Predicting Smartphone Sentiment

Hardik Shahu
Alert Analytics

Background

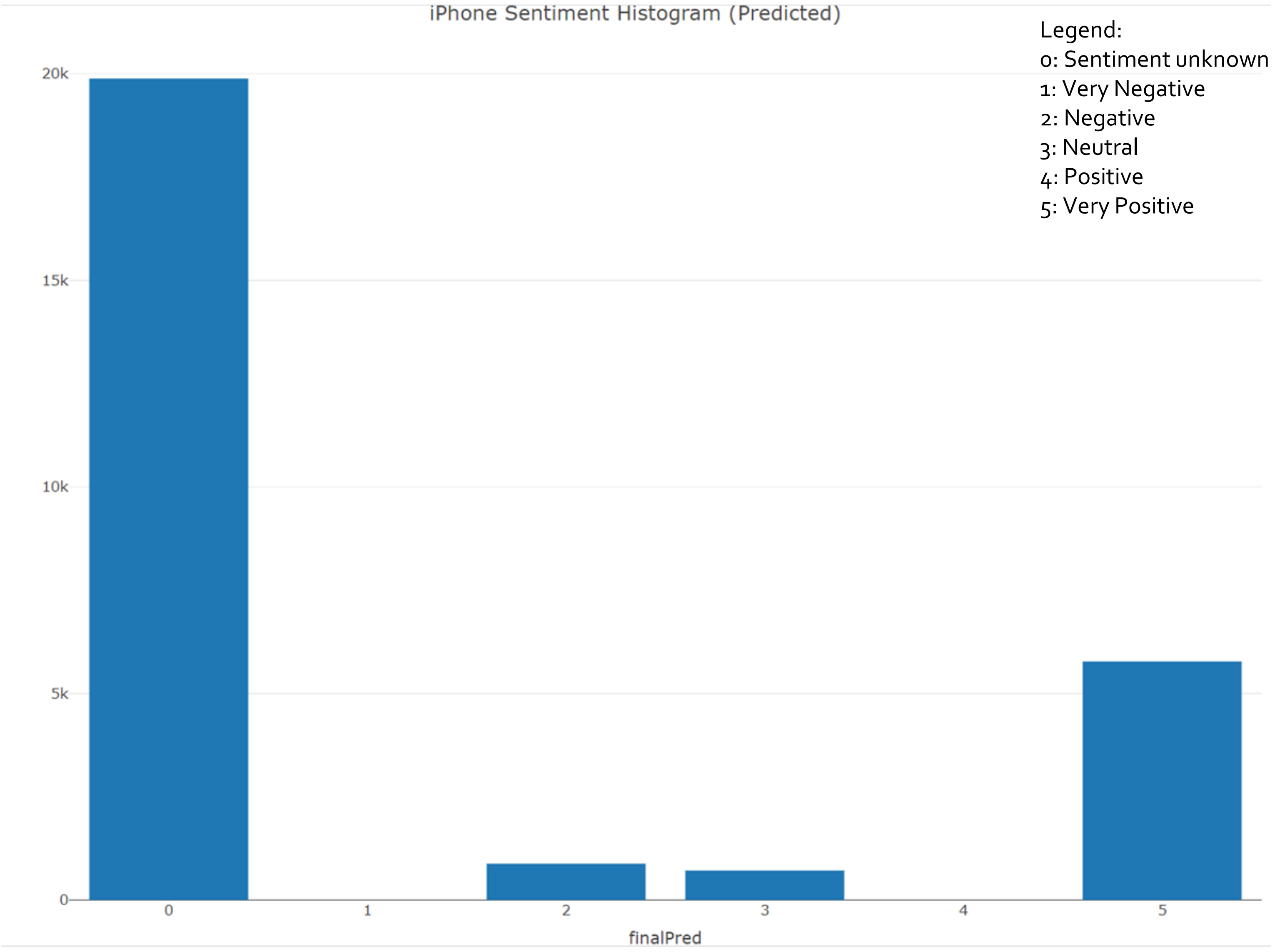
Helio, a prominent smartphone and tablet app developer, has embarked on a collaboration with a government health agency. Their mission: to develop specialized medical applications tailored for aid workers operating in developing countries. With a primary emphasis on delivering technical support services, Helio aims to streamline operations by restricting compatibility to a singular smartphone model. The chosen options are the globally recognized and widely utilized Apple iPhone or Samsung Galaxy, known for their prevalence and accessibility worldwide.

Task Overview

1. **Research Gathering:**
   a) Explore public opinions on Apple iPhone and Samsung Galaxy.
   b) Internet search for user sentiments and reviews.
2. **Data Compilation:**
   a) Consolidate collected information into a comprehensive spreadsheet.
3. **Machine Learning Analysis:**
   a) Utilize a distinct dataset to construct predictive models for user sentiment.
   b) Evaluate models to determine accuracy in predicting opinions on both phones.
4. **Model Selection & Application:**
   a) Choose the most precise sentiment prediction model.
   b) Apply the selected model to the gathered data from internet research.
5. **Final Decision:**
   a) Determine the recommended smartphone based on analyzed sentiments and model outcomes.

