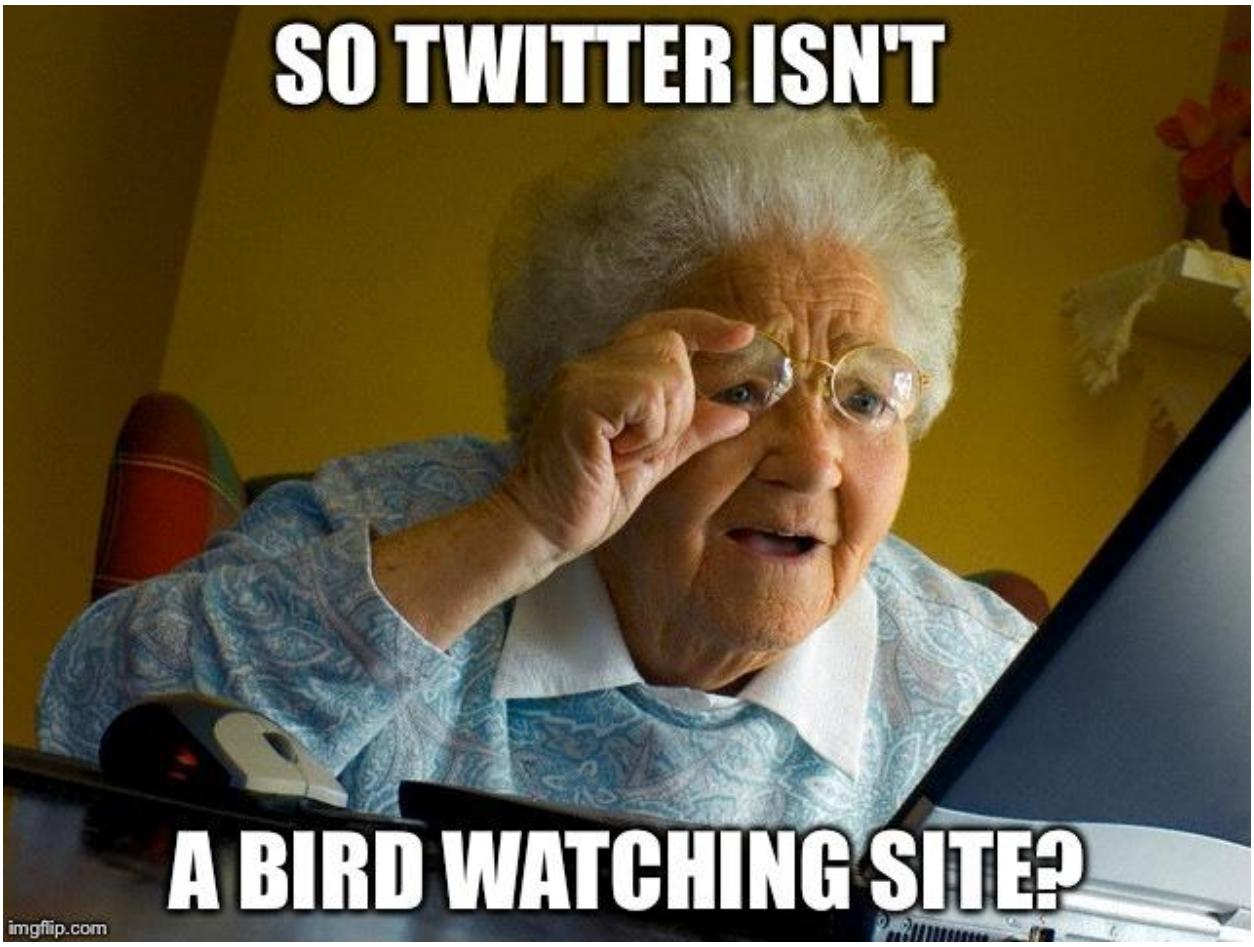
Springboard Data Science Career Intensive Capstone Project

#DSforTweet – Twitter Analytics



By: Ajay Sampath

Mentored By: David Yakobovitch

Sep 23, 2018

1. Introduction

Twitter is one of the most popular social media platforms in the world. The simple, but effective sharing channel has insights from 83% of the world's leaders, 330 million monthly users and 3 billion account holders.

