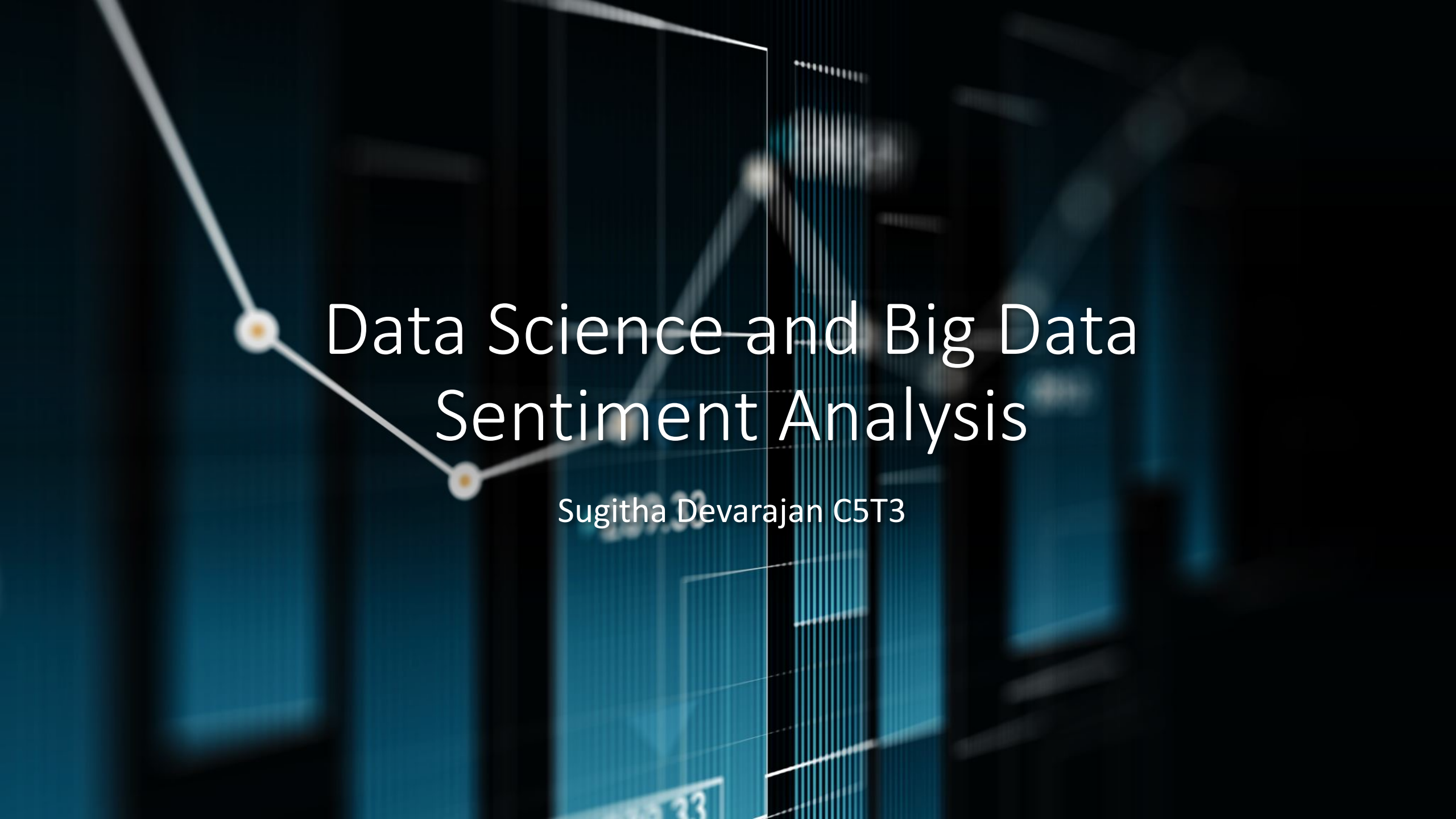


Data Science and Big Data Sentiment Analysis

Sugitha Devarajan C5T3

Background of the Project

- A challenging project to analyze sentiment on the web toward a number of smart phones for Helio, a smart phone and tablet app developer.
- Helio is working with a government health agency to create a suite of smart phone medical apps for use by aid workers in developing countries
- To help Helio narrow the device list down to one device, we examined the prevalence of positive and negative attitudes toward these devices on the web.
- Using AWS EMR we collected the data – The large matrix.
- iPhone and Galaxy small matrix was provided by manually updating the sentiment which we will use to train our model.

