Untitled

July 15, 2021

0.1 Summary of the Data Cleaning Performed above :

0.1.1 Quality Issues :

0.1.2 tweet_json.txt

1. Removing the Contributors Column

2. Renaming id to tweet_id

0.1.3 image-predictions.tsv

1. Renaming following columns p1':'dog_breed_prediction1','p2':'dog_breed_prediction2','p3':'dog_breed_pr

2. Dropping following columns : 'jpg_url', 'img_num', 'p1_conf', 'p1_dog','p2_conf', 'p2_dog','p3_conf', 'p3_dog'

0.1.4 twitter-archive-enhanced.csv

1. Dropping following columns which we will not use for our investigation in_reply_to_status_id", "in_reply_to_user_id", "source","retweeted_status_id", "retweeted_status_user_id", "retweeted_status_timestamp", "expanded_urls"

2. Timestamp has both date and time in a single column so need to split it to two columns Time and Date

3. Time has +0000 added to it so need to ignore that value

4. Date is in string datatype so convert it to datetime format

5. Dropping original columns doggo, floofer, pupper, puppo as already we have now Dog_Lingo

6. On Digging deeper it was found that many of the dogs where name is mentioned as "A" is not a dog its a giraffe/carrot/bow etc Read the text of such rows and get the tweet_id and remove these entries using tweet id

7. Also In many of columns where name has "A/AN/THE" . Text has names of the dog like : "This is a Helvetica Listerine named Rufus."Thus searched such rows and got the valid names from text column

8. In many of the entries in the text column its mentioned these are not valid dog entries "'We only rate dogs'" so removed such entries

9. There are few entries where the numerator rating is in decimal thus find all rows with numerator in decimals using regex, extract numerator & replace the values in original

10 There are few non null values for retweet status id which means these are the retweets thus have removed such entries

0.1.5 Tidiness Issues

1. doggo, floofer, pupper, puppo columns in twitter_archive_enhanced.csv are combined into a single column as this is one variable that identify the stage of dog.

2. Information about one type of observational unit (tweets) is spread across three different dataframes. Thus three dataframes are merged as they are part of the same observational unit.