2020

DATA WRANGLING REPORT

WeRateDogs™ @dog_rates · Apr 16

Say hello to Alice. I'm told she enjoys car rides and smells good. 12/10 would give her everything she could ever want

WeRateDogs™ @dog_rates · Apr 16

A photographer took pictures before and after he told his bunny he's a good boy. Here are the results. 13/10

Mohamed Khaled Galloul

11/18/2020

Gathering Data

In this process I had to deal with 3 different file formats ('.csv','.tsv','.txt') each file has been downloaded or uploaded differently, like this:

1st file: twitter-archive-enhanced.csv

This file was already found on my virtual machine path where my main notebook (wrangle_act.ipynb) was also there.

2nd file: image-predictions.tsv

This file was hosted by Udacity servers so I had to download it programmatically using its URL by using requests library then save it to my virtual machine path.

3rd file: tweet-json.txt *

For this file I had to downloaded it manually using its URL to my local computer first, then upload it to my virtual machine path.

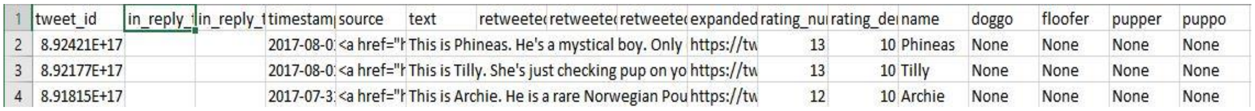
*: this file should be downloaded programmatically from Twitter API after I get an access, unfortunately I couldn't get it so I had to choose the alternative way which mentioned above.

Assessing Data

First, let's consider the two assessment method that I used:

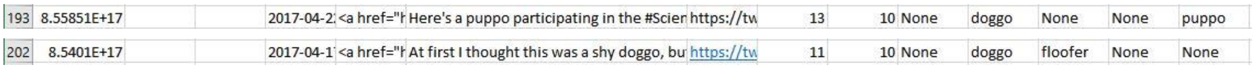
Manually:

1. Using MS Excel to get an overview twitter-archive-enhanced dataset issues



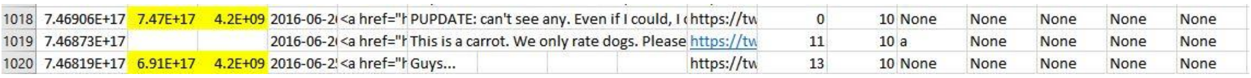
1	tweet_id	in_reply_t	in_reply_t	timestam	source	text	retweete	retweete	retweete	expanded	rating_nu	rating_de	name	doggo	floofer	pupper	puppo
2	8.92421E+17			2017-08-0	<a href="h	This is Phineas. He's a mystical boy. Only				https://tw	13	10	Phineas	None	None	None	None
3	8.92177E+17			2017-08-0	<a href="h	This is Tilly. She's just checking pup on yo				https://tw	13	10	Tilly	None	None	None	None
4	8.91815E+17			2017-07-3	<a href="h	This is Archie. He is a rare Norwegian Pou				https://tw	12	10	Archie	None	None	None	None

Figure 1.a : some issues found in twitter-archive-enhanced dataset in dog stage columns



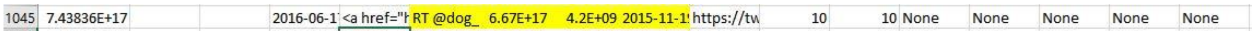
193	8.55851E+17			2017-04-2	<a href="h	Here's a puppo participating in the #Scien				https://tw	13	10	None	doggo	None	None	puppo
202	8.5401E+17			2017-04-1	<a href="h	At first I thought this was a shy doggo, bu				https://tw	11	10	None	doggo	floofer	None	None

Figure 1.b : some issues found in twitter-archive-enhanced dataset in dog stage columns



