Wrangle Report

Chaudhary Sarimurrab

25-Apr-2020

1. Gathering

For gathering the dataset, I followed the suggestions.

First dataset was "twitter-archive-enhanced.csv" downloaded from given link.

Enhanced Twitter Archive

The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, but not everything. One column the archive does contain though: each tweet's text, which I used to extract rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo) to make this Twitter archive "enhanced." Of the 5000+ tweets, I have filtered for tweets with ratings only (there are 2356).

text	rating_ numerator	rating_ denominator	name	doggo	floofer	pupper	puppo
This is Phineas. He's a mystical boy. Only ever appears in the hole of a donut. 13/10 https://t.co/MgUWQ76dJU	13	10	Phineas	None	None	None	None
This is Tilly. She's just checking pup on you. Hopes you're doing ok. If not, she's available for pats, snugs, boops, the whole bit. 13/10	13	10	Tilly	None	None	None	None
This is Archie. He is a rare Norwegian Pouncing Corgo. Lives in the tall grass. You never know when one may strike. 12/10 https://t.co	12	10	Archie	None	None	None	None
This is Darla. She commenced a snooze mid meal. 13/10 happens to the best of us https://t.co/tD36da7qLQ	13	10	Darla	None	None	None	None
This is Franklin. He would like you to stop calling him "cute." He is a very fierce shark and should be respected as such. 12/10 #BarkW	12	10	Franklin	None	None	None	None
Here we have a majestic great white breaching off South Africa's coast. Absolutely h*ckin breathtaking. 13/10 (IG: tucker_marlo) #Bar	13	10	None	None	None	None	None
Meet Jax. He enjoys ice cream so much he gets nervous around it. 13/10 help Jax enjoy more things by clicking below https://t.co/Zr4hWfAs1H https://t.co/tVJBRMnhxl	13	10	Jax	None	None	None	None
When you watch your owner call another dog a good boy but then they turn back to you and say you're a great boy. 13/10 https://t.co/\	13	10	None	None	None	None	None
This is Zoey. She doesn't want to be one of the scary sharks. Just wants to be a snuggly pettable boatpet. 13/10 #BarkWeek https://t.c	13	10	Zoey	None	None	None	None
This is Cassie. She is a college pup. Studying international doggo communication and stick theory. 14/10 so elegant much sophisticat	14	10	Cassie	doggo	None	None	None
This is Koda. He is a South Australian deckshark. Deceptively deadly. Frighteningly majestic. 13/10 would risk a petting #BarkWeek ht	13	10	Koda	None	None	None	None
This is Bruno. He is a service shark. Only gets out of the water to assist you. 13/10 terrifyingly good boy https://t.co/u1XPQMl29g	13	10	Bruno	None	None	None	None
Here's a puppo that seems to be on the fence about something haha no but seriously someone help her. 13/10 https://t.co/BxvuXk0UC	13	10	None	None	None	None	puppo
This is Ted. He does his best. Sometimes that's not enough. But it's ok. 12/10 would assist https://t.co/fBdEDcrKSR	12	10	Ted	None	None	None	None
This is Stuart. He's sporting his favorite fanny pack. Secretly filled with bones only. 13/10 puppared puppo #BarkWeek https://t.co/y70x	13	10	Stuart	None	None	None	puppo

Second was "image-predictions.tsv" downloaded from udacity server.

Image Predictions File

One more cool thing: I ran every image in the WeRateDogs Twitter archive through a neural network that can classify breeds of dogs*. The results: a table full of image predictions (the top three only) alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction (numbered 1 to 4 since tweets can have up to four images).