Data

- Attributes that collect information about the **relevancy of the webpage** toward each device (columns A-E)
- Attributes that collect information about the sentiment toward the **operating system** used on the phone. (columns F-G)
- Attributes that collect information about the sentiment toward a **phone's camera** (columns H-V).
- Attributes that collect information about the sentiment toward a **phone's display** (columns W-AK)
- Attributes that collect information about the sentiment toward a **phone's performance** (columns AL-BF).

Dependent variable

labeled sentiment

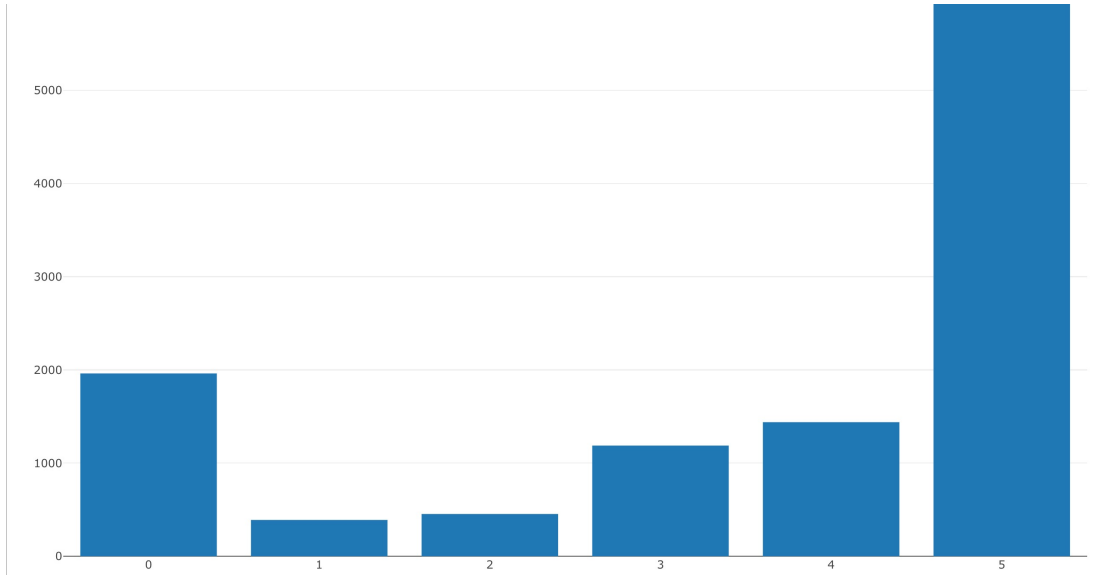
- #0: Sentiment Unclear
- #1: very negative
- #2: somewhat negative
- #3: neutral
- #4: somewhat positive
- #5: very positive