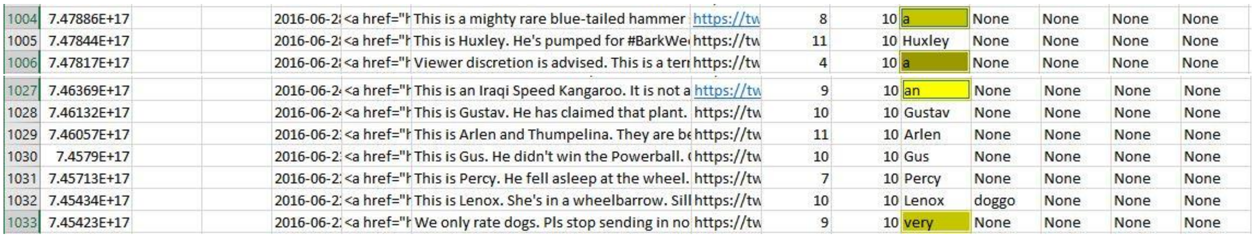
1018	7.46906E+17	7.47E+17	4.2E+09	2016-06-2	<a href="h	PUPDATE: can't see any. Even if I could, I				https://tw	0	10	None	None	None	None	None
1019	7.46873E+17			2016-06-2	<a href="h	This is a carrot. We only rate dogs. Please				https://tw	11	10	a	None	None	None	None
1020	7.46819E+17	6.91E+17	4.2E+09	2016-06-2	<a href="h	Guys...				https://tw	13	10	None	None	None	None	None

Figure 2.a : some replies found in twitter-archive-enhanced dataset



1045	7.43836E+17			2016-06-1	<a href="h	RT @dog_	6.67E+17	4.2E+09	2015-11-1	https://tw	10	10	None	None	None	None	None

Figure 2.b: one example of retweets found in twitter-archive-enhanced dataset



1004	7.47886E+17			2016-06-2	<a href="h	This is a mighty rare blue-tailed hammer	https://tw	8	10	a	None	None	None	None
1005	7.47844E+17			2016-06-2	<a href="h	This is Huxley. He's pumped for #BarkWe	https://tw	11	10	Huxley	None	None	None	None
1006	7.47817E+17			2016-06-2	<a href="h	Viewer discretion is advised. This is a ter	https://tw	4	10	a	None	None	None	None
1027	7.46369E+17			2016-06-2	<a href="h	This is an Iraqi Speed Kangaroo. It is not a	https://tw	9	10	an	None	None	None	None
1028	7.46132E+17			2016-06-2	<a href="h	This is Gustav. He has claimed that plant.	https://tw	10	10	Gustav	None	None	None	None
1029	7.46057E+17			2016-06-2	<a href="h	This is Arlen and Thumpelina. They are be	https://tw	11	10	Arlen	None	None	None	None
1030	7.4579E+17			2016-06-2	<a href="h	This is Gus. He didn't win the Powerball.	https://tw	10	10	Gus	None	None	None	None
1031	7.45713E+17			2016-06-2	<a href="h	This is Percy. He fell asleep at the wheel.	https://tw	7	10	Percy	None	None	None	None
1032	7.45434E+17			2016-06-2	<a href="h	This is Lenox. She's in a wheelbarrow. Sill	https://tw	10	10	Lenox	doggo	None	None	None
1033	7.45423E+17			2016-06-2	<a href="h	We only rate dogs. Pls stop sending in no	https://tw	9	10	very	None	None	None	None

Figure 3 : some wrong or badly extracted names found in twitter-archive-enhanced dataset

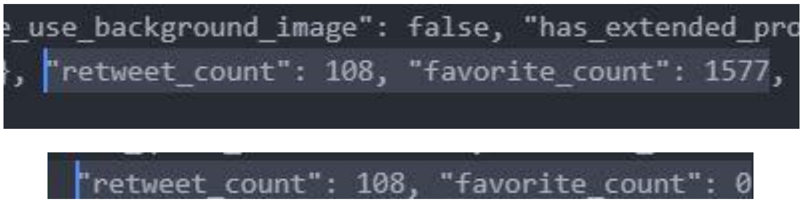
Quality issues:

- Dog stage four columns contain 'None' instead of `NaN` as shown in figure 1.a
- Some tweets are actually retweets and replies not original tweets also have to be deleted as shown in figures 2.a & 2.b
- Name column contain 'a', 'an', 'not', 'his','the' or 'very' ... instead of `NaN` (or the true name if found!) as shown in figure 3
- 'None' is entered for dogs with no names instead of `NaN`

Tidiness issues:

- Dog stage in found in four columns instead of one column as shown in figure 1.a
- Some tweets pictures contain more than one dog (e.g., father and son) which conduct two dog stages for the same tweet as shown in Figure 1.b

2. Using Atom text editor to get an overview tweet-json dataset issues



```
_use_background_image": false, "has_extended_pro
}, "retweet_count": 108, "favorite_count": 1577,
```

```
"retweet_count": 108, "favorite_count": 0
```

Figure 4 : some tweets have 0 for favorite count primary key, instead of the true value found in favorite count key of retweet status primary key.