It is no surprise so many companies want to tap into the potential of Twitter for their own social strategies. The fast-paced nature of this platform means that it's a great way for brands to start building a strong online presence. Tracking twitter metrics could help make insightful decisions about future marketing campaigns.

This project analyses tweets from the United States on one day. The primary purpose of the project is to understand how to scrape twitter, collect data and synthesis meaningful insights from the data. There are several future use cases for more specific tweet analytics and prior knowledge will help deliver solutions efficiently for existing and potential clients.

2. Data

Data was scraped from twitter using the TwitteR package. This is a R package that scrapes data from the twitter website. A snapshot for the code is shown below in *Figure 1*. A twitter account was created, and a key was obtained for obtaining the data.

```
twitter_tokens <- create_token(
  app = "Python Test - Ajay",
  consumer_key = "##################",
  consumer_secret = "############")

rt <- search_tweets("lang:en", geocode = lookup_coords("usa"),n=18000)
page <- max_id(rt)


for (i in 1:100){
  rt1 <- search_tweets("lang:en", geocode = lookup_coords("usa"),n=18000,max_id = page)
  page <- max_id(rt1)
  i = i+1
  print (paste('Finished collecting ',i * 18000,' tweets'))
  rt <- rbind(rt,rt1)
  saveRDS(rt1,paste('Tweets',i,'.rds',sep=''))
  Sys.sleep(15 * 60)
}
```

Figure 1.Snapshot from TwitteR package

2.1. Data Acquisition

The biggest challenge in scraping the website is the rate limit – Twitter allows you to collect only 18,000 every 15 minutes. The R script mentioned before was run for 24 hours to collect a total of 1.7 million tweets. The tweets were limited for the English language and within the United States. The script was paused for 15 minutes as soon as the number of tweets reached 18,000. Luckily the system did not crash while collecting the data.

2.2. Data Cleaning

The twitter data downloaded has two types of data – metrics from an original tweet and metrics for retweets. The number of retweets (1.6 million) were significantly higher than the original tweets (100,000). This was no surprise since most people tend to retweet than having an original post unless you are a celebrity. And since celebrities on an average have a significant number of followers, meaningful insights can be obtained from the data. Hence only the columns involving the metrics for the retweets were used for the analysis.

Each tweet/retweet has 10 variables associated with it such as text, source, favorite count, retweet count, followers count, friends count, statuses count, location, user description and whether the account was verified or not.

The cleaning process involved primarily inspecting and filling the missing values in the required columns. The data set was fairly clean as expected. The null values for the favorite, retweet and statuses count were assumed zero since it makes logical sense. The user description and location columns were u, street and continent were dropped for analysis. The 'Text' column is the most important and a considerable amount of time was spent cleaning the data during the machine learning model phase and will be discussed later.

3. Exploratory Data Analysis

This section takes an in-depth look at some of the variables in the dataset.

3.1 Tweet Source

First, we will analyze the devices used for posting the tweet (mobile device, web etc.). This is interesting data from a network provider perspective such as an AT&T or T-Mobile. If more users are using mobile devices, then network providers can decide on the kind of services they need to provide (e.g. 4g, 4g LTE etc.) to increase customer

base. Combining this data with the usage data such as CDN will enhance the insights and will add further business value for the network companies. However, the data is proprietary. This will be a good future work recommendation for a proof of concept analysis.

For the purposes of this analysis we will consider only 3 sources - Twitter for Android, Twitter for iPhone, Twitter for Web Client.