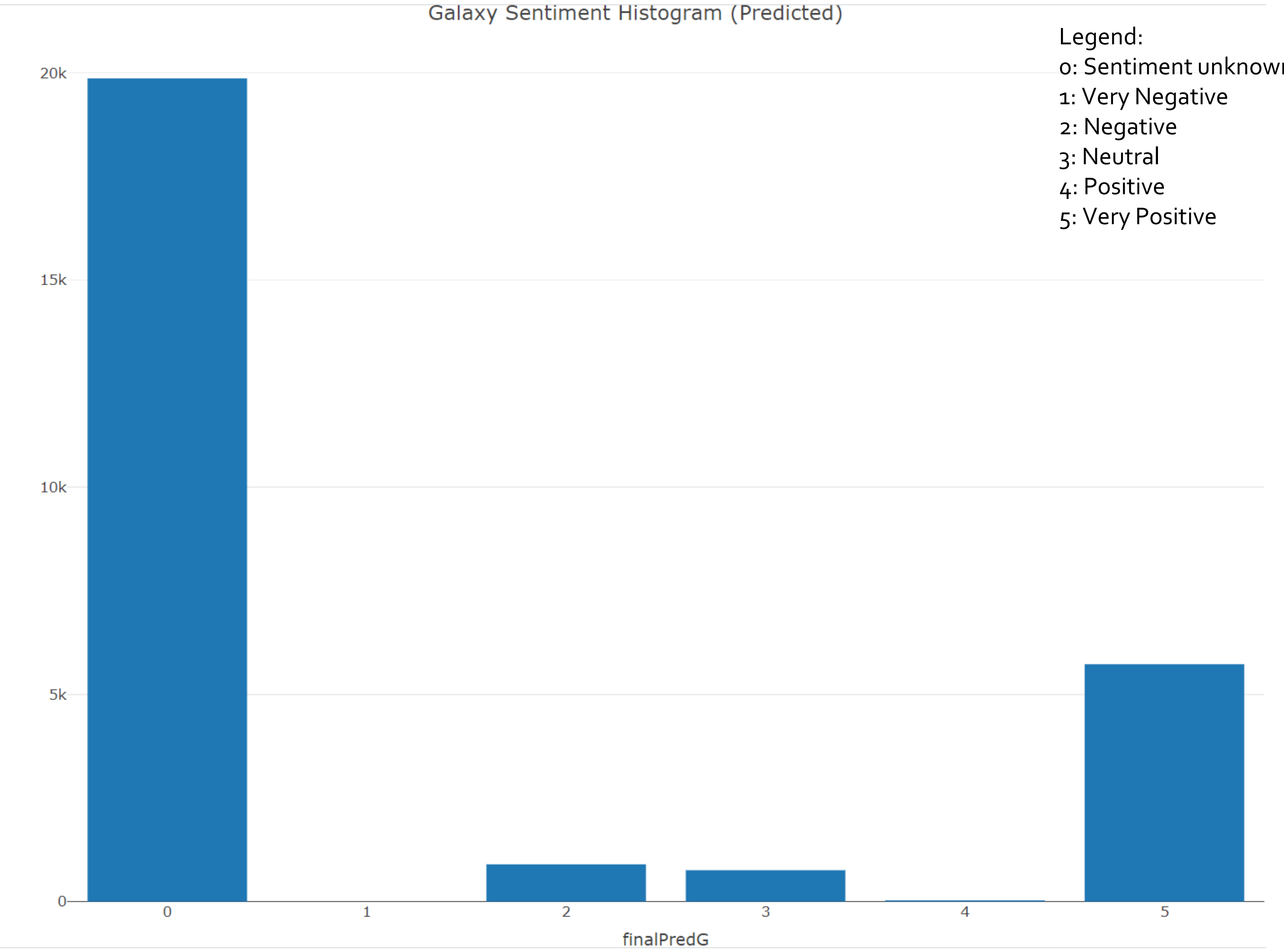
Findings – iPhone

Our investigation into iPhone user sentiments revealed intriguing results. A small minority expressed minor dissatisfaction or remained neutral about their device, while a noteworthy portion showcased very high positivity. Surprisingly, there were none for the very negative category (1) or positive (4). Regrettably, sentiments of most users remains unknown, posing a significant data gap in our analysis.



Findings – Galaxy

Our exploration into Galaxy user sentiments echoed similarities with the iPhone findings. Like the iPhone, a large segment of Galaxy users didn't express their sentiment. A small minority showcased some dissatisfaction, while a moderate portion conveyed high satisfaction with their device. Remarkably, akin to the iPhone, there none for 1. Yet, unlike the iPhone, a very small minority falls in 4.



Legend:
0: Sentiment unknown
1: Very Negative
2: Negative
3: Neutral
4: Positive
5: Very Positive

Results' Analysis

Surprisingly, both devices exhibited an equal sentiment, contrary to our initial expectations of one device being favored over the other. Compounding this unexpected parity is the overwhelming ambiguity in sentiments, representing a significant portion of users.