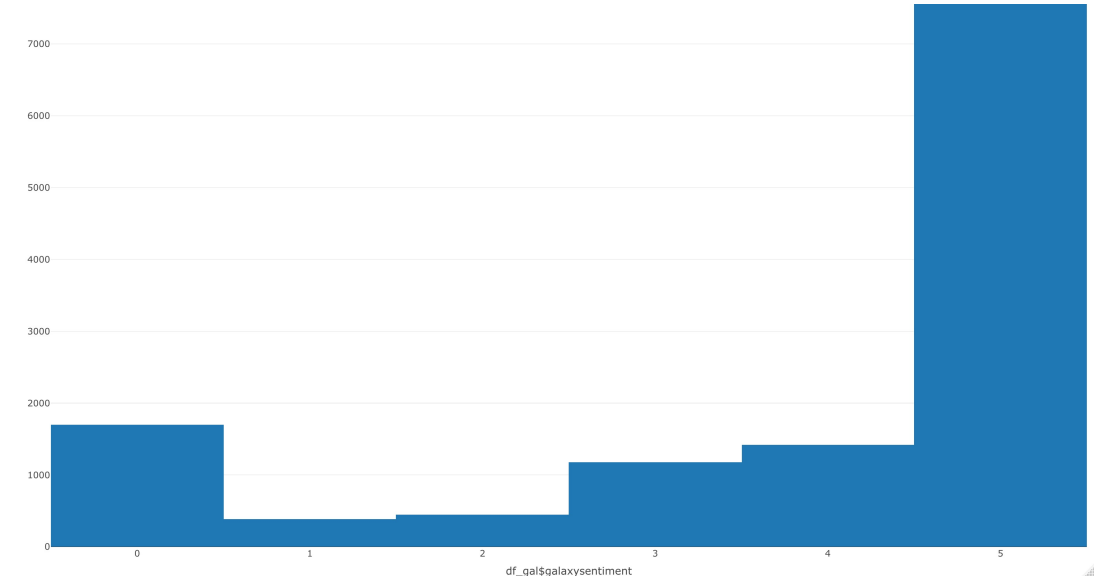
Approach

- Set up parallel processing
- Explore the Small Matrices to understand the attributes
- Preprocessing & Feature Selection
- Model Development and Evaluation
- Feature Engineering
- Apply Model to Large Matrix and get Predictions
- Analyze results, write up findings report

Observations

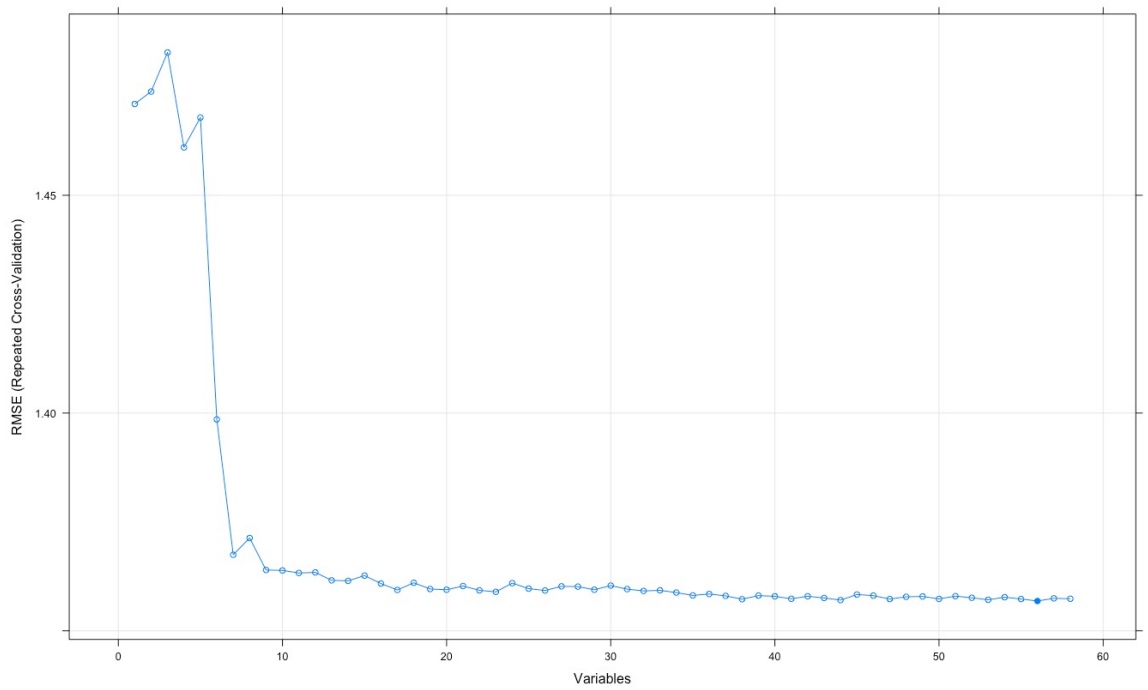
- From the small matrix we can tell that many liked iPhone and galaxy almost the same.
- Most of them are either unclear, neutral, or positive. Negative rating is very less.
- I see duplicates in both iPhone and galaxy small matrix.