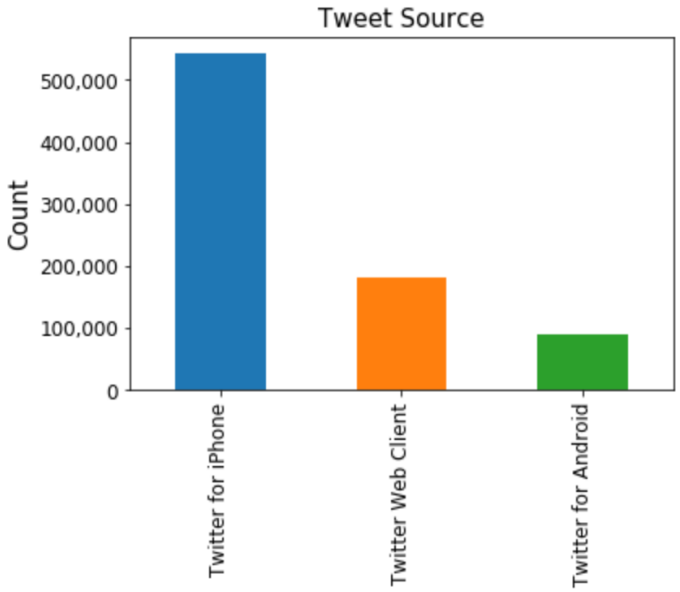


Figure 2. Tweet Source

Based on the analysis, looks like more twitter users have an iPhone. Good for Apple!

3.2 Tweet Metrics

The two main tweet metrics are likes and retweets. The distribution for the variables is shown below in Figure 3. The counts are extremely skewed. There are no surprises here since not every tweet explodes in likes and tweets. The mean for the favorites in the data is 19,437 and 7,626 for the retweets. These are still high. Maybe there are a lot of celebrity tweets in the dataset. There is a tweet that has 4 million likes and 3.6 million retweets. Any guesses who? Yes, the President of United States.

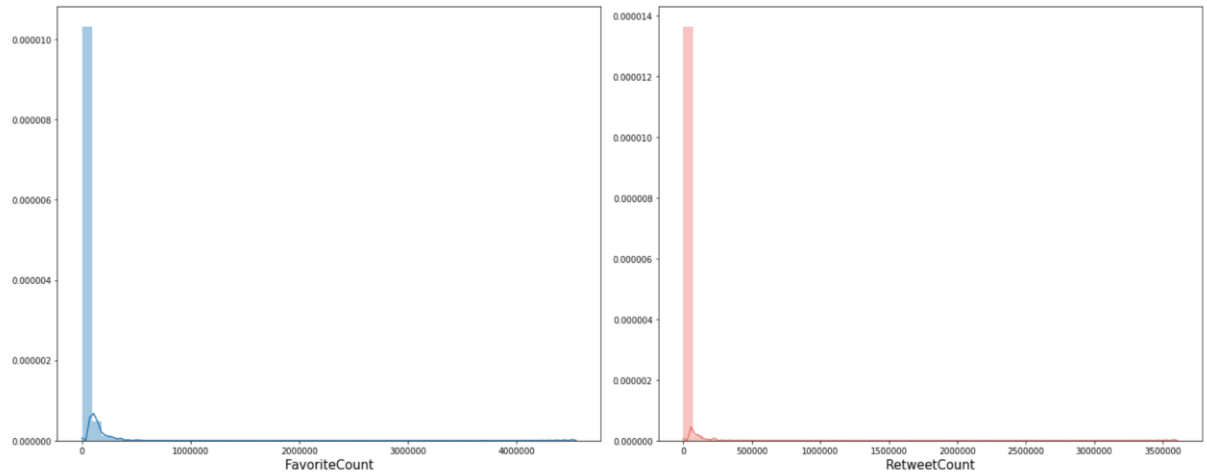


Figure 3. Tweet and Favorite Counts

The correlation plots for the Favorites and Retweets are shown below in Figure 4. There is an overall positive correlation between the favorite and retweet count. This makes sense intuitively. It is expected that more the favorites, the higher the likely hood of retweets and vice-versa.

