Bungie Twitter Analysis
Capstone Project 2 Milestone Report
Sebastian Alvis

Problem Statement

For a company, public opinion is very important. The reach (number of people who see) of their social media posts can often indicate how well a product or idea will be received. The changes in reach can often be correlated to announcements on social media, and can help drive business decisions related to advertising / marketing.

In this project, I am going to analyze the social media presence of Bungie and it's main game, Destiny 2. Part of the analysis will be using natural language processing to make a guess at the opinion of each tweet: whether it is positive or negative.

Dataset Description

The data was acquired from the Twitter API: https://developer.twitter.com/en/docs. I am using the standard, free version, which allows data acquisition from present day to 7 days prior. I acquired a weeks' worth of tweets on the topics "Bungie" and "Destiny 2" (topic data), and acquired the tweets made by the official Bungie (@Bungie) and Destiny 2 (@DestinyTheGame) Twitter accounts in the same time frame (timeline data).

This process required patience, as there was a significant problem with finding duplicate tweets. I could acquire 20k tweets, but only a few thousand of them would be unique. There is a request parameter that enforces a time constraint on the tweets returned in that request, but I was not using it correctly for some time. In the end, I managed to pass it the integer representing the earliest tweet in my data, and the next request would only find tweets as or less recent than that. This was a problem for the topic data, but not the timeline data.

The loop would run, gathering tweets until nothing was returned and I had exhausted the use of the standard search API. I dropped duplicate tweets and found I acquired 21.4k tweets about Destiny 2 and 24.7k about Bungie. The timelines for the official accounts in that week came back with 46 and 107 tweets.

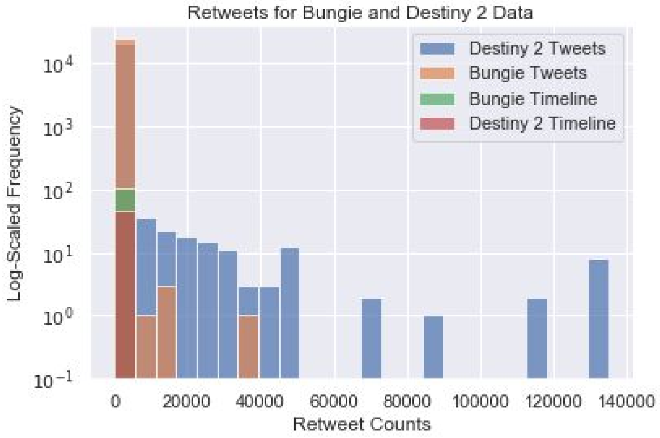
Data cleaning had two main components: missing data and column reduction. The dataframe reduction was about reducing the size of my data by removing columns. The topic data was originally a 320-column dataframe. I went through all the column names, found about 50 columns that were the most relevant, and kept those. These 50 columns were present in the topic and timeline data both, so now my four dataframes all have the same columns (the timeline data originally had about 170 columns). The reason for that many columns was a lot of parameters about the user and the location from which they tweeted, and all the columns for a tweet / user / location were repeated in the case of quoted tweets, retweets, and retweets of quoted tweets. I kept columns for some parameters of base tweets, quoted tweets, and retweets. The last category had single-digit results out of 40k+ tweets, so I deemed it irrelevant.

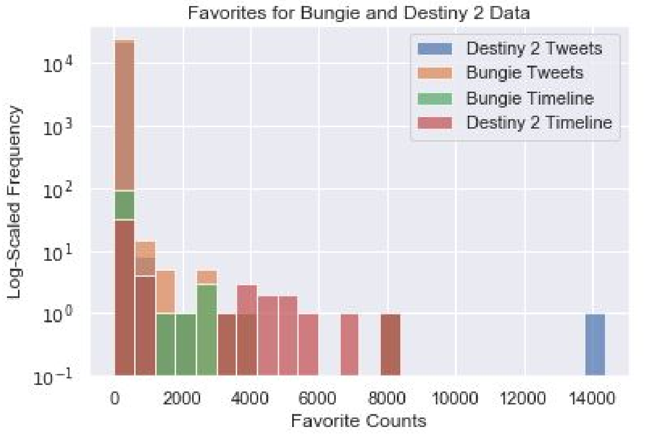
The missing data came in two forms. First was normal missing data: the user and text of the tweet were missing, so I could drop those incomplete rows. The second was misaligned data. recognized the problem because I had some indices that were clearly text instead of an integer, and when digging further, the rest of the columns had improper values or datatypes. I'm unsure if this was a problem in the API or in my saving of the data, but the indices make me think the latter.

