WeRateDogs Project

Wrangling Data gathered from twitter

By: (T.J.Adel) Al-Dajani

Date: January the 1st 2022

Twitter is a social media platform with millions of users that share their pictures opinions and their achievements. Using data gathered from twitter specifically from a user called WeRateDogs, we are going to assess this data and clean it.

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "**they're good dogs Brent**." WeRateDogs has over 4 million followers and has received international media coverage.

For tis analysis data was gathered from three different places. WeRateDogs gave Udacity exclusive access to their twitter archive for this project in the form of a csv file. This file contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017. The second source of data is tweet images that were run through a convolutional neural network to analyze these images and identifies the breed of the dog, and then it was programmatically downloaded as a tsv file.

The third and final source of data is in a file called tweet-json.txt that can be acquired from the twitter API.

Before diving into the analysis, I Tried answering some simple questions like:

What is the highest rating of a dog?

What source is used the most for sharing these tweets?

What is the most common dog rating?

And what stage has the highest average rating?

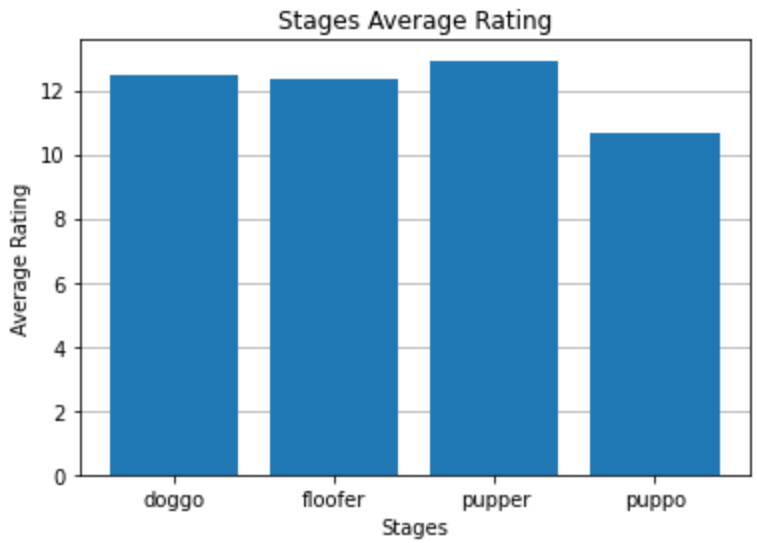
	tweet_id	rating_numerator	rating_denominator	retweet_count	favorite_count	img_num	p1_conf	p2_conf	p3_conf
count	1.994000e+03	1994.000000	1994.0	1994.000000	1994.000000	1994.000000	1994.000000	1.994000e+03	1.994000e+03
mean	7.358508e+17	12.273821	10.0	2766.753260	8895.725677	1.203109	0.593941	1.344195e-01	6.024848e-02
std	6.747816e+16	41.487825	0.0	4674.698447	12213.193181	0.560777	0.271954	1.006807e-01	5.089067e-02
min	6.660209e+17	0.000000	10.0	16.000000	81.000000	1.000000	0.044333	1.011300e-08	1.740170e-10
25%	6.758475e+17	10.000000	10.0	624.750000	1982.000000	1.000000	0.362857	5.393987e-02	1.619283e-02
50%	7.084748e+17	11.000000	10.0	1359.500000	4136.000000	1.000000	0.587635	1.174550e-01	4.950530e-02
75%	7.877873e+17	12.000000	10.0	3220.000000	11308.000000	1.000000	0.846285	1.951377e-01	9.159438e-02
max	8.924206e+17	1776.000000	10.0	79515.000000	132810.000000	4.000000	1.000000	4.880140e-01	2.734190e-01

From the above table we can see that the highest rating of a dog was 1776 this outlier is for a dog named Atticus.

this tweet was sent on July the 4th 2016 stating: “This is Atticus. He’s quite simply American af…”

we can see from this picture the neural network predicted correctly the breed of this dog regardless of Atticus’s costume.

We found that the most common source of tweeting is IPhone having 1955 tweets out of the data set. And the most common rating is 12 having 450 dogs rated 12.



From the visualization above we can see that doggo, floofer and pupper almost have the same average rating but pupper just excels. We can also see that the puppo stage has the least average rating.