Data Wrangling Effort Report

By: Mahmoud Khaled

February, 2021

Data Gathering

In this stage, I collect data from three different sources:

1. “Twitter_Archive_enhanced.csv” file, I downloaded this file manually into my working directory, then used Pandas’ “pd.read_csv” to import it into my Jupyter notebook.
2. “Image_predictions.csv” file, I downloaded this programmatically using Requests library get function then Pandas’ “pd.read_csv” to import into the working directory. This file was made using a neural network to determine the breed of the dog in a picture.
3. “Twitter_json” file, I used twitter REST API via Tweepy library to obtain extra information that’d be useful in my analysis, e.g. retweets count and favorite count.

Data Assessment

- The visual assessment was done on a spreadsheet application.
- Programmatic assessment was done on Jupyter notebook.
- First tidiness issues were assessed, followed by quality issues.
- In the table below, all the issues and their solutions are addressed.

Tidiness

Issue	Solution
Master data frame has values as column names (doggo, pupper, floofer, puppo)	Combined in 1 columns “dog_stage”
The three datasets can be considered of the same type of observational unit.	Merged the three together.

Quality

Master dataframe

Issue	Solution
Invalid data type for timestamp column. (object instead of datetime)	Change type to datetime.
Unwanted entries that do not contain images.	Drop those entries.
Unwanted entries that are not original tweets. (retweet or replies)	Drop those entries.
Tweet_id is an int instead of a string.	Change type to string.
Ratings are extracted wrong when the numerator is a fraction. (ex: 15.5/10 is 5/10)	Re-extract them using regex.
Missing ratings from original tweets.	Set ratings to NaN.
Invalid rating for pictures of multiple dogs.	Calculate the rating for each dog.
Inaccurate parsed ratings that are actually correct but do not follow the specified schema. (ex 1776/10 & 420/10)	Set ratings to NaN.
Rating extracted wrong when it's the second fraction in a tweet. (ex: tweet at index 20)	Fix them manually.
Wrongly extracted / missing names.	Re-extract them using regex.
Master data frame has two values in one column “timestamp” (date and time)	Put each one in separate columns
Predictions_df can be reshaped for better clarifications. Only the first level of predictions is relevant.	Reshape using wide_to_long, and only keep the first level of predictions results, drop the rest.
Some breed names start with lowercase, and others with uppercase.	Make them all lowercase.