Wrangling Report

by Eric Jones

Summary

This is a review of findings in the three files image_predictions.tsv, twitter_archive_advanced.csv, and tweet_json.txt containing Twitter data. After loading these onto a Jupyter notebook, I assessed the files one at a time before making any decisions on merging or sculpting the data into a cleaner form for analysis. Below is a review of each file.

Image_predictions.tsv

This file contains 2,075 rows of data based on predictions of breed of dog in Twitter photos with other non-dog photos. It appears that the image processing model has a training set of more than just dog photos since there are other categories present in columns called “p1,” “p2,” and “p3.” Confidence values of the strength of the prediction are held in columns “p1_conf,” “p2_conf,” and “p3_conf.” It is assumed there was another means to label which category was a dog and what was not with the columns “p1_dog,” “p2_dog,” and “p3_dog” with boolean True/False values. Other columns are the “tweet_id” (a unique key from each tweet), “jpg_url” (a place holding a scaled portion of the original photo), and “img_num” (the number of photos originally in the tweet - up to 4 in this data set).

There is, of course, some quality concern of the predictions themselves since there will be error from the model. However, this is not in my control nor was any information provided on how accurate the model was. It has to be assumed that, given columns designated with “p1” are primary predictions as a .describe() showed much higher confidence values than the other predictions, predictions are to be taken at face value and not questioned. (Examples were found to lead to poor predictions, however.)

Only light ***quality concerns*** were present with breed names being capitalized sometimes and not all or none. To be consistent, all labels were reformatted to lowercase.

twitter_archive_advanced.csv

This file contains 2,356 rows of data and contains a good deal of Twitter data from the WeRateDogs® account, as all “tweet_id” values are from this account. This contains reply and retweet IDs (largely unpopulated); timestamp of the tweet (YYYY/MM/DD HH:MM:SS format); a “source” column, which is where the tweet comes from concerning a platform, such as iPhone - 2,221 rows are iPhone; “text” column where the text of the tweet is stored; “expanded_urls” that can contain more than one url, but seems to be based on original tweets that individual users make and call “@dog_rates” that WeRateDogs® later references - only 2,297 rows available; two columns designated for ratings, “rating_numerator” and “rating_denominator”; “name” of the dog, but has values “None” and values “a,” “an,” and “the,” which are quality issues; and four “doggo lingo”-based columns “doggo,” “floofer,” “pupper,” and “puppo.”

This file seems to have some ***quality issues*** to note:

- “rating_denominator” contained at least one data extraction error when someone was commenting on “24/7” referring to a saying (24hrs a day, 7 days a week) and not a rating; an additional “error,” but more of a fan’s mistake that was corrected, a

denominator of 0 was not an acceptable use of the rating system. The majority of rated tweets have a 10 as the denominator - very few have anything else and was filtered
- “rating_numerator,” as noted above, had at least one erroneous example; most tweets seemed to have a range from 0 to 15, which leaves the rest to filter out
- “Name” contains “None” rather than null as well as examples of values “a,” “an,” “the,” and “such” as ‘names’; this column seems to have 957 unique names with very little distribution, which does not yield to a meaningful aggregation and analysis and was dropped

Tidiness issues are as follows:
- “expanded_urls” contain more than one url; later inspection shows up to 5, but with twitter.com only, up to 4; this did not seem to have any immediate value and was left alone
- “doggo,” “floofer,” “pupper,” and “puppo,” unfortunately are not flag columns and only contain “None” or the respective label corresponding to the name of the column; “doggo,” by definition, is all dogs and is redundant; 1,976 rows contain “None” across all four columns lending to little value for these labels; if there was more use/data, then this should be changed to either a Y/N or 1/0 value or one column - since for this data set, no lingo overlaps with the exception of “doggo” - with only the lingo label or null value; this column was dropped

Side note:
- “source” only has four unique values, 2,221 of which are for iPhone, which is not particularly useful and was dropped

tweet_json.txt

This file contains the JSON information per tweet with several fields. However, only three were chosen: “tweet_id,” “retweet_count,” and “favorite_count.” The latter columns are self-explanatory. The total row count is 2,354.

The only concern here is the fact that “retweet_count,” and “favorite_count” are not integers, but are strings. This ***quality issue*** is easily changed by calling .astype(int).

Overall

Overall, a ***quality issue*** would be that all three files do not contain the same number of rows. This results in reducing the number of rows in total to 2,073. It follows that the number of replies narrows down from 78 to 23 and retweets 181 to 79. Since retweets are being asked to be filtered, then additionally these 23 replies were filtered resulting in 1,971 rows left over. Further filtering the rating denominator to only values or 10 as well as filtering the numerator to less than 20 reduces the row count slightly to 1,949. Month was extracted from the timestamp for additional insights.

Other, but uncontrollable, quality issues would be the quality of the predictions are not known and have evidence of wrong predictions, and various other columns that play minor roles, such as, “img_num,” do not have accurate data to whatever degree. Since this data set

has more than just dog tweets, that also muddies the analysis of looking at rated dogs - even the predictions contain non-dog labels.

The ***tidiness issue*** is simply combining all three of these files to have one master data set.

Result

This data set, called “twitter_df” (saved as “master_archive.csv”) is the cleaned set with columns as follows: tweet_id, timestamp, month, text, expanded_urls, rating_numerator, retweet_count, favorite_count, p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_conf, p3_dog. For whatever purpose this data set may serve, I believe this to be as complete as possible.

My data set (“my_df”) was narrowed down for my analysis - I don’t have knowledge of advanced language processing techniques to utilize the above data set to its fullest potential. After filtering on “p1_dog=True,” since p1 seemed to have the stronger predictions, the row count reduced to 1,446 containing only columns “month,” “rating_numerator,” “retweet_count,” “favorite_count,” and “p1.” It was in best interest to create an additional data set based on a .value_counts() of the top 10 breeds that are tweeted the most, ranging from 29 to 134 counts. This data set, called “top10_dogs_df,” has the following columns: breed (distinct names from p1), rating_numerator, retweet_count, and favorite_count with a total of 629 rows.