The WeRateDogs Project- An analysis of data wrangled from Twitter

January 29, 2019

By Arham Ansari

According to an article by The Globe and Mail, Twitter today has almost 200 million users worldwide. Approximately 460,000 new Twitter accounts are opened daily. More than 140 million tweets are sent daily. That's one billion weekly tweets! For this project, the goal was to capitalize on Twitter's vast amounts of tweet data, utilizing the Twitter API to exploit the Twitter data of the user @dog_rates, aka WeRateDogs. WeRateDogs is a very popular Twitter account with over 4 million followers and has received international media coverage. WeRateDogs gained its popularity by rating people's dogs with a good-natured comment about the dog.

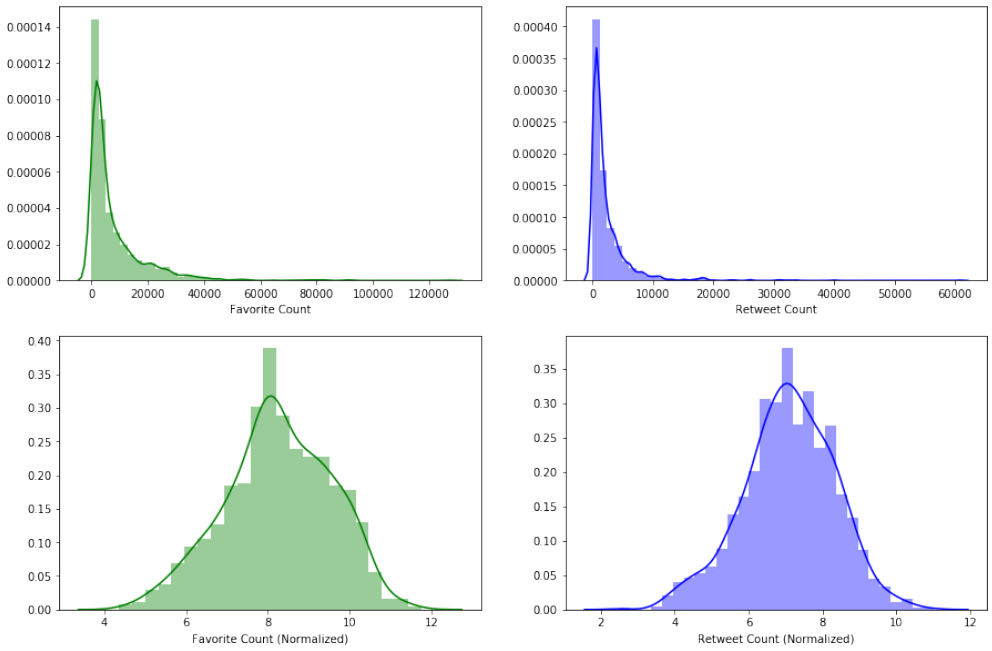
For this analysis I gathered data from three different sources. WeRateDogs gave Udacity exclusive access to their Twitter archive for this project in the form of a csv file. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017. Each tweet image was run through a convolutional neural network with the purpose of analyzing the images to correctly identify the dog breeds. The convolutional neural network predictions were programmatically downloaded using the Requests Python library as a tsv file. And finally, using the tweet IDs from the WeRateDogs archive I queried the Twitter API for each tweet's JSON data using the Python's Tweepy library I stored each tweet's entire set of JSON data, which I would later use to analyze the tweet's retweet and favorite counts.

Before diving into the statistical analysis, I began by answering some basic questions. What are the most common dog names in the dataset? What does the tweet say about the dog with the lowest rating (i.e. 0/10)? Using the Dog Breed Classifier, what do the dogs with the lowest rating look like and was the classifier able to accurately predict the dog's breed?

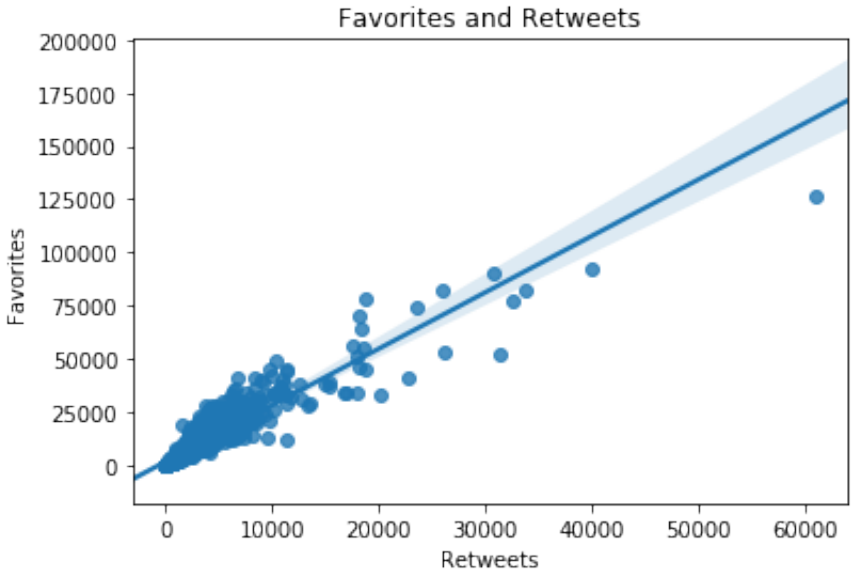
I discovered the most common dog names within the WeRateDogs dataset, excluding the NaN values, are Oliver, Winston, Tucker and Penny.

Now let's dive into a statistical analysis of the Dog Ratings! The mean numerator value is 12.84. The most interesting result is the rating_numerator maximum value of 1776. The rating_numerator outlier is a dog named Atticus.

The doggo with the highest favorite count also has the maximum retweet count. On further investigation I found out that his name is Stephan; he had a rating of 13/10. The tweet said, "This is Stephan. He just wants to help". The Dog Classifier did really well in predicting Stephan's breed. Stephan appears to be a Chihuahua/Corgi mix and the classifier pegged Stephan as a Chihuahua with a predication confidence equal to 0.51. Below is a picture of Stephan; the most popular do in the dataset.



Visualisation:



Favorite,Retweet- The original distributions for both favorites and retweets have long positive tails. Extremely popular tweets are extremely rare. The normalized graphs (the bottom two) again show similar distributions. They are roughly normal except for the spike in values for the normalized favorite count. That may be due to my jitter work and not an actual attribute of the data. Or there are actually a bunch of tweets with one favorite count.

Regession- There is a strong relationship between retweet and favorite counts. As a tweet gains Favorites, one can expect to see Retweets to increase and vice versa. It looks like it may be a nonlinear relationship.