Wrangle data steps: gather, assess, and clean

Report

By Muyul Alsubaie

12/22/2022

The dataset used in this project is the tweet archive of a twitter account @dog_rates WeRateDogs, the report will document the steps of gathering, assessing, and cleaning the data.

Data gathering

In this project, data were gathered in various formats and by diverse manners.

- Twitter Archive file was provided by Udacity as a csv file, I download it directly.
- Image prediction file wad provided as tsv file; I requested a library to download the tweet image prediction file.
- I downloaded tweet_json.txt via the Twitter API using Tweepy library.

Data assessing

After gathering the data, I began the data assessing process, focusing on both quality and tidiness issues.

Quality issues

On Twitter Archive:

- Retweets and replies should be deleted because original ratings with images are all that we really need.
- Missing values in columns: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, and expanded_urls.
- timestamp should be date instaed of str.
- tweet_id should be object instaed of int64.

On image Predictions:

- jpg_url has 66 duplicated values
- P1, P2, and P3's type of dog breeds contained both capital and lowercase letters.
- Names column need cleaning in Tarchive_clean .
- No rating standard.

Tidiness issues

- json_T should be part of the twitter_archive table.
- All tables should be part of one dataset.

Data Cleaning

Data was cleaned using the following methods and steps:

- retweeted status id and reply to status id columns, was removed using isnull().
- Missing values in columns: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, and expanded_urls in Tarchive_clean, I drop these columns and the useless coulmuns in the analysis using dropna() method.
- I fixed various datatyps using astype() method.
- jpg_url has 66 duplicated values in image_predictions_clean, I dropped duplicates using drop_duplicates() method
- P1, P2, and P3's type of dog breeds contained both capital and lowercase letters in image_predictions_clean, lower () method was used to uniform the letters.
- Names column need cleaning in Tarchive_clean, using drop() method and specifying a condition I managed to remove unwanted values.
- Devide rating_numerator by rating_denominator to get standardized rating.
- Merging tables using merge() method, because all tables should be part of one dataset.