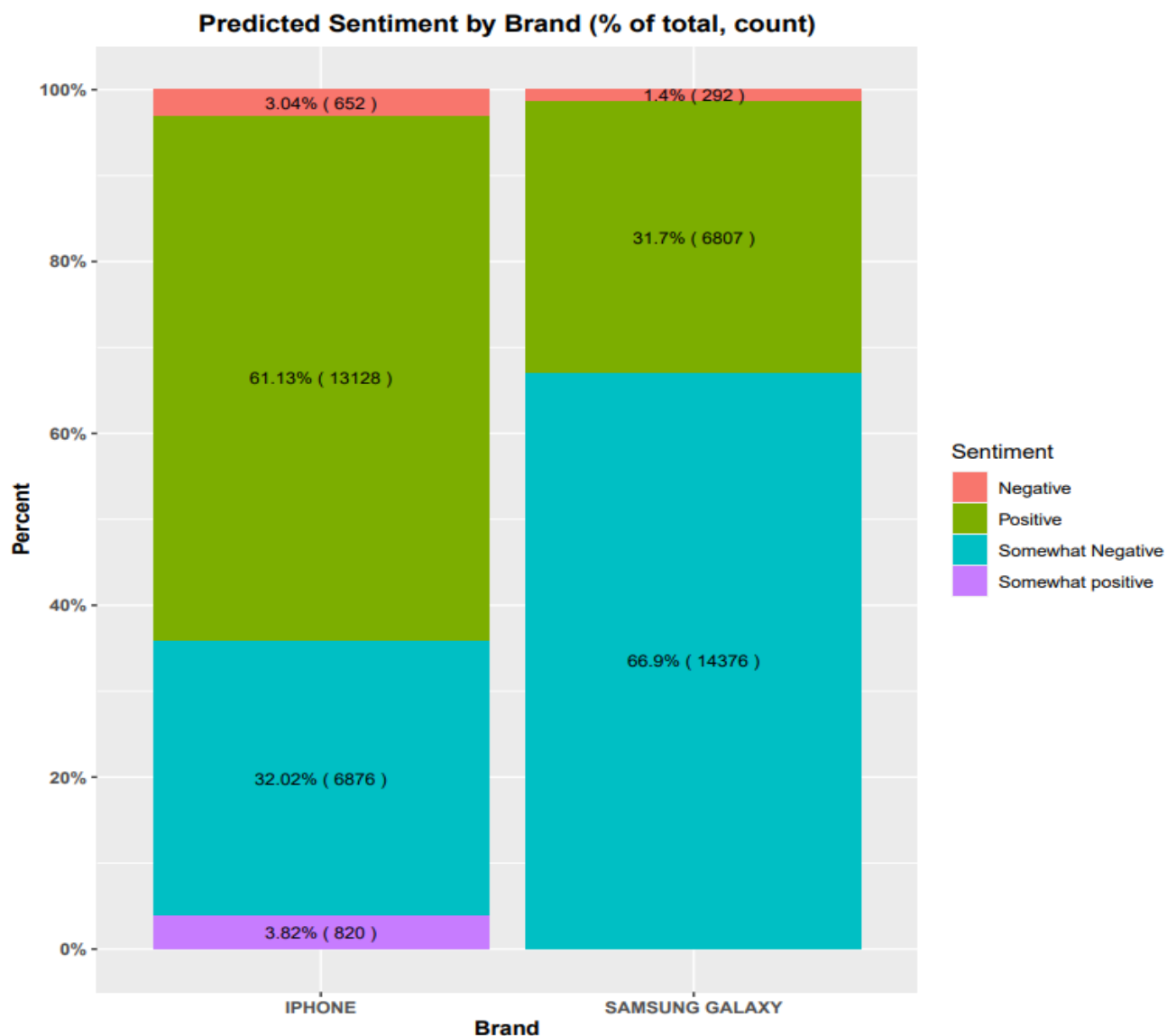
SENTIMENT PREDICTIONS FOR APPLE IPHONE AND SAMSUNG GALAXY – LESSONS LEARNED

During testing the selected models showed the following performance:

MODELS	PERFORMANCE METRICS			
	IPHONE		SAMSUNG GALAXY	
	ACCURACY	KAPPA	ACCURACY	KAPPA
K- Nearest Neighbor	0.7787	0.6099	0.8153	0.5923
Support Vector Machine (radial kernel)	0.7485	0.5524	0.7436	0.3927
C5.0	0.7858	0.6221	0.8126	0.5856
Random Forest	0.7999	0.6502	0.8299	0.6264

Final predictions

Below is the chart with sentiment predictions for both handsets.



SENTIMENT PREDICTIONS FOR APPLE IPHONE AND SAMSUNG GALAXY – LESSONS LEARNED

Summary

It was difficult to achieve high Kappa score in this project. Sentiment categories were manually entered into the training/testing data sets. Therefore, it was impossible to check those categories for accuracy.

Predictions for category “2” (somewhat negative) seem high because category “2” was only 3.5% of total observations. Synthetic Minority Oversampling Technique can be used to fine tune the ratio between the observations to build a model with higher Kappa score.

Please let me know if you have further questions.