Wrangle and Analyze Project Report

Sawsan Alshaghel

Outline

Introduction

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. WeRateDogs has over 4 million followers and has received international media coverage.

Gathering Data

The data was gathered in various ways; first from CSV file, second from TSV file, and last from API.

1.Twitter Archive File ()

This file was given by Udacity, It have 2356 entries and 17 columns, this file has information about basic tweetes properties :

tweet_idi,n_reply_to_status_id,in_reply_to_user_id,timestamp,source,text,retweeted_status_id,retweeted_status_user_id,retweeted_status_timestamp,expanded_urls,rating_numerator,rating_denominator,name,doggo,floofer,pupper,puppo

2.Image Prediction File

The tweet image predictions, i.e., what breed of dog (or another object, animal, etc.) is present in each tweet according to a neural network.

This file contains 2075 entries and 12 columns

tweet_id ,jpg_url ,img_num ,p1 , p1_conf , p1_dog , p2 , p2_conf , p2_dog, p3, p3_conf, p3_dog ,

Twitter API — JSON File

By using the tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data using Python's tweepy library.

The JSON Object consists of

{

- "created_at",
- "id",
- "id_str",
- "full_text,
- "truncated",
- "display_text_range": [,],
- "entities": {

"hashtags": [],"symbols": [],"user_mentions": [],"urls": [],

"media": [{

"id","id_str","indices":
[,],"media_url","media_url_https","url","display_url","expanded_url","type",

"sizes":{

"large": {"w","h","resize"},

"thumb": {"w","h","resize"},

"small": {"w","h","resize"},

"medium": {"w","h","resize"}}}]},

- "extended_entities": {

"media": [{

"id","id_str","indices": [,],"media_url","media_url_https”,"url": ,"display_url”,"expanded_url","type","sizes":{"large":{"w","h","resize"},"thumb":{"w","h","resize"},"small":{"w","h","resize"},"medium": {"w","h","resize"}}}]},

- "source",
- "in_reply_to_status_id",
- "in_reply_to_status_id_str",
- "in_reply_to_user_id",
- "in_reply_to_user_id_str",
- "in_reply_to_screen_name",
- "user": {

"id","id_str","name","screen_name","location","description","url",

"entities": {

"url":{"urls": [{"url ","expanded_url ","display_url ","indices": [,]}]},"description": {"urls": []}},"protected","followers_count","friends_count","listed_count","created_at","favourites_count","utc_offset,"time_zone","geo_enabled","verified","statuses_count","lang","contributors_enabled","is_translator","is_translation_enabled","profile_background_color","profile_background_image_url","profile_background_image_url_https","profile_background_tile","profile_image_url","profile_image_url_https”,"profile_banner_url”,"profile_link_color","profile_sidebar_border_color","profile_sidebar_fill_color","profile_text_color","profile_use_background_image","has_extended_profile","default_profile","default_profile_image","following","follow_request_sent","notifications","translator_type"},

- "geo",
- "coordinates",
- "place": null,
- "contributors",
- "is_quote_status",
- "retweet_count",
- "favorite_count",
- "favorited",
- "retweeted",
- "possibly_sensitive",
- "possibly_sensitive_appealable",
- "lang"

}

Assessing Data

Assessing data means has to assess issues of tow types:

Quality: content issues. Low-quality data is also known as dirty data. The Data Quality Dimensions are Completeness, Validity, Accuracy, and Consistency

Tidiness: issues with a structure that prevents easy analysis. Untidy data is also known as messy data. Tidy data requirements:

Each variable forms a column.

Each observation forms a row.

Each type of observational unit forms a table.

I configured these problems:

Quality issues

- **twitter-archive-enhanced**

1-Invalid data: source column has html tags

2-Invalid data: Timestamp is datetime datatype

3-missing data: /2278/ NAN values in (in_reply_to_status_id, in_reply_to_user_id) columns and /2175/ NAN values in (retweeted_status_id,retweeted_status_user_id, retweeted_status_timestamp) cloumns.

4-Unnecessary data (retweeted_status_id,retweeted_status_user_id,retweeted_status_timestamp,in_reply_to_status_id, in_reply_to_user_id)columns

5- Invalid data:/59/ NAN values in (expanded_urls) column.

6- Unnecessary data: There are 137 duplicated values and they are retweeted tweets.

7-invalid data: there are errors in dogs name (745 'none',55 'a',7 'an' ,8 'the')

8-Inaccurate data: denominator values(50,20,80,11,0,2,7,15,16,40,70,90,110,120,130,150,170) in rating_denominator column are not equal to 10

- **image-predictions**

9- Unnecessary data: 66 duplicates jpg_url.

- **Twitter API**

10-Missing data: (contributors, coordinates, geo) columns are empty, place columns have a one not null value.

11-Invalid data: source column has HTML tags

12- column id doesn't express tweet_id

Tidiness Issues

- **twitter-archive-enhanced**

1-columns (Doggo, Floofer, Pupper, Puppo) are type of dog.

- **image-predictions**

2- image-predictions.tsv , twitter-archive-enhanced, and tweet JSON , they are one table.

Cleaning Data

Cleaning means acting on the assessments we made to improve quality and tidiness.

Improving quality doesn't mean changing the data to make it say something different — that's data fraud. Quality improvement means Correcting when inaccurate, Removing when irrelevant, and Replacing when missing.

Similarly, improving tidiness means transforming the dataset so that each variable is a column, each observation is a row, and each type of observational unit is a table.

I cleaned through these steps:

Quality

- **twitter-archive-enhanced**

1- Delete html tags in source column

2- Convert Timestamp type to datatime

3- drop (retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, in_reply_to_status_id, in_reply_to_user_id) columns

4-drop null values in expanded_urls

5-drop duplicated values in expanded_urls (delete rows which text column begins with "RT @dog_rates:" (retweeted))

6- replace dog names ('none', 'a', 'an' , 'the') with NAN

7- correct error values in rating_denominator if exist:

- replace 24/7 with nan
- replace 4/20 with 13/10
- replace 1/2 with 9 10
- replace 7/11 with 10/10
- replace 50/50 with 11/10

- **image-predictions**

8- delete 66 duplicates jpg_url.

- **Twitter API**

1- delete(contributors, coordinates, geo, place) columns

2- Delete html tags in source column

3- rename id = tweet_id

Tidiness

- **twitter-archive-enhanced**

1- make one column type instead of (Doggo, Floofer, Pupper, Puppo) columns

- **image-predictions**

2- Merge image-predictions with twitter-archive-enhanced, and tweet JSON , in one table.

Conclusion:

There are many quality issues, there are many columns with missing values especially in tweet JSON file, the columns line stage of dog and name has many missing values. We could extract more available values like whom users have posted the most popular photos and what is the most stage of dogs that people prefer.