Wrangling WeRateDogs Datasets

By Ahmed ElSabbagh

WeRateDogs is a strange fun page on Twitter which rates dogs on completely baseless and fun format, supposedly every dog is rated out of 10, the scores however can reach up to 1776 for reasons that will be cleared.
The page also gives a “Dogtionary” classification that describes how old and lovable the dog is, which could be random as the classifications can be mixed together. We will dive into the wrangling process I have taken to wrangle this data.

Wrangling steps

In my effort to wrangle the dataset, I had to follow these seemingly simple steps, which will be discussed one at a time.

1. Gather the data
2. Assess the data
3. Clean the data

Gathering the data:

I had to gather the data from three main sources.

1. The original archive provided by WeRateDogs.
2. Image classification results for each dog that predicts their type or if they are dogs.
3. Some unmentioned data from Twitter API.

Archives

The original archive was provided in CSV format containing 2356 data tweets, the were already wrangler with text analysis techniques and provided for me to use.

Image Model

The image prediction model was provided in TSV format containing 2075 tweet photo predictions, however it had to be downloaded programmatically using “Requests” library from python.

Twitter API

There were some relevant data that was not in the original archives and thus had to be obtained through Twitter API. The developer account was set up and security data to access the account was provided for me. The data was stored in a Python file and imported using the magic “%run” command, thus ensuring me privacy.

After looping through each tweet to obtain Tweets Counts and Favourites Counts required approximately 20 minutes for nearly 2090 tweets.

Assessing the data

The assessment for each individual dataset resulted in the following:

Archive

1. Data is untidy because of the dogtionary labels
2. Also untidy, unwanted retweets and replies (already removed before using the API)
3. Many names are None, some could be misinterpreted as an, a, the, etc.
4. Time stamp not date-time type
5. We have 5 empty columns, another has some missing values
6. Source has anchor tags on it
7. Very unusual rating numerator and denominators (After taking a look, apparently these oversized rates are given for the presence of multiple dogs at the same time)
8. 3 duplicate expanded urls
9. 2 overvalued numerators, one of them is Snoopdog rated 420/10 and another is a patriotic dog with score 1776/10.
10. Some tweets contain floats which were chopped (9.75/10 becomes 75/10)

Predictions

1. Many predictions don't indicate it's dog at all
2. Predictions dataset is smaller than the archive, even after removing retweets and replies
3. Extra columns for prediction are unnecessary, it is making the data untidy
4. Duplicate photos

Twitter API data

- API dataset smaller than archive, even after removing retweets and replies, some tweets are missing

Cleaning the data

The steps taken to clean the data were as follows:

1. I removed all replies and retweets. And the unnecessary columns related to them.
2. Formatted timestamp for archives to datetime instead of text.
3. Unified the dogtionary into one column.
4. Removed the bad names.
5. Manually removed irrational nominators, fixed wrong text mining results. And scaled nominators for group dogs to 10 scale and removed the denominator column.
6. For predictions: I removed all secondary predictions and only kept the most confidant, removed the confidence itself and changed a dog pred column to is_dog.
7. Stored each result in a separate dataframe and csv file.
8. Merged and stored all the dataframes into one and a csv file.
9. Because of removed tweets, the final result is 1961 tweet.