@WeRateDog Data Wrangling Project

BY Lydia Fang

February 18, 2019

Introduction

@WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. WeRateDogs has over 4 million followers and has received international media coverage. Via Udacity, we were able to achieve the basic tweet data (tweet ID, timestamp, text, etc.) from @WeRateDog for all 5000+ of their tweets as they stood on August 1, 2017. More interestingly, all the posted images have been classified through a neural network to identify dogs'breed, and get different prediction probability for each dog. In this report, I was trying to exploring this huge dataset in different ways, and investigated the patterns and the relationship among people's rating, retweet, favorite, and prediction probability etc.

Exploration1 : The "most" dogs

First, I want to simply check which dogs were the "most" dogs online, such as the highest rating, the most favorites, most retweets, and the most confident prediction.

As expected, the dog who got the most retweets (60,715) also got the most favorites (126,596). Yes, "This is Stephan. He just wants to help. 13/10 such a good boy". He is a Chihuahua (Figure 1).

Figure 1. The most retweet and favorites dog Stephan

However, this boy was not the highest rated one. Then who got the highest rating online? Well, this rating numerator was way larger than the rating denominator of 10. Surprisingly, the highest rating of 165 was for a bunch of dogs. I also want to ask “Why does this never happen at my front door... ”

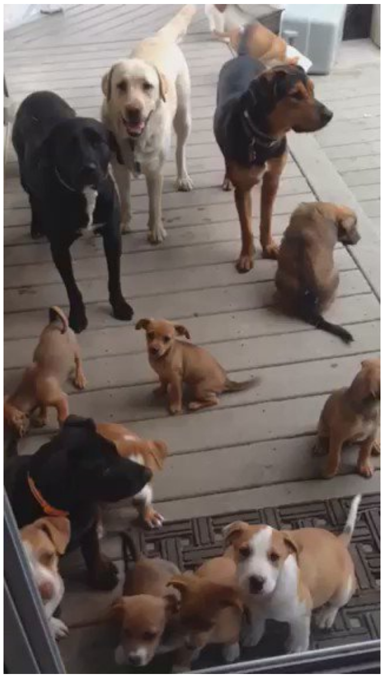

Figure 2. The highest rated picture

Next, I am interested in how does the neural network work? Who got the highest p1 prediction pr obability? In order to investigate this question, I removed all Non-p1dog items. Well, this big dog (Figure 3) got the highest p1 prediction probability (0.999956), and it had been identified as “Fals e” in the other two prediction level (p2 & p3). There was no name and stage information of this do g, but from the text of the post, we know “This is an Irish Rigatoni terrier named Berta....”.

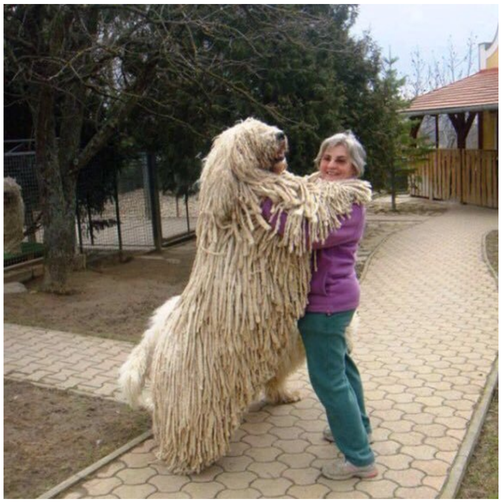

Figure 3. The highest p1 prediction dog

Exploration2: relationship among rating numerator, favorite count, retweet count, and p1 confidence

2.1 The relationship between favorite count and retweet count

As expected, the dog who got the most favorite count also got the most retweet count, and the correlation coefficient is very high (r = 0.93, $p < 0.0001$). The correlation scatter plot is shown in Figure 4. Since the favorite count is highly correlated with the retweet count, I only used favorite count for other analyses next.

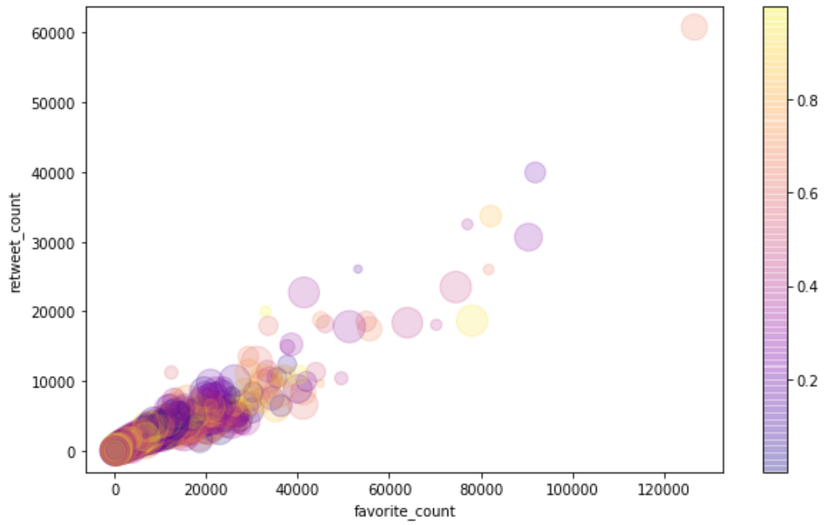

Figure 4. Correlation between favorite count and retweet count

2.2 Is there any relationship between favorite count and rating numerator?

To answer this question, I investigated the correlation between the rating numerator and the favorite count. However, from our early assessing, there were some obvious rating outliers, to assure the results more accurate, I removed the data points whose rating numerator larger than 15. As shown in Figure 5, the rating numerator was positively correlated with the favorite count (r = 0.41, $p < 0.0001$). And we can see that most ratings were in range 10-13.

