wrangle_report

December 16, 2018

Real-world data can be disorderly and unclean, making analyses difficult. In this project, I have collected tweet data and details for the twitter archive of the user @dog_rates, also known as WeRateDogs. WeRateDogs rates dogs with some humorous content about them. The ratings are out of 10, but these fun dogs all manage to score above 10. Since the twitter data consists primarily of tweets, additional data including favorite/retweet count data, and image predictions to identify dog breeds, were collected by querying Twitter's API and from alternate sources.

Data gathering: Sources: The data were collected from three sources. The basic tweet texts from 2015 to 2017, along with the information gleaned from it including the dog's name, stage and ratings were downloaded from the Udacity server as a csv file and saved as twitter-archive-enhanced.csv, and loaded into a dataframe using Pandas. Additional data were obtained by querying Twitter's API using the Tweepy library after creating a developer account, using the associated secure keys. The tweet_id indicated json files were downloaded and stored in a text file tweet_json.txt, and parsed to extract the favorite and retweet counts into a dataframe. Image_based breed predictions for the dogs, openly available from a url was downloaded as a tsv (image-predictions.tsv) and loaded into a third dtaframe.

Data wrangling: Upon assessment using the Pandas and Numpy fucntions, the following data tidiness and quality issues were identified. The prospective solutions are noted alongside.

Tidiness: 1. Type of dogs split into 4 columns in df_archive; join into one 2. jpg_url and img_num columns in df_impreds are not necessary for the planned analyses; remove these columns 3. df_addl should be part of df_archive; merge to the archive 4. Add df_impreds also to the df_archive for easier analyses

Quality: 1. in_reply_to_status_id and in_reply_to_user_id columns are reply tweets; remove rows 2. retweeted_status_id, retweeted_user_id and retweeted_time_stamps columns represent retweets; remove rows 3. Timestamp as string; change to datetime format 4. Source column has 4 unique values which represent the actual sources, lost in the html format; Pick out the source from the html link 5. expanded_urls column doesn't make sense; remove 6. Rating denominator should be 10 for all; change to 10 7. Rating numerator has low and high values; begin at 11, cap at 15 8. Datatype of tweet_id is int; change to string 9. Dog_types has some mixed types of dogs; consolidate to 'mixed_types'

Data Analyses: After merging the three dataframes and fixing the cleanliness issues, it was seen that there were 2356 instances and 19 attributes. These were the tweet_id, timestamp, source, text, rating_numerator, rating_denominator, name, dog_type, favorite_count, retweet_count and the 9 image prediction details columns. Of the 4 sources used to render the images, iPhone, Vine,

Web_client and Tweetdeck, iPhone was found to be the most frequent, with 2221 source listings. There were 690, 1183 and 483 tweets for the years 2015, 2016 and 2017 respectively. The maximum number of tweets during 2016 is expected since the tweets were collected from November 2015 to August 2017. From the information contained in the texts, the dogs were classified into dog stages, ranging from pupper to doggo to floofer, and alternative intermediate or composite stages. The pupper was seen to be the most frequent stage among the classifiable dogs (from the tweet texts) of the categories encountered viz. 'pupper', 'doggo', 'puppo', 'floofer' and 'mixed_types'. The highest favorite count was 140764, for an unnamed puppo, apparently a super supportive one participating in the Women's march in Toronto. The highest retweet count was 61150, for Stephan, a helpful dog predicted to be a Chihuhua with 50% confidence level. A scatter plot analysis showed that there was a strong correlation between the favorite and retweet counts for the dogs. The dogs received ratings between 11 and 15. The most frequent ratings were in the range of 11-12, numbering aroung 900. The average confidence levels for dog predictions based on their images were 60% for their first prediction, and lower for the subsequent two breed predictions.