Feature Selection

- Examine **Correlation** most of the feature have less correlation with the dependent variable. A new data set was created with most correlation.
- Removing the **near zero variance** will be helpful so a new data set was created removing near zero variance.
- **RFE** is a form of automated feature selection. The top 5 variables (out of 56): iphone, googleandroid, iphonedisunc, samsunggalaxy, iphonedispos



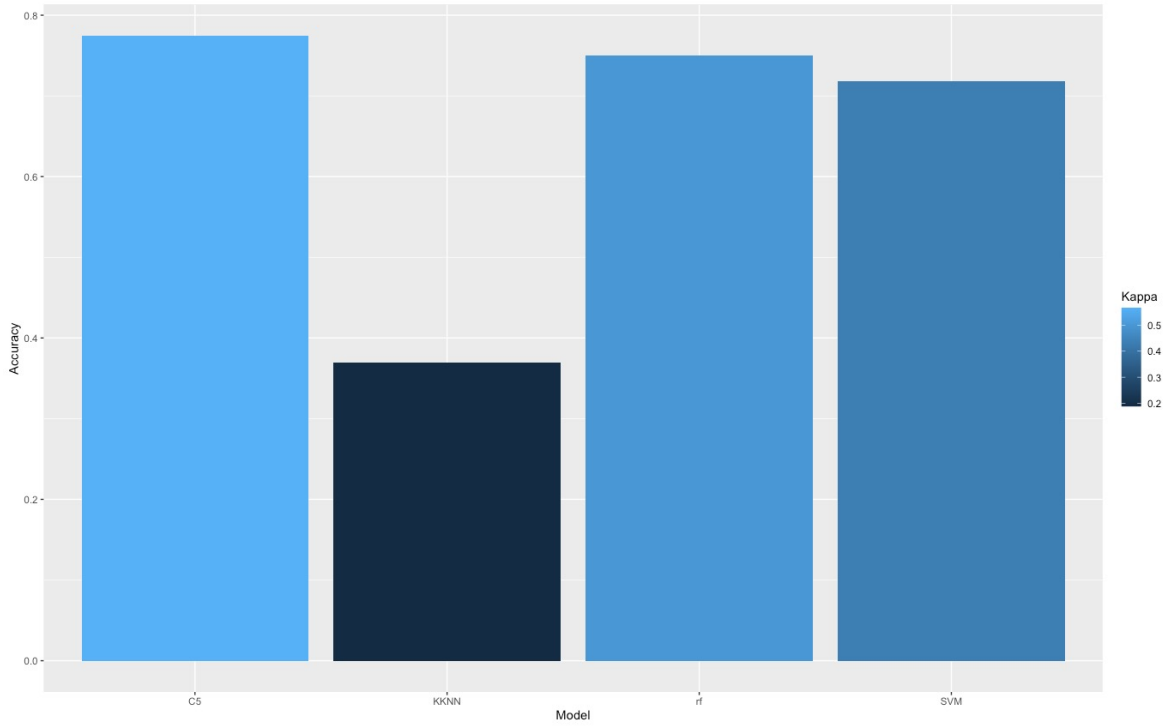
Preprocessing the data

- One, we recoded the sentiment to the following
  - 1: negative
  - 2: somewhat negative
  - 3: somewhat positive
  - 4: positive
- Second, we converted the dependent variable to factor.

Models

I used RF,C5,SVM, and KKNN models. I noticed RF takes more computational time. KKNN was not helpful in predicting because of the lowest accuracy. SVM, since this is classification problem, I tried the linear kernel, but the kappa was very low. On the other hand, C5 model had faster computation and slightly high kappa and accuracy.

- C5 is the choice made



Training the different feature selected data set and recoding the dependent

Since C5 model is being selected, now we check the postresample() for all the feature selected data.