After reducing the column number and removing missing data, I saved the dataframes as new csv files to keep reproducibility, and data wrangling / cleaning was complete.

Initial EDA

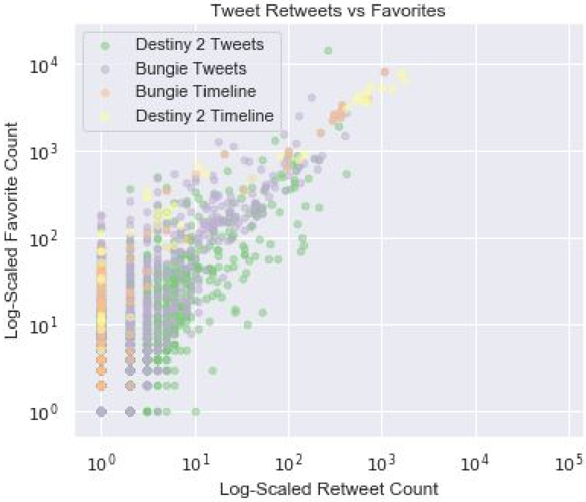
The first metrics that I can measure tweet engagement with are retweet and favorite counts. I have four dataframes, so I can plot these distributions for each one and overlay them all to get a sense of what the data looks like.



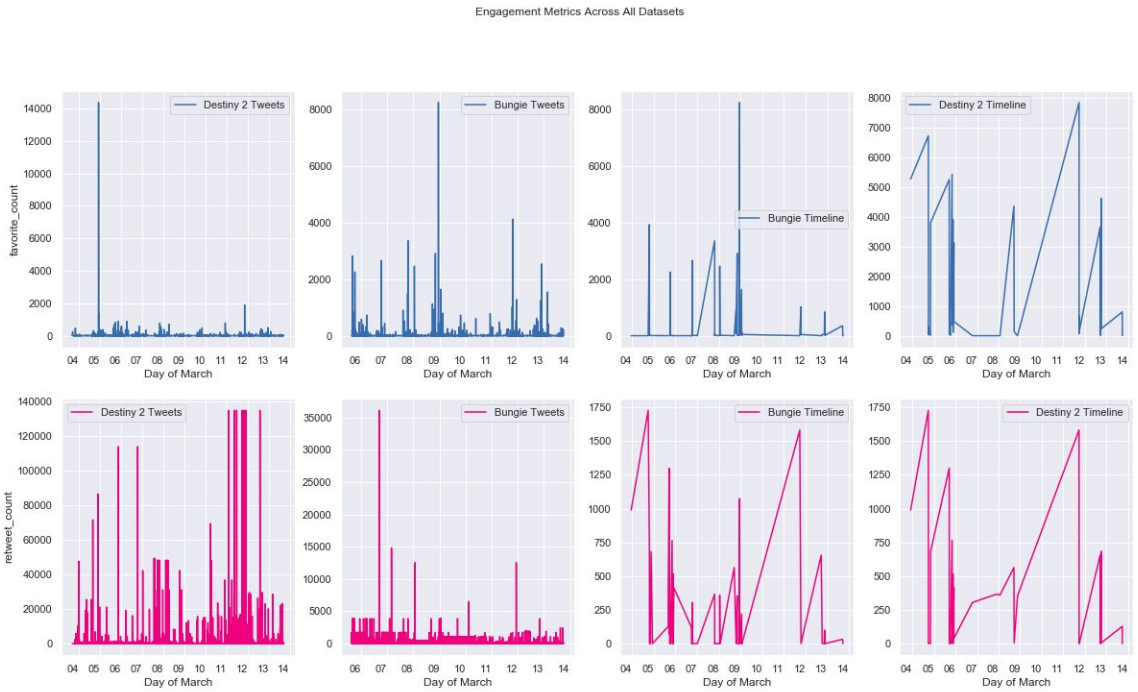


In these plots, the Destiny 2 Tweets and Bungie Tweets both have very similar values for the leftmost bin. If you look closely, you can see that the orange bar is slightly taller than the blue bar hiding behind it.

What is notable for these plots is that the overwhelming majority of the data sits in the first bin. Additionally, the official accounts don't get a significant number of retweets ever (compared to the rest of the data), but they do get a competitive number of favorites.

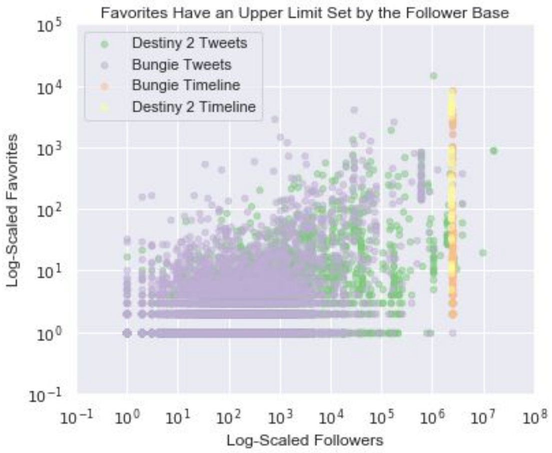