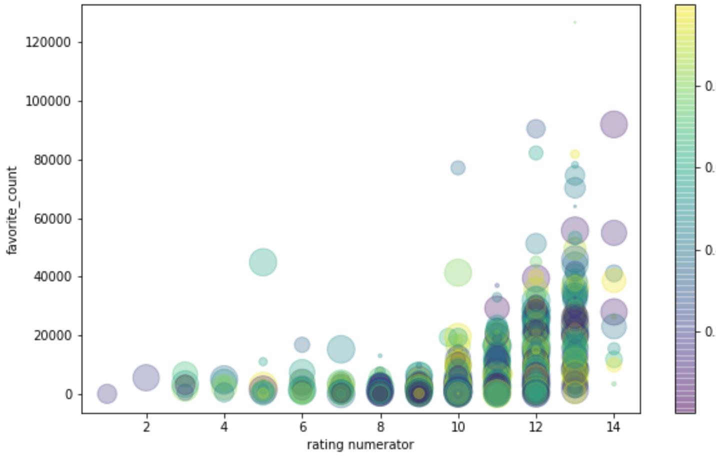

Figure 5. Correlation between favorite count and rating numerator

Exploration3:Dog stage and dog breed

Next, I want to explore the differences between different dog stages and dog breeds.

3.1 favorite count and p1 confidence distribution in different dog stage

As shown in Figure 6A & 6B, we can see that the averagely, doggo stage dogs had the most favorite count (mean = 16830.55), while puppo dogs had the highest p1 confidence (mean p = 0.70)

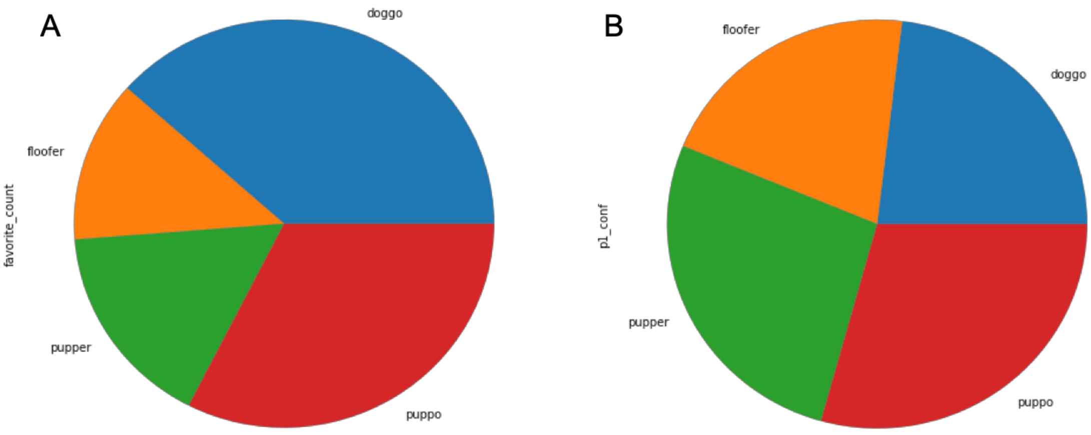

Figure 6. Favorite count and p1 confidence distribution in different dog stages

3.2 favorite count and p1 confidence distribution in different dog breed

Finally, I want to explore the differences between different dog breed. Since there are many different breeds, it is not necessary to investigate all of the them. Therefore, I just checked the first 20 breed of dogs who had the most favorite counts and p1 confidence. As shown in Figure 7A & 7B, we can see the breed of standard poodle had the most favorite count while the komonodor had the highest p1 confidence.

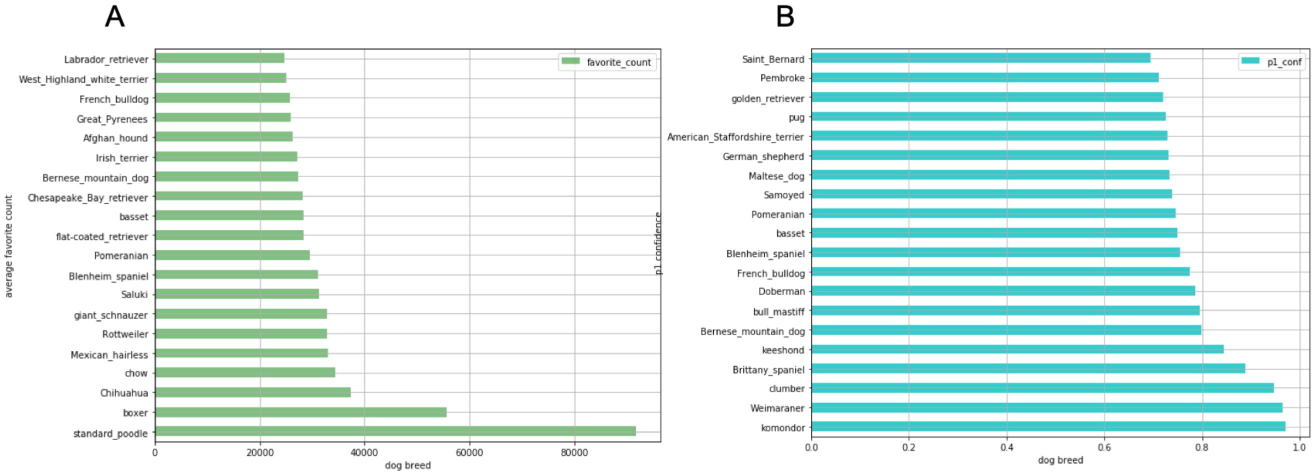

Figure 7. First 20 dog breed that had the most favorite count and highest p1 confidence