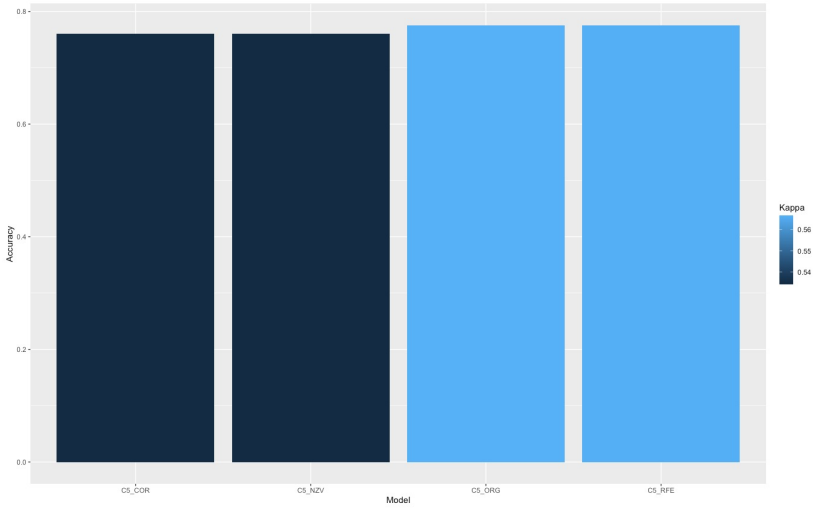
Then we do recode, by reducing the classification of the dependent variable we then see the improvement in accuracy.

1: negative

2: somewhat negative

3: somewhat positive

4: positive



	Model	Accuracy	Kappa
1	C5_RC	0.8544987	0.6438301
2	C5_RC_RFE	0.8542416	0.6430659

Galaxy small matrix – training using C5

- Taking the same approach, we convert the dependent variable to factor, re-code the sentiment column, and pick the C5 model to get the best accuracy and kappa.

Accuracy	Kappa
0.8411674	0.5874237

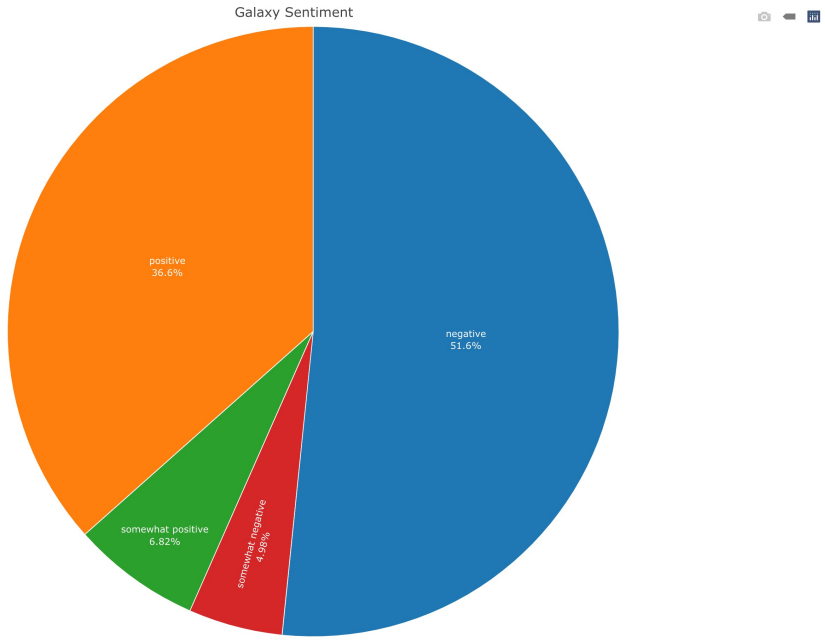
Predicting the Large Matrix using C5 model

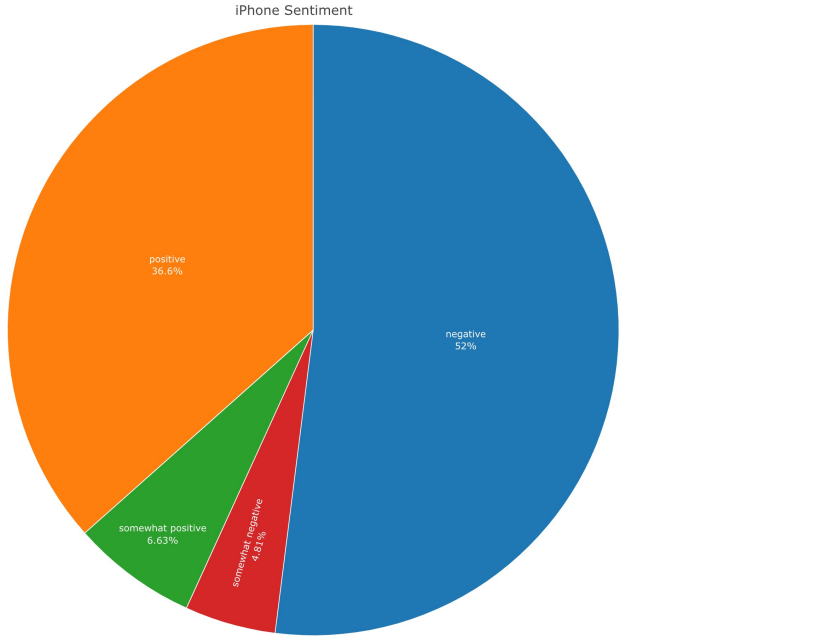
summary(preds_final_galaxys entiment)