Quality issue:

- Some favorite_count values found to be '0' as the primary key of the tweet object have a value of zero. So the alternative approach is to get `favorite_count` from `retweet_status` as both retweet_count are equal **(solved while gathering)** as shown for tweet 32 in figure 4 .

Tidiness issue:

- This dataset should be merged with twitter-archive-enhanced to add to master data set.

Programmatically

1. In twitter-archive-enhanced dataset as ‘archive_df’

Quality issues:

for `rating_denominator` column using ‘.value_counts() ’ method

- `rating_denominator` can be 'multiples of 10' based on the number of dogs in the same picture in the tweet.
- `rating_denominator` other values of '2','7','11' , '15' , '16' and others can be found. (manaual fixation needed in this case)

for `rating_denominator` column using ‘.value_counts() ’ method

- `rating_numerator` contain retweet count instead of rating. (manaual fixation needed in this case)
- `rating_numerator` contain some wrong rating values 'BELOW 10' (other animals than dogs) or 'OVER10' .

for `timestamp`, `expanded_url` columns using '.info() ' method

- `timestamp` column should be of type datetime.
- `expanded_url` blank spaces should be dropped as they don't contain any image to predict.

2. In image-predictions dataset as 'image_predictions_df'

Quality issues:

By using '.columns()' :Undescriptive column names: `p1`, `p1_conf`, `p1_dog` instead of 'prediction1','confidence1','is_dog1'

Dog breeds in `p`'s columns either in uppercase or lowercase

Tidiness issue:

By using '.info()' method :(p1,p2,p3), (p1_conf,p2_conf,p3_conf), and (p1_dog ,p2_dog ,p3_dog) are in 3 columns each instead of one column for each feature

Cleaning Data

Cleaning process for each data set will be as following:

1. For twitter-archive-enhanced dataset as `archive_df`

   a. Making a copy of archive_df called archive_df_cleaned
   b. Replacing each 'None' with `NaN` in [`doggo`, `floofer`, `pupper`, `puppo`] columns.
   c. Replacing ('a', 'an', 'not', 'his','the', 'very', ..) or any name found to be starting with lowercase with `NaN`, as all non-dog names found

to be starting with lowercase while real dogs names where starting with uppercase.

d. Replace 'None' with `NaN` in `name` column
e. Gather [`doggo`, `floofer`, `pupper`, `puppo`] columns into one column called ‘dog_stage’.
f. Drop empty cells in the `expanded_url` column cause they refer to no pictures
g. Drop 79 replies and 183 retweets by using: `in_reply_to_user_id` and `retweeted_status_user_id` columns then drop ALL related columns to replies and retweets
h. Drop `tweet_id` from `archive_df` if not found in `image_predictions_df` .
i. Change `timestamp` dtype from object to datetime
j. Replace multiples of 10 found in `rating_denominator` to 10 and also replace the corrosponding `rating_numerator` with correct values based on the number of dogs.**

** P.S.:

`rating_numerator` contains retweet count instead of rating and contains some wrong rating values 'BELOW 10' (other animals than dogs) or 'OVER10' like 9.75 is 75/10 for example!!

Generally, both would require to be cleaned manaully (which would be out of my cleaning process scope),so based on that we would consider valid numerator values to be with in [10,14] interval

2. For image-predictions dataset as `image_predictions_df`

a. Making a copy of image_predictions_df called image_predictions_df_cleaned
b. Rename `p1` to 'prediction1' ,`p1_conf` to 'confidence1,`p1_dog` to 'is_dog1'
c. Group each feature of (`p`,`p_conf`,`p_dog`) under one column instead of three.

d. Standardize all dog breeds to start with uppercase.
e. Merging `image_predictions_df_cleaned` with `api_df_cleaned` to form 'twitter_archive_master' dataset

3. For tweet-json dataset as `api_df`

a) Making a copy of api_df called api_df _cleaned
b) Merging `archived_df_cleaned` with `api_df_cleaned` to form 'twitter_archive_master' dataset

Storing Data

Into one file

- `twitter_archive_master.csv` *

***:for analysis purposes we will use `twitter_archive_master` data frame that contains only `archive_df_cleaned` and `api_df_cleaned`, and `image_predictions_df_cleaned` seperately to produce clearer insights for both datasets, rather than using `master_df`.**