Report on Wrangling of WeRateDogs (@dog_rates) Tweet Archive

July 4th, 2022
Victor Oladoja
Udacity Student

Dear Udacity Reviewer,

Good day,

As per the project to wrangle WeRateDog twitter dataset, I was given a csv file - twitter_archive_enhanced- that contains tweet_id, timestamp, ratings, dog name, dog stage, and other columns. Because the information is not enough for my analysis, I acquired a file - image_prediction-from udacity that contains information on tweet_id and dog breeds and I also queried Twitter API where I obtained additional data - twitter_add_data - such as retweet count and favourite (likes) count together with their tweet_id. The two additional files are obtained programmatically for reproductivity.

On assessing twitter_archive_enhanced.csv dataset, I discovered that some of the ratings are wrong with absurd ratings such as 1776/10 and the data contains duplications inform of retweet and reply. More so, dog stage variable is spread across four columns, timestamp column is in object data structure, there is no column for year which is important for my analysis, and many columns present are not needed for my analysis. For the dataset in image_prediction file, some of the predicted breeds are not dog breed in the three predictions with most probability. Other observable units only have correct dog breed name in their second or third predictions. Also, I noticed that many of the columns could be collapsed to a single column for clarity. For additional file I got from querying Twitter API, I went for tweet_id, favourite_count, and retweet_count data because they are what I need. Thus, the dataset is very clean from the get go.

I cleaned the tables as shown in the attached Jupyter Notebook. I re-extracted ratings, cleaned invalid dog name and cleaned name, retweet and reply rows. I converted timestamp data to datetime type and removed unnecessary columns. For image_prediction table, I removed all rows with no prediction that is a dog breed name. Then, I created a loop to iterate through the dataframe and picked dog breed from first prediction with a condition that the predicted name is for dog, second prediction if the former condition is not met, or the third prediction. The list I got from above was used to create dog breed column and I dropped all other columns apart from tweet_id and dog breed columns. Furthermore, I merged tweet_add_data with cleaned twitter_archive_enhanced tables because they have similar observations recorded. Lastly, I merged all the data together on the condition that the tweet_id of combined_twitter table is present in the image_prediction table to form a master table for insights and visualization. Through this, I was able eliminate all ratings that were from retweet.

Should you need further clarification, please do contact me. Thank you.

Yours sincerely,
Victor Oladoja
oladojavictor@gmail.com