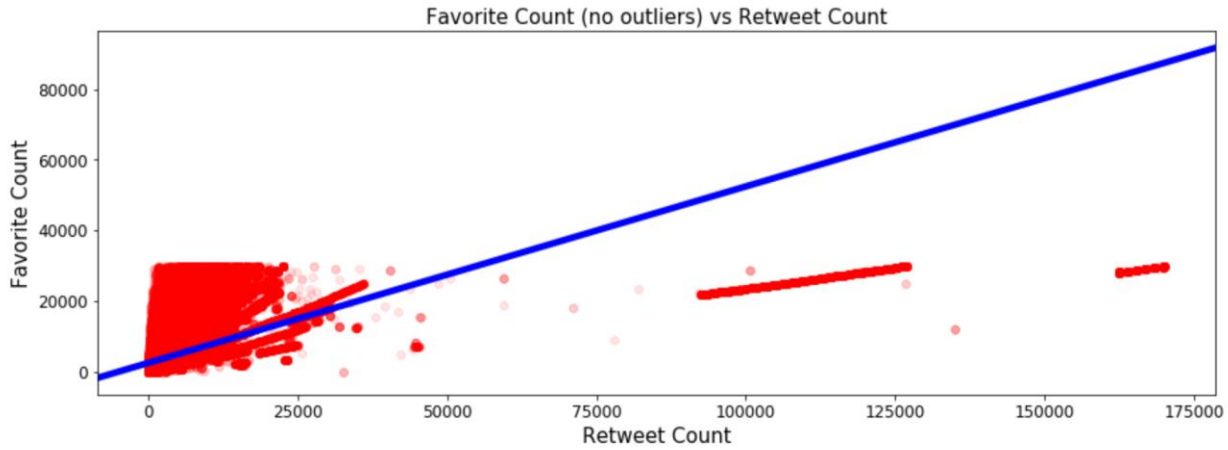


Figure 4. Favorite vs Retweet Count

3.3 User Metrics

The three main user metrics are followers count, friends count and statuses count. The summary statistics for the variables are shown below in Figure 5. The mean value for the followers count is approximately 1.8 million with a standard deviation of 8 million. It looks like the twitter scraper used for the analysis got data from a wide variety of accounts and is a good sample set for analysis. One person has close to 101 million followers!!! And another user has posted a total of 9 million tweets!!

	FollowersCount	FriendsCount	StatusesCount
count	996,647	996,647	996,647
mean	1,846,998	14,483	44,191
std	8,712,657	65,758	86,435
min	0	0	0
25%	2,135	324	4,629
50%	18,780	871	17,022
75%	167,401	3,037	45,217
max	101,802,394	2,122,181	9,108,873

Figure 5. User Metrics – Summary Statistics

We will analyze the relationship between followers count and statuses count. The plot is shown below in Figure 6.