1	2	3	4
10331	996	1365	7320

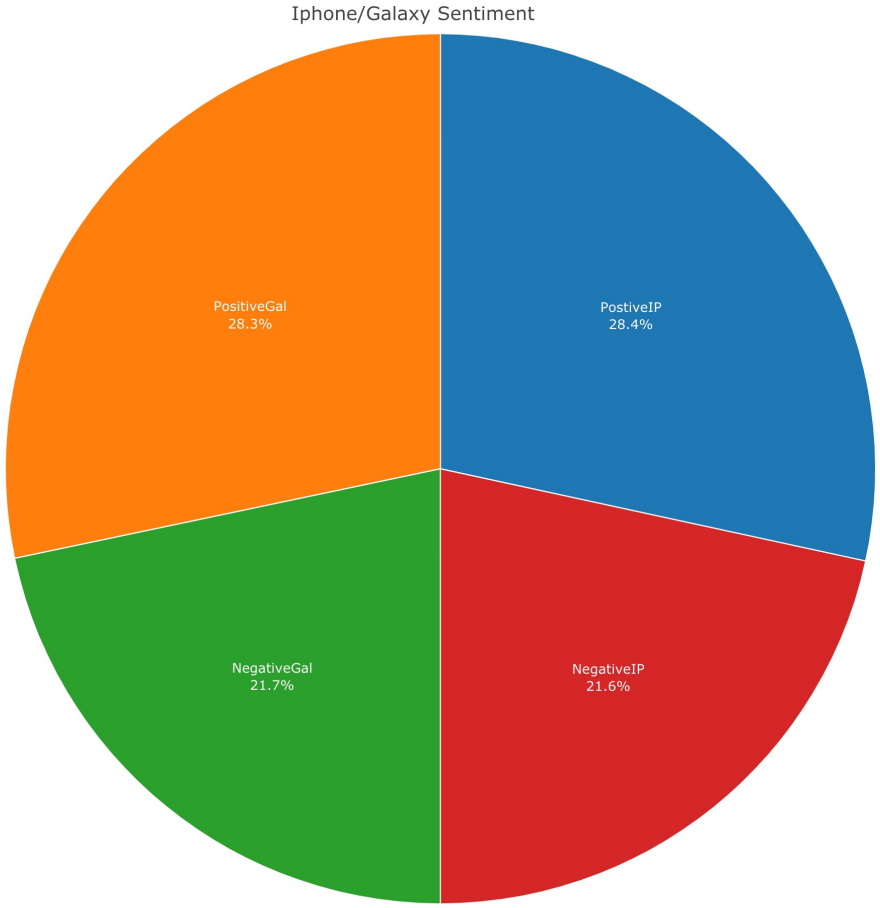
summary(preds_final_iphone sentiment)

1	2	3	4
10405	963	1326	7318





iPhone Vs Galaxy prediction



Recommendations

- Since the sentiment prediction of both iPhone and Galaxy are similar now it comes to the point of evaluating them based on cost and which gives the best value.
- According to fox business from 2020, iPhone in recent times started making affordable models to fit the needs of developing countries. https://www.foxbusiness.com/technology/most-popular-smartphone-world
- But Android phones are still largely used in the world according to https://www.which.co.uk/reviews/mobile-phones/article/apple-iphone-vs-samsung-galaxy-mobile-phones-aZL5V5m4UGbw
- Conclusion, for now it is safe to say the choice would be Galaxy Android for the purpose to be used in developing countries.