I wondered if retweet and favorite counts were related, and found that it only appears that way if you log-scale both of them. I think there are a couple high outliers on each axis that make plotting these quantities difficult.

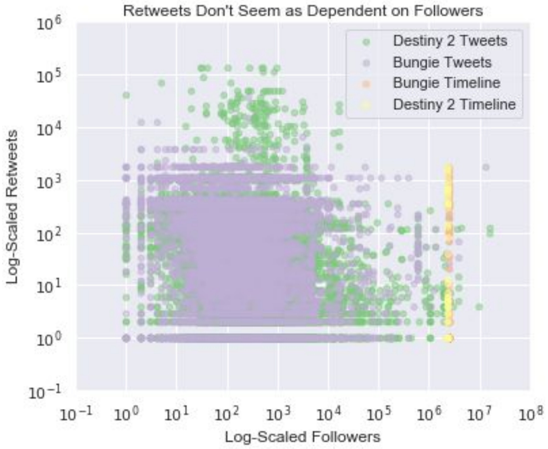


Make title and axis labels bigger

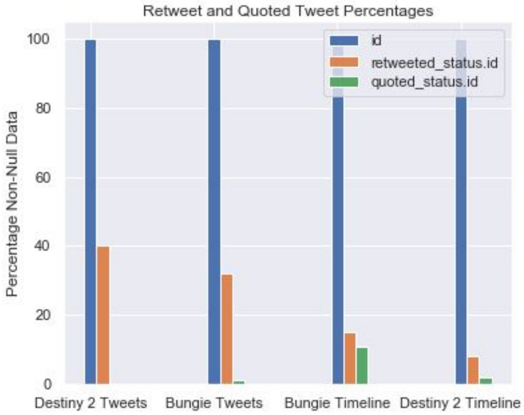
The left two columns are the main bulk of tweets, and there isn't much time correlation between the retweet and favorite counts. The official accounts, however, have a decent time correlation. Part of this is probably because the official account have large follower bases that regularly interact with their tweets and that there aren't as many tweets in general. It also helps that since those columns are one Twitter account each, a lot of variables about the user posting are held constant.



Retweets and favorites correlate some, but I figured both of them would correlate with the size of the follower base of the posting user. For favorites, pictured above, it seems that the follower base sets the upper limit of the favorite count, but an uninteresting tweet could always have fewer favorites that this limit.

