WeRateDogs Tweets Archive – Wrangle Report

By Manish Kumar
31st July 2020

This report summarizes the process carried out to Wrangle the data from WeRateDogs twitter archive, i.e. Gather, Assess and Clean Data.

Gathering Data

Data gathering is the first step in the process if Data Analysis. The data is not always available at one single place. Its usually scatter and need to be gathered from different sources. For the sake of this project the data was gathered from 3 different sources.

1. The twitter archive of the WeRateDogs which was available for download from Udacity servers and was manually downloaded.
2. The image predictions dataset, which was programmatically downloaded from Udacity servers
3. Tweets data in JSON format, which was pulled from Twitter API's with the help of Tweepy library.

Assessing Data

Data assessment is the process of assessing the data and find Tidiness and quality issues with the available data. All the 3 different datasets were individually assessed for any such issues and below are the findings of the same: -

1. **Twitter Archive Dataset**
    a. The columns like doggo, floofer, pupper, puppo are not necessary. Instead one column category (a different name can be used as well) can serve the purpose
    b. Some tweets are re-tweets/replies which are not required for the analysis as they're not original tweets.
    c. There is one record for which the rating denominator is 0, which is invalid.
    d. There are two records for which the rating denominator is less than 10, which is not in line with the standard of 10.
    e. Not all columns are necessary for the analysis.

2. **Image Predictions Dataset**
    a. Column names p1, p2, p3 must be renamed to a more descriptive name.
    b. Not all tweets part of master database has image predictions which mean missing records.
    c. Around 324 predictions are not dogs according to either of the predictions and are thus irrelevant for analysis.
    d. The prediction mean for the first prediction is approx. 0.6 with a standard deviation of approx. 0.2 which is not a great number. For the sake of analysis if all of the predictions accuracy is less than 0.5 then we can ignore such records which will

result in a reduced dataset. But while analysis we will us the prediction that has maximum confidence probability and is a dog.

e. There are around 2075 unique tweets but only 2009 unique image URL's that shows missing predictions.

3. **Tweets Dataset**
    a. Columns like contributors, coordinates, geo has no values and thus are of no use.
    b. id column must be renamed to tweet_id for joining purpose. Also, the id column is present twice id, id_str.
    c. There are some re-tweets/replies which are not needed for analysis.
    d. Not all the columns are necessary for the analysis and thus while merging only the required ones must be considered.

Cleaning Data

All the issues listed above for the different datasets were addressed while cleaning the dataset.

- The columns doggo, floofer, pupper, puppo were merged into one column "category".
- The re-tweets/replies were removed to exclude from the analysis.
- The records in the master archive that have a rating denominator or less than 10 were removed as it not in line with the standard rating out of 10.
- Column names p1, p2, p3 in the prediction dataset were renamed and even merged into one single columns "prediction" based on the maximum confirmation accuracy and the prediction being a dog.
- The predictions which were not dogs according to any of the 3 predictions were dropped.
- The re-tweets/replies were removed from the tweets dataset. Also while merging only the favorite_count, retweet_count were taken as they were only relevant columns for the analysis.

Though the above assessment and cleaning process may not have addressed all the issues, but we can always re-iterate. For the sake of this report, only one iteration was considered.