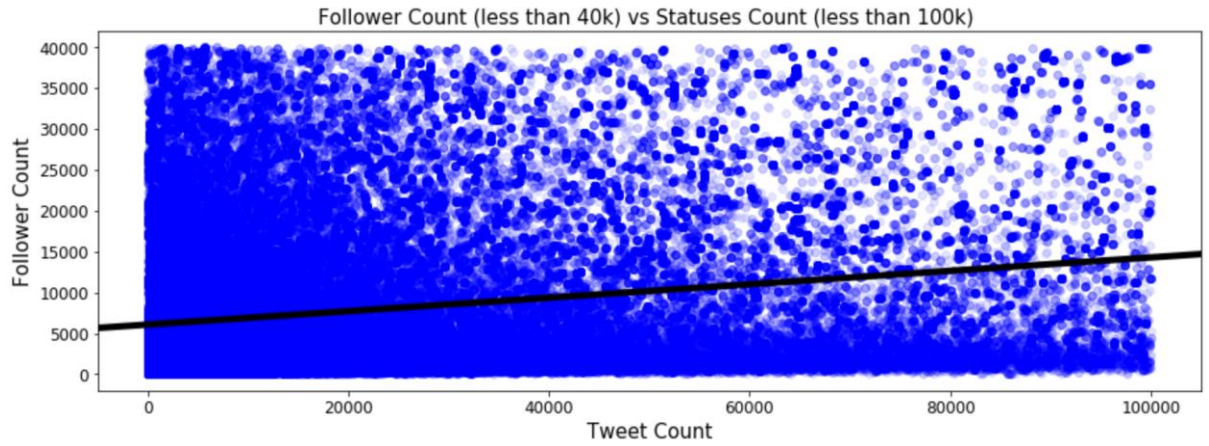


Figure 6. Favorites vs Statuses

The above plot shows only a slight positive correlation between the followers and tweet count. A tweet count of 100k is high. So, it may not matter if you are constantly tweeting to gain followers. As the age old saying goes, 'Quality matters over Quantity'.

3.4 User vs Tweet Metrics

In this section we will analyze the relationship between the user and tweet metrics. The correlation plot is shown below in Figure 7. It is interesting to see that the number of likes and retweets have no correlation with the followers count. I would have expected to see more retweets and likes if you had more followers. I guess, as stated earlier, the tweet must really matter!! We will get into the text analysis later.

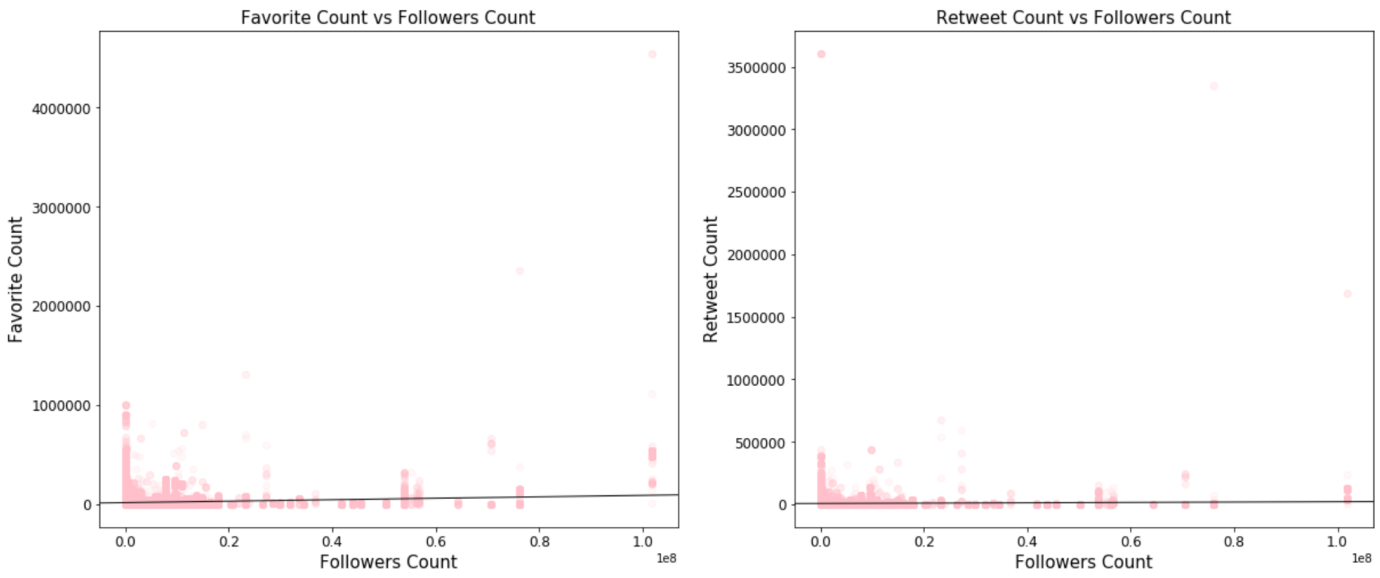


Figure 7. User vs Tweet Metrics

3.5 Text Analysis

Since we made a statement earlier that a quality of text matters, we will analyze the tweet column in detail in this section. Typical cleaning methods such as removing punctuations, converting to lowercase, remove hyperlinks, spaces and numbers were applied to use the text column for analysis. Two separate columns were created to extract the hashtags and mentions using regular expressions.

