Wrangling Efforts: WeRateDogs Twitter Account

By: Constance Martineau

Date: July 6th, 2018

# Objective

* Cleaning and preparation of data – with a focus on text data;
* Basic initial feature extraction for further analysis and model building;

# Data Sources

The data gathered for the project came from three different sources:

* **WeRateDogs Twitter Archive Dataset:** 2356 tweets that were provided from user @dog\_rates for their own twitter account - WeRateDogs. Archive contains basic tweet data (tweet ID, timestamp, text etc.) *where ratings are included* as at August 1, 2017. Format is “.csv”.
* **Image Predictions Dataset:** Every image in the WeRateDogs Twitter Archive was classified via a neural network by an instructor at Udacity. The top three predictions, along with their p-values, are included within the dataset along with the image URL, and the image number that corresponded to the most confident prediction. Format is “.tsv”.
* **Twitter API Call:** Retweet Count and Favorite Count were gathered from API call.

# Inclusion Criteria

Inclusion criteria were developed to screen data:

1. Tweets must include image;
2. Tweets cannot be a retweet;
3. Tweets must include some sort of rating;

# Results

The final cleaning resulted in a single table:

**Twitter Archive Master**

2000 observations across 35 columns

# Data Cleaning Efforts

Multiple quality and tidiness issues were found amongst the three datasets. Details of the issues identified, and solutions developed are found in the following tables.

## Quality

|  |  |  |
| --- | --- | --- |
| TABLE | ISSUE DESCRIPTION | SOLUTION |
| Archive | Retweets are found within the dataset | Subset all rows that contain a null value within the retweeted\_status\_id column, and the drop the retweeted\_status\_id, retweeted\_status\_user\_id and retweeted\_status\_timestamp columns. |
| There are columns that need to be converted to a more useable dtype | Convert columns to appropriate dtype |
| Missing data in expanded\_urls, names, doggo – puppo columns | Drop rows that do not contain a value in the expanded\_urls column |
| Extract name from text column |
| Extract dog type from text column |
| The rating\_numerator, rating\_denominator and name columns contain some incorrect values | Extract the proper rating\_numerator from the text column |
| Extract the proper rating\_denominator from the text column |
| Extract the proper name from the text column |
| Image Predictions | There are additional characteristics, such as gender, that can be extracted from the text column | Infer gender based on rules |
| Extract basic features from text column |
| There are tweets within the archive that don’t have a corresponding image prediction observation | Dealt with as part of the merge outlined above |
| API Data | There are columns that need to be converted to a more useable dtype | Dealt with as part of dtype conversion outlined above |
|  | The archive includes tweets that have since been deleted – therefore, we won’t be able to retrieve their retweet or favourite counts | Dealt with as part of the merge outlined above |

## Tidiness

|  |  |  |
| --- | --- | --- |
| TABLE | ISSUE DESCRIPTION | SOLUTION |
| Archive | The “doggo” to “puppo” columns are observations of the same variable | Combine the doggo, puppo, floofer and pupper columns into one |
| In multiple columns, nulls are presented as “None” | Replace None with np.nan |
| The source column contains extended links. | Replace the source observations with corresponding label – Twitter for iPhone, Vine, Twitter Web Client and TweetDeck |
| Image Predictions | Dog breeds and other images are specified as part of the same variable. | Extract dog breed and make it its own variable |
| All | The data sources are currently stored in three separate dataframes | Merge all three dataframes into one using the Tweet IDs |

## Feature Extraction

|  |  |  |
| --- | --- | --- |
| TABLE | COLUMN | VARIABLES TO BE EXTRACTED |
| Archive | Text | 1. Number of words; 2. Number of characters; 3. Average word length; 4. Number of stopwords; 5. Number of hashtags; 6. Number of uppercase words; 7. Clean words from text; 8. Gender; |
| Image Predictions | P1 | 1. Dog Breed; 2. Non-Dog Images |