WeRateDogs tweet Archive Analysis: Data Wrangling Process

Michael Smales

August 2022

Introduction

WeRateDogs is a popular Twitter account which rates pictured of dogs submitted by other Twitter users. For this analysis, an archive of WeRateDogs tweets dating between 2015 and 2017 was gathered, assessed and cleaned, preparing it for subsequent analysis.

Three component data sources were used:

- An archive of the tweets, provided by the Udacity team
- Tweet's performance stats (i.e. likes and retweets), sourced from the Twitter API
- Dog breeds, based on an image classification algorithm, again provided by the Udacity team

All wrangling was carried out using Python in a Jupyter Notebook environment in the VS Code editor.

Wrangling: Gathering Phase

The three data sources were gathered and each was loaded into a separate dataframe.

The enhanced tweet archive was provided as a .csv file by Udacity. The file was manually downloaded and read into a dataframe using pandas `.read_csv()` function.

Image prediction data was provided as a .tsv file on a server provided by Udacity. The Python `requests` library was used to download the file, and then write it to a local .tsv. file.

Like and retweet data was gathered form the Twitter API, using the Tweepy access library. A Twitter Developer was account set up to obtain API credentials. Requests were made one-by-one using the tweet_id, and the results stored in a json-style object. The data was written then written to a local tweet_json file using the json library's .dump() function.

Wrangling: Assessment Phase

Each of the 3 files was loaded into a separate dataframe for assessment. Each dataframe was visually assessed by loading the head of the dataframe and by using VS Code's data viewer functionality. Assessment of the data tidiness (i.e. structure) was made in this stage. Following that, programmatic assessment was carried out using Python and Pandas functions.

Issues found from both phases were noted in the Jupiter notebook, and are reproduced here for reference:

Quality issues:

- The tweet archived includes retweets and replies
- Dog named 'a' should either be named 'None' or their actual name
- Dogs named 'the' should be named 'None'

- Missing retweet / like data for some tweets
- Missing image predictions for some tweets - **not addressable**
- Some expanded urls are missing
- Source is encoded as an HTML tag, but should be a categorical
- Expanded urls generally contain repeats of the same url separated by a comma
- Timestamp is string but should be a datetime
- Doggo, floofer, etc are encoded as string
- Tweet 810984652412424192 should be dropped: it is not a rating
- Tweet 666287406224695296 rating should be corrected to 9/10 (incorrectly picked up as 1/2)

Tidiness issues

- Doggo, floofer, pupper and puppo should be encoded as a single categorical column
- The tweet archive contains information both about the dog (e.g. the dog's name) and the tweet (e.g. the source of the tweet). These should be separated into two separate tables

Wrangling: Cleaning Phase

Firstly, copies of each of the original three dataframes were made. Subsequently, each issue noted above was cleaned one by one using Python and Pandas functionality, following a define-code-test process. This stage resulted in 3 datasets:

- A dog_ratings table
- A tweet_stats table
- An image prediction table (containing the dog image predictions)

A separate dog_id index was added to the dog ratings table as a primary key, for ease of use in the future. The tweet_id was retained so that the datasets could easily be joined.

Storing Data

The three datasets were saved locally as .csv files for future analysis.