REPORT

On Data Wrangling Steps of WeRateDogs Twitter Data

Luiz Felipe Iurk
[Email address]

Abstract

The dataset wrangled in the project is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. This is a Twitter account that rates people's dogs with humorous comments about their dogs.

This project aims to wrangle the raw data following three steps: Gathering, Assessing and Cleaning.

And, as a last step, creating insights and visualizations of the extracted data.

Gathering

Three sets of data were provided to start this report. Each set of data was provided from a different sources as following:

- **WeRateDogs Twitter Archive**: This set of data was previously provided by Udacity in a .csv file under the name twitter_archive_enhanced.csv. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) of 2356 tweets as they stood on August 1, 2017.
- **Image Predictions**: This set was stored as a .tsv file and was stored at Udacity's online storage folder under the name image-predictions.tsv. The file contains the result of each tweet of the machine learning algorithm that returns the dog breed based on one of the 4 possible picture of each tweet.
- **Additional Tweet Information**: This set of data was supposed to be fetched through the Twitter API Tweepy and stored in a .json format.

Assessing Data

After the data was gathered, the data was assessed in two ways : Visual and Programmatically.

In both cases the objective was to evaluate quality and tidiness issues.

Among the issues found in the quality aspects, four of them are identified in this report:

1. Completeness: Missing data
2. Validity: The data don't conform to a defined schema
3. Accuracy: Wrong data that is valid. It is in the schema, but don't correspond to the reality
4. Consistency: Lack of Standardization

Among the issues found in the tidiness aspects, three of them are identified in this report:

1. Each variable forms a column

2. Each observation forms a row
3. Each type of observation unit forms a table

Cleaning Data

After assessing the data visually and programmatically, the cleaning process is done by following three steps:

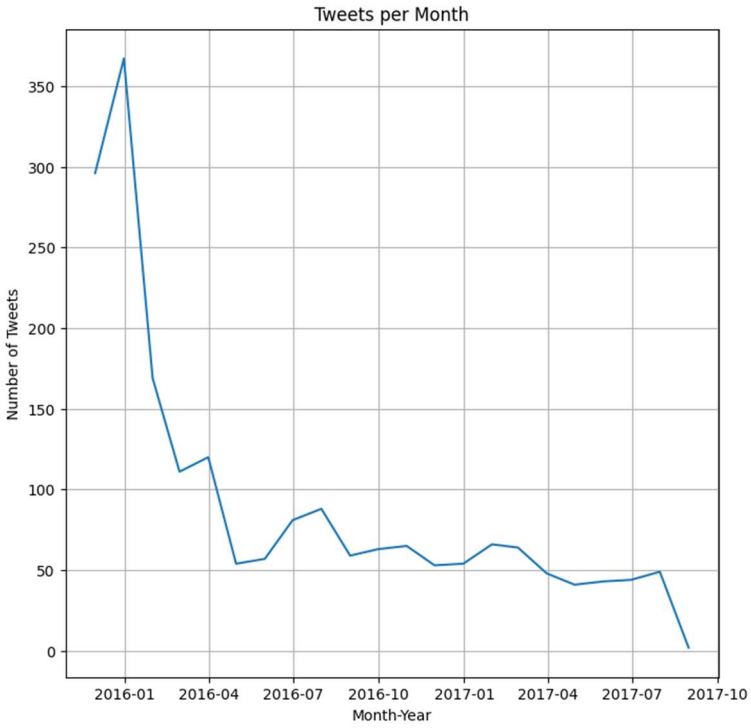
1. Define: Describe with action verbs exactly what will be done in the section;
2. Code: Shows all the python code lines used to achieve the definition;
3. Test: Shows that the defined action works.

Visualization

Several analysis could be done with the given data, but in order to keep it short and objective, only 4 aspects were analyzed:

1. Total Number of tweets over time

The number of tweets decreases substantially over time. Did people lose interest in WeRateDogs?