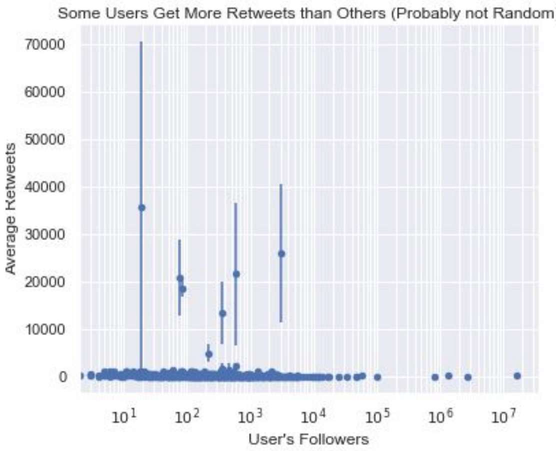


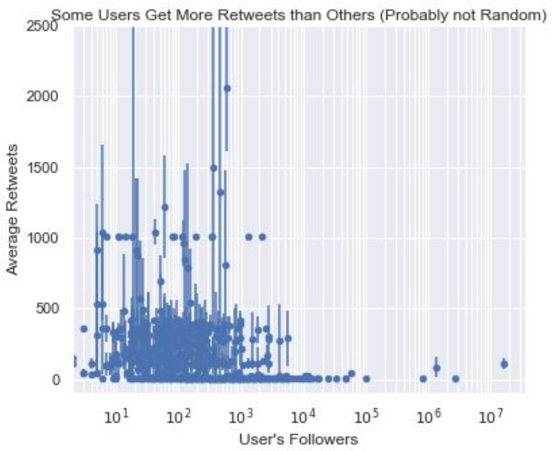
This trend does not hold for retweets. There doesn't appear to be a correlation between follower base and retweet counts, although this may be because a retweeting a retweet can get more retweets on the original tweet (if that makes sense).



Speaking of retweets, I realized I didn't have a sense of how many retweets were in my data. Here, the blue bars are tweets, and 100% of my data is tweets. Retweets constitute roughly 30% of the overall data, differing depending on what dataset you look at. Quoted tweets are much rarer.

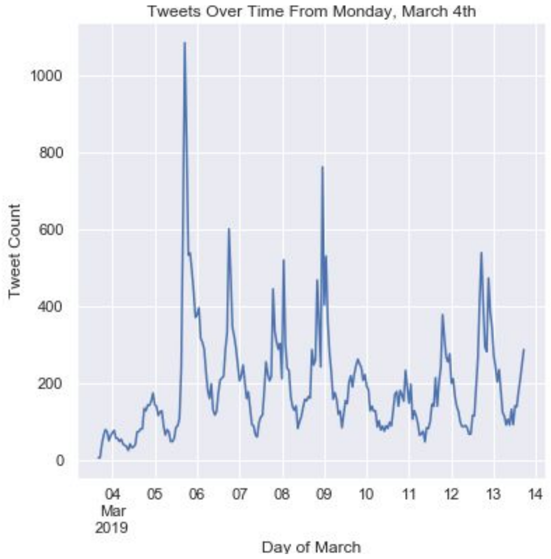
Looking at things from a single tweet perspective is natural, since each row is a tweet, but it would be good to see if certain users get more or less engagement than others, and if there are factors that help.





