The WeRateDogs Project: Analysis of data wrangled from Twitter

Goal: To capitalize on Twitter's vast amounts of tweet data, utilizing the Twitter API to exploit the Twitter data of the WeRateDogs.

WeRateDogs is a very popular Twitter account with over 4 million followers. 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. 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 (i.e. "like") counts.

Before diving into the statistical analysis, I began by exploring the datasets and looking for basic questions like 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.

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 dog 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.

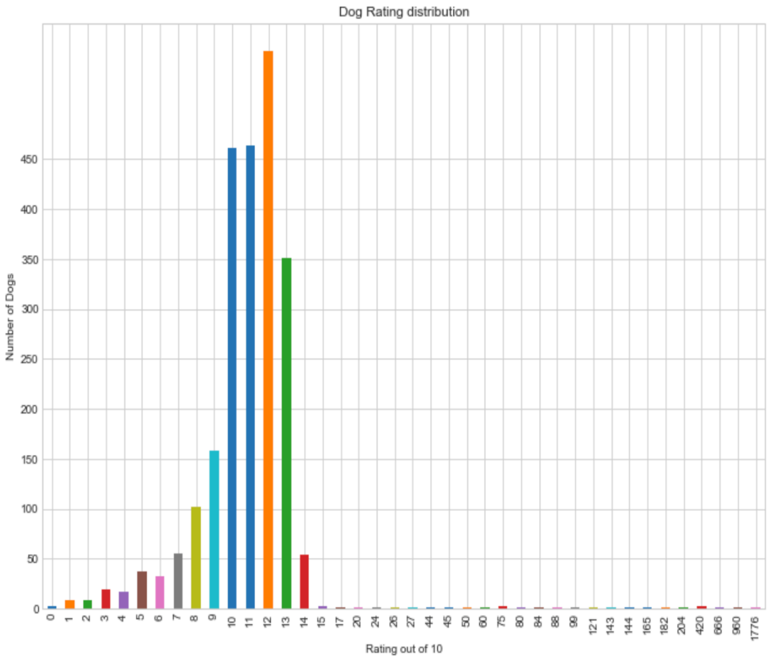
Now it's time to dive into the favorite and retweet count data. The statistical analysis denotes a large positive (right) skewed distribution in both categories indicated by the large standard deviations. The results also indicate people will favorite a tweet more often then they will retweet the original tweet as shown by the larger favorite count.

I also saw a strong correlation between the favorite and retweet data with a Pearson correlation coefficient, r, equal to 0.92. The strong correlation makes logical sense because the popular a tweet the higher the favorite and retweet count should be.

With this new insight into the tweet data I was curious, what time of day do people tweet about dogs? Also, what time of year is the most active time for people to tweet about their pets?

It seems the most popular times to tweet about dogs are 6am, 12am and 8pm, respectively. Notice that from the statistics report there are a greater number of people "liking" a tweet as compared to the number of retweets. Interestingly, tweet activity is cyclical in nature with relative maximums in January and May with an absolute maximum in December.

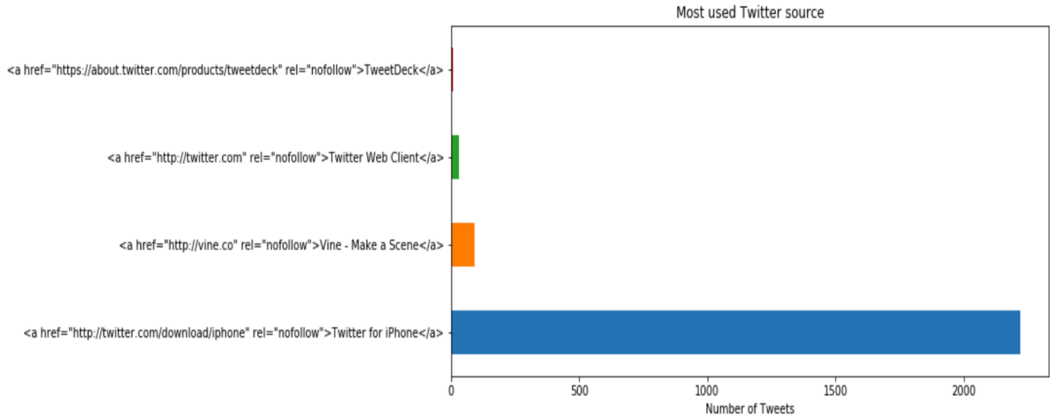
**Dog Rating Distribution:**



Most of the dogs are rated 12/10 (455 out of 1994 dogs).

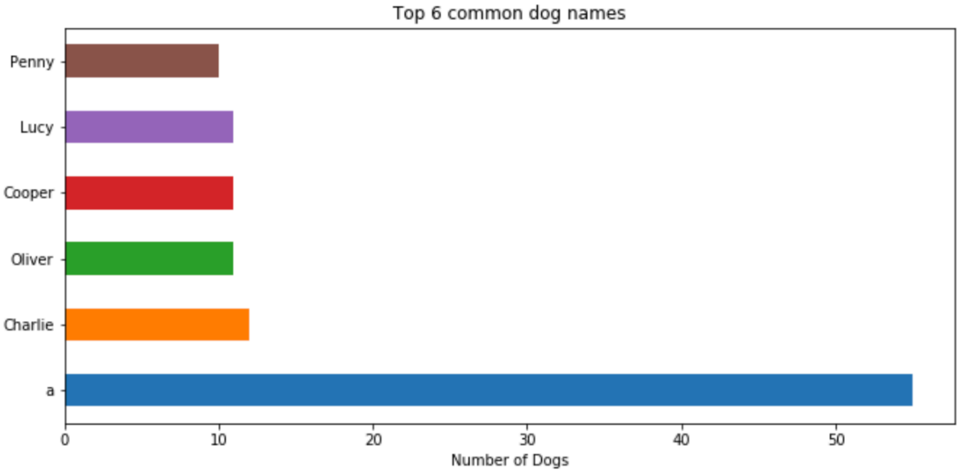
Excluding the 2 rating outliers (420 and 1776), the highest rating received by any dog is 14/10. However, only 2% (i.e. 37) dogs got this rating.

**Most Used twitter source:**



Out of the 1994 tweets, 1955 were posted from iPhone. It seems that only mobile device WeRateDogs uses to post tweets is an iPhone.

**Common Dog Names:**



It seems Charlie is the most common dog name. A close second will be Lucy, Cooper and Oliver.

The favorite and retweet data is strongly correlated and the most active time of year for people to tweet about dogs is in December.

A Chihuahua mix named Stephan has the highest favorite and retweet count.