Introduction

The dataset of interest is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. 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 are almost always greater than 10. WeRateDogs has over 4 million followers and has received international media coverage.

Gather (Data Source)

The first source of data is the Twitter archive sent downloaded by WeRateDogs and sent directly to Udacity. This archive contains basic tweet data, such as tweet ID, timestamp, text, etc., for over 5,000 of their tweets as of August 1, 2017. The dataset was pre-processed in a way that each tweet's text was analysed, and the information about the dog's rating, dog's name, and 'stage' (i.e. doggo, floofer, pupper, and puppo) were included in their respective columns in the dataset. Only the tweets with ratings were retained in this dataset, resulting in only having 2,356 rows of data, instead of the original 5,000+

The second source of data is image prediction provided by Udacity. Basically this dataset contains the top three predictions of the image (classifications of breeds of dogs), alongside the tweet ID

The third source of data is the additional data procured by using tweepy in Python, a Twitter API. Since the pre-processed dataset online contains the very basic information of each tweet, Twitter API was used to gather two pieces of additional information of each tweet, the retweet count and the favourite count.

Assess and Clean

The three datasets were assessed both visually and programmatically for quality and tidiness issues. The quality issues identified were erroneous column data types, inconsistent representation of dog names and predicted breeds of dogs (i.e., lower/uppercase, underscore

instead of space), inclusion of retweets in the dataset (only the original tweets should be included for later analysis), and columns that deemed irrelevant for later analysis and could be dropped.

There were also a few tidiness issues. The timestamp column in the pre-processed dataset should be separated into date and time columns. The dog stage columns should be combined into one single column instead of separately represented in four coluns.

After all the quality and tidiness issues were dealt with, the three separate datasets were merged into one final cleaned dataset. Note that inner joins were performed so that tweet IDs that were present in all three datasets were included.

Final Merged Data Set

The final merged dataset contains the following columns:

- tweet_id – Tweet ID
- date – Date of the tweet
- time – Time of the tweet
- source – Source of the tweet (i.e. iPhone, web, etc.)
- text – Original text of the tweet
- rating_numerator – The numerator of the rating
- rating_denominator – The denominator of the rating
- name – The name of the dog
- stage – The stage of the dog
- p1 – First predicted breed of the dog or object
- p1_conf – Probability of correct first prediction
- p1_dog – Whether the first prediction is a dog or not
- p2 – Second predicted breed of the dog or object
- p2_conf – Probability of correct second prediction
- p2_dog – Whether the second prediction is a dog or not
- p3 – Third predicted breed of the dog or object
- p3_conf – Probability of correct third prediction
- p3_dog – Whether the third prediction is a dog or not

- retweet_count – Number of retweets
- favourite_count – Number of times ‘favourited’