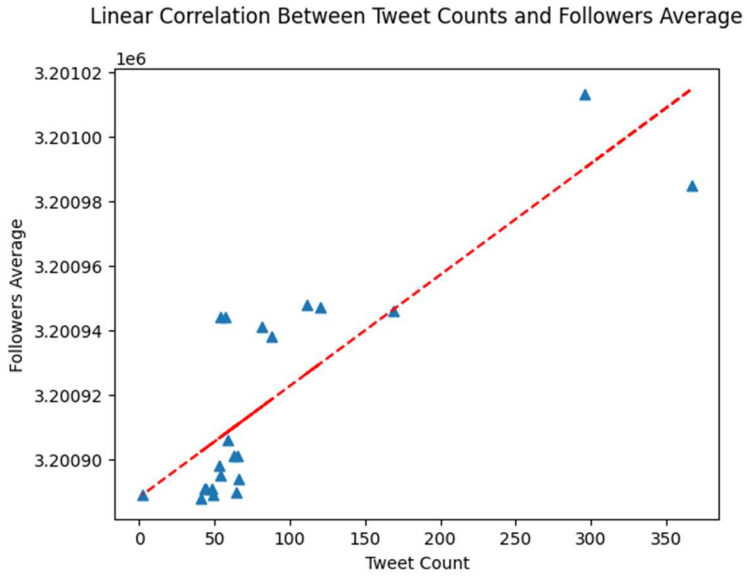


We notice that, in the beginning, there is a peak of number of tweets and it suddenly drops to 100 tweets per month, until it reaches a level of 50 tweets per month, with not so sharp peaks of increase. In my opinion, this is a clear sign of decrease of popularity, since there is no trend of increase after July/2017.

2. Correlation between tweets and average of followers

One valid guess is: “With higher number of followers, we should have more tweets”. Which makes sense, because, more people, more tweets.
A way to taste this assumption could be through a linear correlation, despite the fact that more elements can affect the number of tweets. The results are as follows:

	tweet_count	***followers_count***	***retweet_count***
tweet_count	1.000000	0.834819	-0.651202
followers_count	0.834819	1.000000	-0.859924
retweet_count	-0.651202	-0.859924	1.000000



Considering that the amount of tweets vary over the month, so does the amount of followers.

The Pearson's Ranking classifies the correlation above as a High Positive Correlation, which in other terms can mean, the more tweets we have, the more followers we will get.

However, it's important to be careful in this statement because Statistical Correlation Not Necessarily Describes Causality reference.

Therefore, a deeper investigation must be done to prove this statement

3. Top 10 most predicted breeds

113 Distinct breeds were identified by the algorithm with an average confidence level of 54.99%. The top 10 most predicted are:

Dog Breed	Number of Records
Golden Retriever	158
Labrador Retriever	108
Pembroke	95
Chihuahua	91
Pug	62
Toy Poodle	51
Chow	48
Samoyed	42
Pomeranian	42
Malamute	33

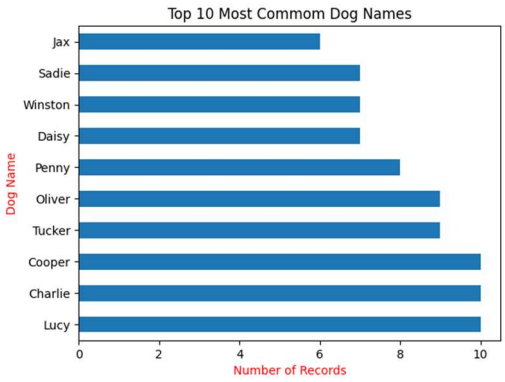
In the analysis above, it was discarded predictions that are not a dog.

4. Top ten 10 most common dog names

Most of the records from the dataset could not identify the dog's name. So, filtering out the "non-dogs" and the unidentified names, we end up with this top 10 list:

Name	Number of Records
Lucy	10
Charlie	10
Cooper	10
Tucker	9
Oliver	9
Penny	8
Daisy	7
Winston	7
Sadie	7
Jax	6

Showing it in a bar chart for better visualization:



Conclusion

The wrangling process of this data is far from being perfect. However after all the cleaning process, the data seems to be tidy enough to make valid assumptions.

The accuracy of the machine learning algorithm shows that there is room for improvement, because many dogs had the breed wrongly identified.

Regarding the free text extraction, this requires some improvement. But this requires a lot of time and to identify all the possible scenarios. In real life, this is an endless job.