Wrangling Efforts of WeRateDogs Twitter Data

Beshir Aman
Term2: Project 3

The aim of this report is to discuss mainly wrangling efforts made on WeRateDogs Twitter data prior data analysis part. There are three main wrangling steps: gather data, assess data quality and tidiness, and finally clean data by removing corrupt records and fixing tidiness issues.

Gather Data

In this step, the data need to be collected from three different data sources stored in different file formats. These data files are as follows:

Enhanced WeRateDogs Twitter Archive Dataset

An enhanced version of WeRateDogs Twitter Archive data was provided as a Comma-Separated Value (csv) file. The file was read locally and stored as a data frame ‘df_archive’.

The Tweet Image Predictions Dataset

This dataset has what breed of dog (or another object, animal, etc.) is present in each tweet according to a neural network. The file is stored in Udacity's server as a Tab-Separated Value (tsv) file. The file was downloaded programmatically from a given website URL and saved locally. The data was stored later as a data frame ‘df_predictions’ for further analysis.

Extra Tweet Information Dataset

Extra tweet's features such as retweet count and favorite ("like") count were fetched from Twitter using an API for each tweet ID in the WeRateDogs Twitter archive. Then, the data was stored locally as a text (txt) file. Afterwards, data was eventually prepared as a data frame ‘df_info’ to be wrangled and analyzed later.

Assess Data

Each dataset was assessed visually and programmatically, here are the findings:

1- Enhanced WeRateDogs Twitter Archive Dataset

The data have 2356 rows and 17 columns. The main features are: tweet id, timestamp, text, dog rating, and dog stage. The following quality issues were identified:

- Data contains retweets entries mixed with original tweets. Retweets are not of interest.
- Records where ‘in_reply_to_status_id’ values are not null, are not original tweets and don't have much information.
- The following columns are irrelevant to upcoming analysis: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp
- The timestamp column should be of type datetime instead of object type
- Some tweets have incorrect dog 'stage' information. Manual investigation should be conducted on cases were multiple stages are identified.
- Some tweets contain rating denominator different than 10 and not accurate.
- Some rating numerators have decimals are inaccurate.

In addition, there are a couple of tidiness issues:

- There is a column for each dog stage (doggo, floofer, pupper and puppo) instead of one column.
- Rating is divided into two columns: numerators and denominators. It can be combined into one as a ratio between numerator and denominator.
- All other datasets should be merged with this dataset

2- The Tweet Image Predictions Dataset

The data have 2075 rows and 12 columns. Each record includes an image URL and top three algorithm predictions where each prediction provides: whether it is a dog or not, dog breed, and confidence level. The following quality issues were identified:

- There are tweets with no images
- Some tweets have the same images

3- Extra Tweet Information Dataset

The data have 2345 rows and 3 columns. It mainly includes retweet count and favorite (like) count for each tweet in twitter archive. The following quality issues were identified:

- Some records are related to retweets, which are not of interest

Clean Data

First, the three data sets are copied. Then, the issues identified in the previous section are resolved and cleaned. Each issue is well defined, fix is coded, then the solution is tested. Finally, all datasets will be merged and saved as csv file named as ‘twitter_archive_master.csv’.

The following are the major issues resolved in order:

1- The retweet records are removed.
2- The following columns were dropped: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, source, expanded_urls, and name
3- The 'timestamp' column type was converted to datetime64 type
4- Some erroneous rating numerators and denominators are corrected.
5- Rating Ratio column was created
6- Dog stages were merged into one column
7- Dog breed predictions and confidence values were aggregated into prediction_type and prediction_confidence columns
8- The following columns were only kept: tweet_id, jpg_url, prediction_type and prediction_confidence
9- All cleaned three datasets were merged using tweet_id into df_master
10- Records with no images were removed
11- Final dataset was saved as ‘twitter_archive_master.csv’