The WeRateDogs Project using twitter data
By Kamal Chahal

Introduction

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! In this project we used twitter API to import data of the user@dog_rates from weRateDogs. WeRateDogs is a very popular Twitter account with over 6 million followers and has received international media coverage. One of those coverage were about the quote *"they're good dogs Brent",*this was an exchange in which WeRateDogs shut down a person having an issue
with its rating system in humorous ways. WeRateDogs gained its popularity by rating people's dogs with a good-natured comment about the dog. The rating system is based on a fraction, with the denominator fixed at 10 and the numerator is almost always a number greater than 10. In theory the rates should be 1 to 10.
This rating helps us understand following questions:
1. Overall Find lowest rated dogs without filtering breed type?
2. What is the most common used names for dogs?
3. Which dog type (breed) is most common dog in the tweet dataset?
4. What dog type has the highest average rating?
5. How favorites vs retweets vs ratings co-related?

WeRateDogs has over 5000+ tweets. I was able to analyze around 1250+
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.

Below are the visualization & insights:

1. Visualization & insights - find lowest rated dogs

2pics were wrongly identified by Algorithm as Hen & fan. 2 lowest rated dogs. I did the analysis of low rated dogs. There was not a direct reason why some breeds received high ratings over remaining. It could be because they were more commonly owned or color or more commonly breed owner's comments or maybe they were easy to identify by the Algorithm intelligence which was classifying them.

2. <u>Visualization & insights - ratings vs time & favorites counts , retweets count vs time plots</u>

Favorites and retweets had a strong positive correlation. This means that if a tweet is good in general then it will get more retweets and favorites. Yet there is no correlation between rating and retweets or rating and favorites & all 3 factors increases with time. It can be because the dogs are not actually getting better. It can be that 'lower quality' dogs are given funnier captions. In this case, it is the caption that is getting more retweets and favorites, rather than the dog itself.
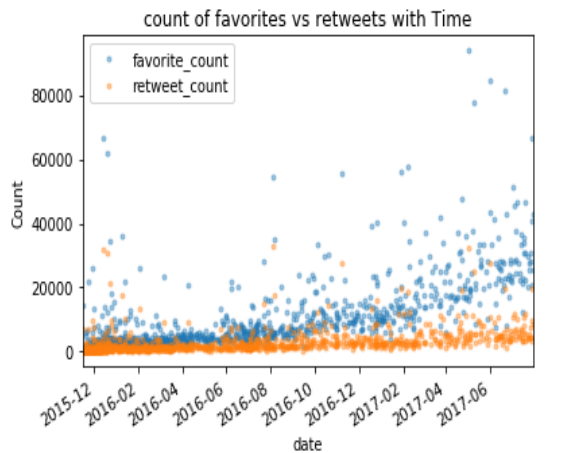
```
#Below is the perason correlation for favorites , retweet and rating _ numerator.
```

```
twitter_archive_master[['favorite_count', 'retweet_count','rating_numerator']].corr(method='pearson')
```

	favorite_count	retweet_count	rating_numerator
favorite_count	1.000000	0.915874	0.017996
retweet_count	0.915874	1.000000	0.027699
rating_numerator	0.017996	0.027699	1.000000

3. Visualization & insights – Commonly used dog names:

Oliver & Winston are more used dog names.

```
[('None', 418),
 ('Oliver', 8),
 ('Winston', 7),
 ('Tucker', 6),
 ('Lucy', 6),
 ('Cooper', 6),
 ('Penny', 5),
 ('Bella', 5),
 ('Toby', 5),
 ('Sadie', 5)]
```

4. Visualization & insights – Most rated & popular dog as per breed analysis:
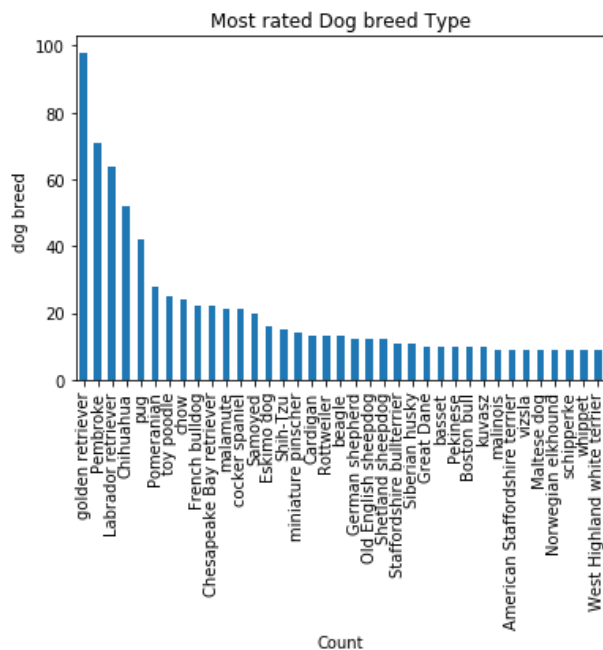
The most rated dog was golden retriever with more than 98 ratings. Mean ratings for golden retriever is 12.918367.

5. Visualization & insights – Lowest Rated dog as per dog_breed analysis:

Japanese_spaniel has the lowest rating while Clumber have the highest average ratings but only 1 person rated Clumber.