wrangle_report

Eileen Hertwig

March 11, 2019

1 Gathering

To get all the data, three different steps were necessary. The twitter archive was provided by *WeRateDogs* and just needed to be uploaded to the workspace and opened as a pandas dataframe. The second file contained the image predictions and could be downloaded, which I did programmatically using the *Requests* library. To get additional information about the retweets and favorites counts, I used the twitter API *tweepy*.

2 Assessing

For the visual assessment, I looked at all three tables to get an overview of the structure and to see if I can spot any problems with cleaniness or tidiness this way. Afterward, I used programmatic assessment to catch more problems. These are the issues I identified: **Quality:** *archive table*

- retweets are included
- replies are included
- missing values in various columns (`in_reply_to_status_id`, `in_reply_to_user_id`, `retweeted_status_id`, `retweeted_status_user_id`, `retweeted_status_timestamp`, `expanded_urls`)
- id-columns (`in_reply_to_status_id`, `in_reply_to_user_id`, `retweeted_status_id`, `retweeted_status_user_id`) should not be floats
- timestamp-columns (`timestamp`, `retweeted_status_timestamp`) should not be objects
- 745 times 'None' as a dog name
- many incorrect dog names (55 times 'a', 8 times 'the', 7 times 'an' as dog name)
- some very low and some very high values in the numerator of the rating

- denominator is not always 10
- many missing values in dog stages columns
- many posts are not ratings, but posts saying "we only rate dogs, stop sending other pictures" or similar

predictions table

- some dog types are capitalized, other are spelled with lower case letters
- less entries than in the archive table

counts table

- less entries than in the archive table

Tidiness:

- the data is scattered over three tables
- dog-stages should be represented in one column and not in four

3 Cleaning

To start the cleaning, I made a copy of all three tables so the raw data is still available.

3.1 Missing Values

As a first step I tried to address the issues with missing values.

The `retweeted` or `reply` columns should actually all be NaNs, since we only want original ratings in our data, so all rows that contain non-null values in these columns are removed and these columns droped.

The issue with the missing values in the `expanded_urls` column cannot be addressed right now, since this data is not available for these tweets.

Not all tweets contain information about dog stages, but I was able to fill some gaps by searching the text fields for the words "floofer", "doggo", "pupper", and "puppo" (as well as the capitalized and plural versions of these dog words).

3.2 Tidiness

To continue with the dog stages, I addressed the issue with the four different columns for next. The data would be tidier if we only had one column called `dog_stages` with "floofer", "doggo", "pupper", and "puppo" as possible entries. The type of this column should be 'category'. So, I reorganized the data in that way and droped the four original columns.

To reorganize the amount of tables I decided to keep all the information in one table. One could probably argue that we have information about the dog and about the tweet, but it is sometimes really difficult to keep these apart. Therefore, I merged all three tables into one called `master`.

3.3 Cleaniness

For comparison of dog types, it would be better if all are either capitalized or spelled in lower case. Because it was mixed before, I capitalized the beginning of each word in the `p1`, `p2`, and `p3` columns.

For the types of the columns, the issue with the id-columns does not need to be addressed anymore, since they have already been removed. The timestamp column, however, still remains and was changed to be of type 'datetime'.

I removed all tweets that are not ratings but contain phrases like "we/I only rate dogs", "this is not a dog", etc., since apparently a lot of people keep sending pictures of animals or things that aren't dogs.

To fix the rating columns, I searched all text fields for the string '/10' because most ratings are some value out of 10. If this string could be found the denominator is set to 10 and the string before that is the numerator (for a few cases I had to fix this string because there was no space in between the text and the rating). If the '/10' string was not found, I just looked for '/' for ratings with a different denominator. These weren't many, so for all these ratings I looked at the text to find out why the rating denominator was different. In one case it was '24/7', so obviously not a rating. In all other cases, it was apparent from the text, that more than one dog was described by this rating and usually the numerator was also quite high. To make the rating easier to analyze I scaled the rating values to one dog each, so I divided the rating numerator and denominator by the amount of dogs this rating refered to. Now all denominators are 10 (or missing). The rating_denominator has been assigned the type 'int', while the rating_numerator has been assigned the type 'float' (floats were necessary when dividing by the number of dogs).

To fix the dog names, I firstly identified some obviously wrong names, like 'None', 'a', 'an', 'the', 'such', 'life'. I replaced all of these with NaNs. Then I looked for some phrases that are usually followed by dog names in the tweets, like 'named', 'name is', 'Meet', 'meet', 'is called'. For all cells with missing dog names I searched the text fields for these phrases. If the word following these phrases was not included in the bad names I identified at the start, I assigned it as the dog name. There are still missing values (since some tweets just don't include the dog name), but some cells could be filled this way.