Delving into the "how" of these results leads to a logical realization. The ubiquitous nature of smartphones has transformed them into essential but unremarkable tools for most users. This functional perception might explain the prevalence of the large unknown segment in sentiment analysis.

Results' Analysis Cont.

Examining the known sentiments sheds light on a coherent pattern. We expected a moderate to very positive reviews for either phone as users who will tend to have bias on the product they purchased.

However, the absence of anticipated negative sentiments is puzzling. Typically, reviews tend to be polarized due to the inclination of individuals to vocalize strong opinions. While we anticipated a larger positive sentiment, the scarcity of negatives, even mild ones, suggests a subtler dissatisfaction that leads users to quietly switch devices rather than express discontent.

Confidence

We hold significant confidence in the accuracy of these findings. The model selected for prediction showcased robust performance during rigorous testing, bolstering our trust in the outcomes. Our process incorporated meticulous precautions to mitigate errors and ensure precise predictions. However, it's essential to acknowledge the slim possibility of oversight or error, although we deem this likelihood to be remote based on our comprehensive approach and testing procedures.

Our Recommendations

Regrettably, we are unable to recommend a single device due to two primary reasons. Firstly, a significant portion of users did not express their sentiments, making it challenging to ascertain a clear preference. Secondly, among the users with discernible sentiments, both devices exhibited comparable levels of satisfaction.

Our suggestion, albeit potentially demanding, is to extend support for both devices. Although this may increase workload for Helio and the agency, it presents a pragmatic approach given the circumstances. A silver lining emerges from this decision: aid workers will now have the freedom to choose between two models, ensuring a tailored experience. For instance, if someone doesn't favor the iPhone, they might find satisfaction with the Galaxy, thereby offering flexibility and personalization.

Lessons Learned Report - iPhone

Our analysis involved testing four variations of the dataset with distinct preprocessing approaches. Two datasets had features removed based on near-zero correlation and variance, while the "Recursive Feature Selection" (RFE) automated feature selection process identified optimal combinations. The fourth dataset remained unaltered.

We evaluated these datasets using four classifiers: C5, Random Forest, Support Vector Machine (SVM), and KKNN (K-nearest neighbors). Upon assessment, it became evident that Random Forest and C5 outperformed the other classifiers significantly. Notably, the Original and RFE datasets showcased superior performance across the classifiers.

Ultimately, our decision to select the final model hinged on accuracy and kappa The highlighted dataset, exhibiting the highest among the variations tested, was chosen as the most suitable for our purposes.

IPHONE								
DataFrame	**C5**		**Random Forest**		**SVM**		**kknn**	
	Acc	Kappa	Acc	Kappa	Acc	Kappa	Acc	Kappa
iphoneDF	0.7727	0.5589	0.7759	0.5658	0.6931	0.3695	0.3281	0.1613
iphoneCOR	0.7007	0.4035	0.7051	0.4111	0.6828	0.3451	0.2042	0.0571
iphoneNZV	0.7569	0.5233	0.7551	0.5223	0.7147	0.4258	0.3166	0.1443
iphoneRFE	0.7719	0.5573	0.7756	0.5663	0.7170	0.4308	0.3382	0.1685

Lessons Learned Report - Galaxy

In a similar approach to the iPhone dataset analysis, we replicated the setup for the Galaxy dataset. However, in this instance, we streamlined the process by exclusively utilizing the default dataset, leveraging our previous success with its performance. This decision was influenced by the extensive training duration, prompting us to take the calculated risk of relying solely on this dataset, assuming its potential to yield favorable outcomes.

Fortunately, our gamble paid off, as the default dataset showcased commendable performance, aligning closely with our expectations. Despite registering approximately 1% lower accuracy compared to our best-performing iPhone dataset, it notably delivered satisfactory results.

Hypothetically, if the default dataset failed to meet our satisfaction, contingency plans included exploring alternative datasets to ensure a comprehensive assessment.

Galaxy								
DataFrame	**C5**		**Random Forest**		**SVM**		**kknn**	
	Acc	Kappa	Acc	Kappa	Acc	Kappa	Acc	Kappa
galaxyDF	0.7642	0.5273	0.7608	0.5259	0.7006	0.3641	0.7133	0.4696
galaxyCOR	**NA**							
galaxyNZV								
galaxyRFE								

Lessons Learned Report - Finale

This project served as an invaluable learning experience for us. Leveraging AWS for data acquisition and employing parallel processing to expedite model training were notable highlights, contributing significantly to our skill development.

Overall, the project progression was relatively smooth with few obstacles encountered, and those we faced were promptly addressed. However, the primary challenges revolved around the substantial time required for model building and the implementation of parallel processing. Despite our efforts to streamline processes using parallel computing, the time investment remained considerable. Additionally, we encountered instances where the parallel processing encountered issues, either underutilizing assigned cores or disrupting the script's functionality, necessitating periodic restarts of R for resolution.