A word cloud was created to check the most used words in the tweets and shown in Figure 8.

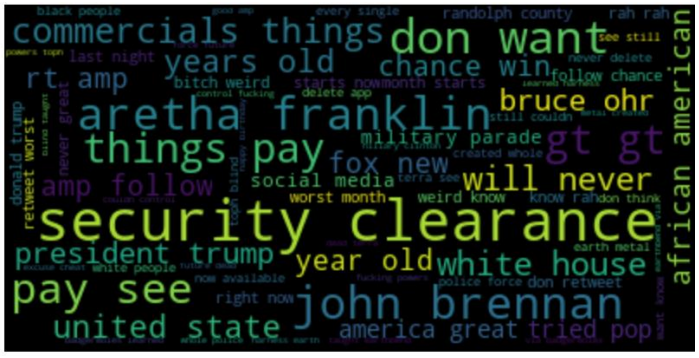


Figure 8. Tweet WordCloud

From the word cloud we can see a few words that stand out such as 'aretha franklin', 'john brennan', 'security', clearance' and 'president trump'. This makes sense because of the two important events that occurred on the day when the tweets were collected -

death of Aretha Franklin and president trump revoking the security clearance for the ex-CIA director John Brennan.

A bag of words model was created using a count vectorizer and the 10 most common hashtags are plotted in Figure 9.

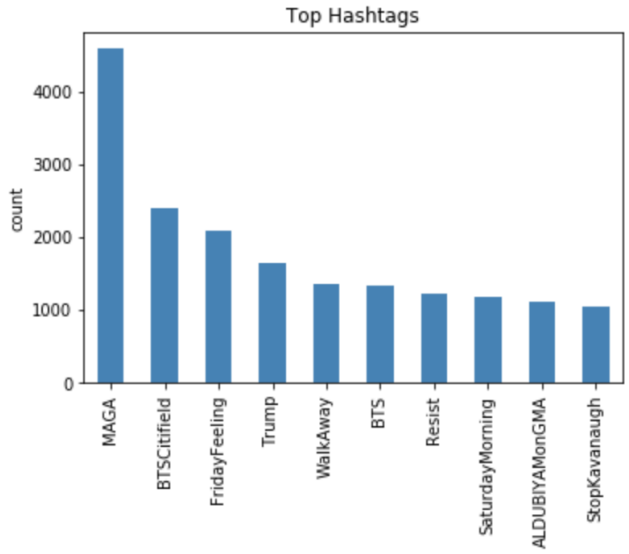


Figure 9. Bag of Words – Top Hashtags

MAGA had a substantial number of retweets. No surprises there! I guess MAGA will always be in the top 10 any day until we have a change in the President. The interesting one is 'ALDUBIYAMonGMA'. Initially I thought this was junk. But this was an actual trending topic that day. It is a movie of some kind. Surprising 'Aretha Franklin' is not showing up in the top 10. It was an unfortunate day of her passing away.

4. Prediction Modeling

Machine learning techniques were applied to predict the favorites count and retweet count (target variables analyzed independently). The first part of the analysis using all the columns except the text column - 'FavoriteCount', 'StatusesCount', 'FollowersCount',' FriendsCount', 'IsVerified', 'TextLength', 'HashtagCount', 'MentionsCount'. The second part used only the text column to analyze predict the target variables. The favorite counts and retweet count were divided into 4 categories as below and analyzed as a multi-class label problem.

0 - 100: 'Low', 101 - 1000: 'Medium' , 1001 - 10000: 'Medium_High' , Greater than 10001: 'High

The counts for the categories are shown in Figure 10. The favorite category count has a somewhat equal distribution for the different categories. For the retweet categories, the lows are significantly higher than the other categories. This may skew the accuracy for the predictions.