tweet_id	jpg_url	img_num	p1	p1_conf	p1_dog	p2	p2_conf	p2_dog	p3	p3_conf	p3_dog
892177421306343426	https://pbs.twimg.cor	1	Chihuahua	0.323581	TRUE	Pekinese	0.0906465	TRUE	papillon	0.0689569	TRUE
891815181378084864	https://pbs.twimg.cor	1	Chihuahua	0.716012	TRUE	malamute	0.078253	TRUE	kelpie	0.0313789	TRUE
891689557279858688	https://pbs.twimg.cor	1	paper_towel	0.170278	FALSE	Labrador_retriever	0.168086	TRUE	spatula	0.0408359	FALSE
891327558926688256	https://pbs.twimg.cor	2	basset	0.555712	TRUE	English_springer	0.22577	TRUE	German_short-haired_pointer	0.175219	TRUE
891087950875897856	https://pbs.twimg.cor	1	Chesapeake_Bay_retriever	0.425595	TRUE	Irish_terrier	0.116317	TRUE	Indian_elephant	0.0769022	FALSE
890971913173991426	https://pbs.twimg.cor	1	Appenzeller	0.341703	TRUE	Border_collie	0.199287	TRUE	ice_lolly	0.193548	FALSE
890729181411237888	https://pbs.twimg.cor	2	Pomeranian	0.566142	TRUE	Eskimo_dog	0.178406	TRUE	Pembroke	0.0765069	TRUE
890609185150312448	https://pbs.twimg.cor	1	Irish_terrier	0.487574	TRUE	Irish_setter	0.193054	TRUE	Chesapeake_Bay_retriever	0.118184	TRUE
890240255349198849	https://pbs.twimg.cor	1	Pembroke	0.511319	TRUE	Cardigan	0.451038	TRUE	Chihuahua	0.0292482	TRUE
890006608113172480	https://pbs.twimg.cor	1	Samoyed	0.957979	TRUE	Pomeranian	0.0138835	TRUE	chow	0.00816748	TRUE
889880896479866881	https://pbs.twimg.cor	1	French_bulldog	0.377417	TRUE	Labrador_retriever	0.151317	TRUE	muzzle	0.0829811	FALSE
889665388333682689	https://pbs.twimg.cor	1	Pembroke	0.966327	TRUE	Cardigan	0.0273557	TRUE	basenji	0.00463323	TRUE
889638837579907072	https://pbs.twimg.cor	1	French_bulldog	0.99165	TRUE	boxer	0.00212864	TRUE	Staffordshire_bullterrier	0.00149818	TRUE
889531135344209921	https://pbs.twimg.cor	1	golden_retriever	0.953442	TRUE	Labrador_retriever	0.0138341	TRUE	redbone	0.00795775	TRUE

Tweet image prediction data

Third was "tweet_json.txt" file using Twitter API tweepy. tweet_dataf d_frame with tweet ID, retweet count, and favorite count. I fetched each tweet one by one and write into a file called **tweet_json.txt** and put the code in twitter_api.py. Then I created the data frame by reading the file **tweet_json.txt.**

2. Assessing

Assess data frame visually and programmatically and results into Quality and Tidiness issues.

For Assessing the data frames I used data_f.name, data_f.rating_denominator.value_counts() , data_f.rating_numerator.value_counts(), .isnull() function , image_dataf.info() and tweet_dataf.info().

Then by looking deep into it, I summed up the following quality & tidiness issues.

Quality and Tidiness Issues-

A retweet is not an original tweet instead it a tweet again posted by other user.

QUALITY ISSUES

data_f

1. As we can see by retweeted_status_id, there is 181 retweets
2. Missing data in expanded_urls (Tweets without images) for example For row number 30,55,64 etc
3. All images aren't dog images.

4. As we can see by rating_numerator and rating_denomenator that all ratings are not correct
5. As we can see by data_f.name that there are some incorrect dog names (a, an, the etc.) For example for row number 2327, 2333, 2334, 2335 etc
6. As we can see by data_f.name that there are some missing values in dog names as "None".
7. Inaccurate datatype for tweet_id(int64) and timestamp(object)
8. Very Inefficient to understand the source from source column
9. Dog stage type --> categorical

image_dataf

10. There are some missing records. image_dataf.shape Showing 2075 data instead of 2356 which means there can be retweets,None, replies etc.
11. Some images are not dogs for example row number 29 in image_dataf is a fish.
12. Lowercase breed_names in p1, p2, p3 and underscore is used instead of space for example row number 0,1,2 etc

tweet_dataf

14. rename id to tweet_id so can merge all the data frame.

TIDINESS ISSUES

15. Delete or drop insignificant columns for example **retweeted_status_id**,**retweeted_status_user_id** etc.
16. Merge all three data frames.
17. Combine dog stages into one column that is doggo, floofer, pupper, and puppo.
18. Combine rating_numerator and rating_denominator into one_column.

I took me so much time to put down these issues.

3. Cleaning

- I obtained a copy of each of three data frame to work with, by this techniques I can safely work for analyzing and cleaning.
- Next step was cleaning the data frames and for each action I defined the problem then solved it using code and then later testes it.
- After that I merged all three data frame into clean_data_f.
- And solved the issues mentioned above in reasonable order.
- I did cleaning , re-extracting and renaming the ratings, dog stages, and cleaning the tweets with the non-dog images such as fish and human etc.

- Finally I saved my clean_data_f to **twitter_archive_master.csv** and then analyzed using bar and graphs.

The project was very abstract as I want to create that image-predictions.tsv file using NN by myself. I will do it later.