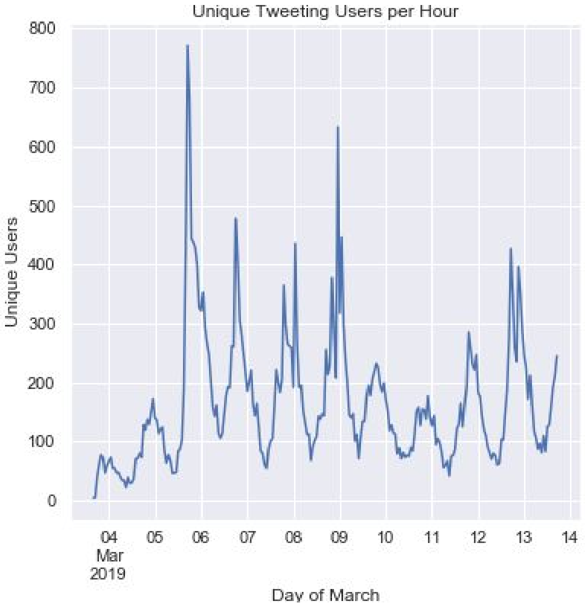
The left plot has a larger scale on the y-axis. The error bars are the 95% Confidence Intervals on the average number of retweets for the given user. Since there are many sets of error bars that do not overlap, I conclude that there is less than a 5% chance for the difference in certain users' average retweets to be purely random. Similar plots can be found for favorite counts.

Now, since this is time-ordered data, I am interested in when tweets generally get posted.



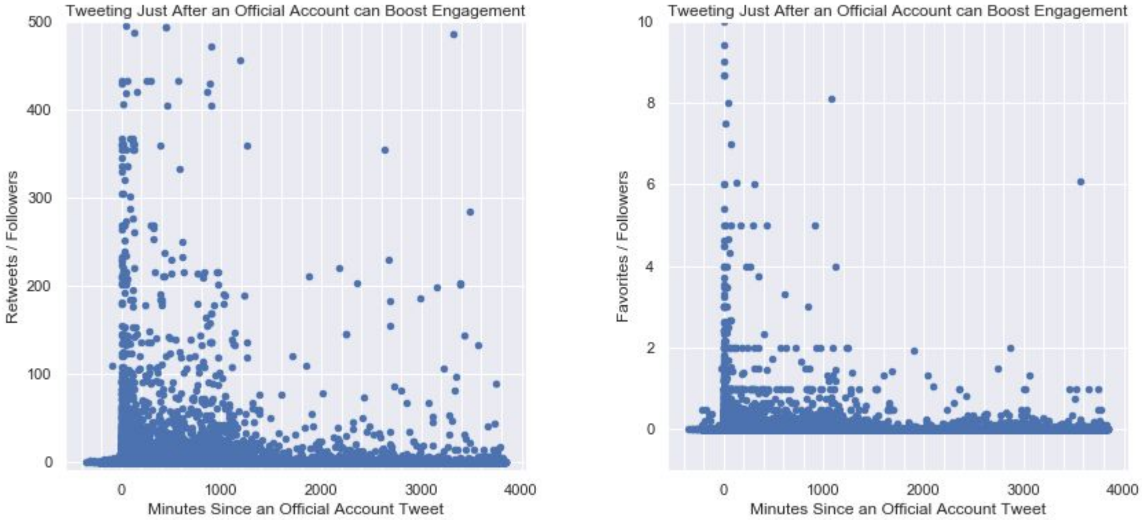
The first tweets I have are from Monday, March 3rd, 2019. The small first couple peaks are actually artificial. The Destiny 2 tweets I acquired go back to March 3th, but the Bungie tweets only go back to March 5th. This is just what I was able to get from the API.

The grid lines are midnight, and we can see that most tweets are posted in the evening hours. The lower peaks on March 10-11 are Saturday and Sunday evenings. The official accounts tend not to post tweets on the weekends, which may be part of the reason why there are fewer tweets going out then.



The number of unique users who post in a given hour strongly matches the pattern of tweets in a given hour. Most people probably post 1-2 times in an hour, so the distributions are very similar.

There is a thought that the official Bungie and Destiny 2 accounts can generate Twitter chatter by posting, but we can check this visually. I constructed a column to represent the time a tweet was posted minus the time of the most recent posting from one of the official accounts. I would expect that a smaller time difference will result in more engagement (retweets / favorites). The following plots support this theory.


Tweeting Just After an Official Account can Boost Engagement
Retweets / Followers
Minutes Since an Official Account Tweet
Tweeting Just After an Official Account can Boost Engagement
Favorites / Followers
Minutes Since an Official Account Tweet