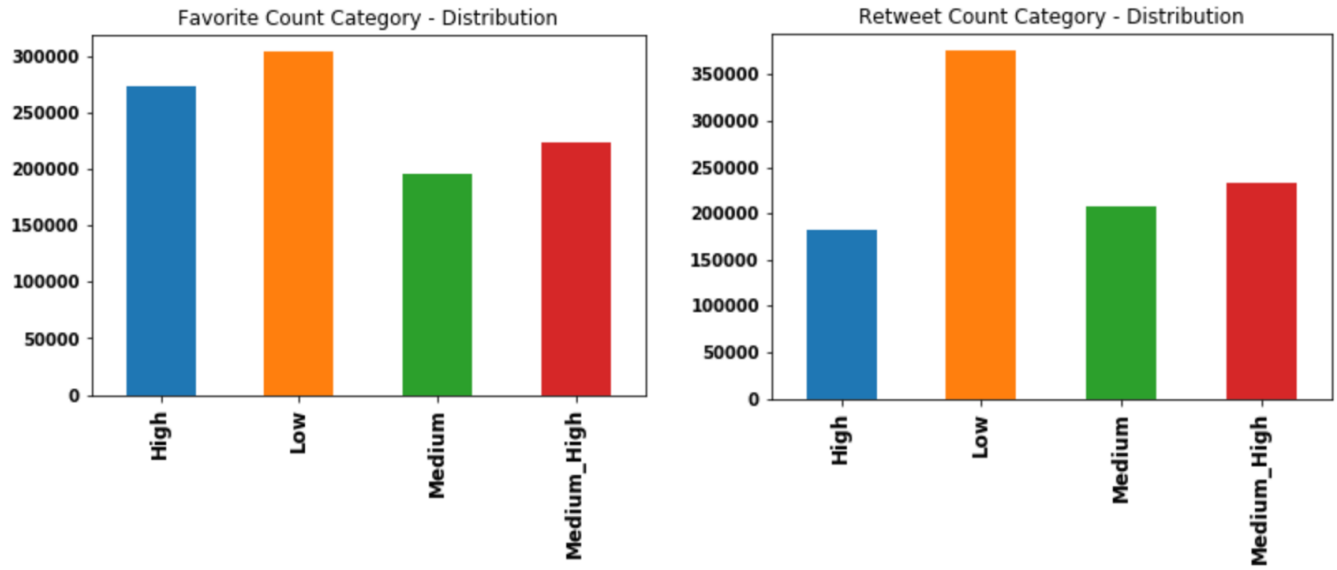


Figure 10. Favorite and Retweet Category Counts

The accuracy summary for the different algorithms and methods are summarized in the table below.

Algorithm	**Accuracy**
Logistic Regression	76.9 %
Logistic Regression with standard scaler	78.1%
Decision Tree and Standard Scaler	93.5 %

The Decision Tree classifier with the standard scaler was used to predict the retweet counts. The accuracy from the model was 95.3%.

The favorite counts categories were predicted using only the text column. A multinomial Naives-Bayes algorithm was applied for the prediction since it is the most common for text-based classification problems. Both count vectorizer and tfidf vectorizer were used for hyper tuning the model. The accuracy summary for the two methods are summarized below.

Algorithm	Accuracy
Multinomial NB – Count Vectorizer	77.6%
Multinomial NB – tfidf Vectorizer	80%

5. Conclusions

A total of 1.6 million tweets between Aug 16 and Aug 17 were analyzed for this study. One of the key findings was that the followers or favorite count does not really dictate the number of retweets and likes. It really depends on the content and other factors such as audience sentiment after reading the tweets. The decision tree classifier gave an accuracy of 93.5% for the favorite count prediction and 95.3% for the retweet count. The multinomial NB model with a tf-idf vectorizer showed an accuracy of 80% for the favorite count prediction using just the text data.

Further analysis is required on the data such as sentiment analysis and Named Entity Recognition. It is also planned to incorporate the exploratory data analysis into a dashboard using Dash by plotly. This will help providing an end-to-end solution for clients from insights to machine learning.