WeRateDog

Wrangle report

This data set contains data of 1992 WeRateDogs tweets since its beginning (November 2015) to August 2018. This data has 27 columns and it can be sliced in to 3 main categories of columns. This data is an output of 3 other data sets:

- ***General tweet data***. Columns in this category are unique tweet id, date, time, full_text, hastags and links used in the tweets, source, retweet and favorite counts, etc.;
- ***Unique WeRateDogs data.*** This data set contains rating values as well as given dog associasions (doggo", "floofer", "puppo" or "pupper");
- ***Computer Neural network algorithm data***. Computer algorithm tries to predict dog breed from the pictures of tweets. This data also contains indicators which shows if algorithm value is true or not and how 'confident' an algorithm is about its decision.

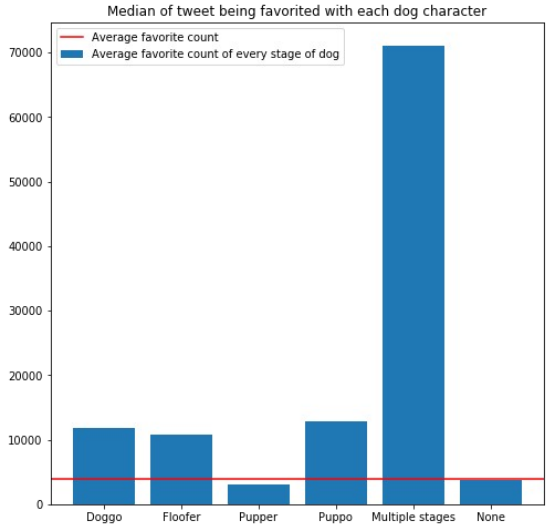
When all data was joined, these issues were found and fixed:

- Information about one type of observational unit (tweets) is spread across three different files/dataframes. These three dataframes was merged into 1 table.
- ***doggo, floofer, pupper, puppo***: columns in twitter_archive_enhanced.csv should be combined into a single column as this is one variable that identify stage of dog.
- The output contains rows that doesn't have values from neural network algorithm.
- Filter tweets to original ones (no retweets). Delete ***retweeted_status_id***, ***retweeted_status_user_id***, and r***etweeted_status_timestamp*** columns.
- ***rating denominator***: remove rows where rating denominator is lower than 10..
- ***source*** : replace values with more meaningful ones.
- ***p1_dog, p2_dog, p3_dog***: The data type of these columns should be changed to boolean values in order to interpret values correctly.
- ***timestamp***: filter timestamp column to date and time information only.
- Extract @ links to separate column for easy data filtering.
- Extract # links to separate column for easy data filtering.
- ***p1, p2, p3***: make all values uppercase in columns.
- ***text***: clean from multiple spaces and new line(\n) signs.

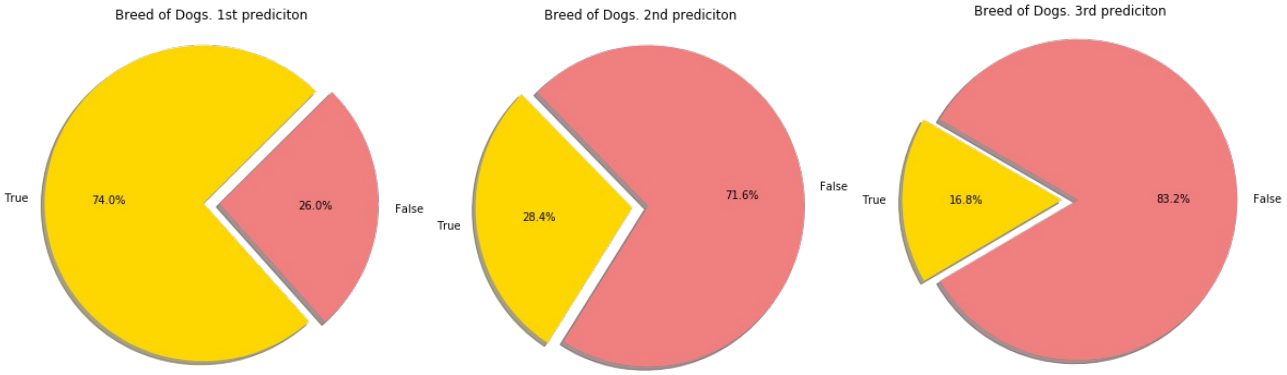
Observations that was found in this data set:

- Tweets which dog_stage associated with *"doggo", "floofer", "puppo, and multiple stages"* labels tend to have higher favorite counts. However graph might be incorrect due to big gap between tweets with and without stages of dog (Picture 1).
- With each model prediction (assuming that "guess" was false), the rate of being true is decreasing exponentially (Picture 2).
- The confidence rate of each next prediction tend to be lower. Looking to histograms it seems that values are decreasing exponentially. It was expected to have lower values on every next attempt such a decrease of confidence values is a surprise (Picture 3).

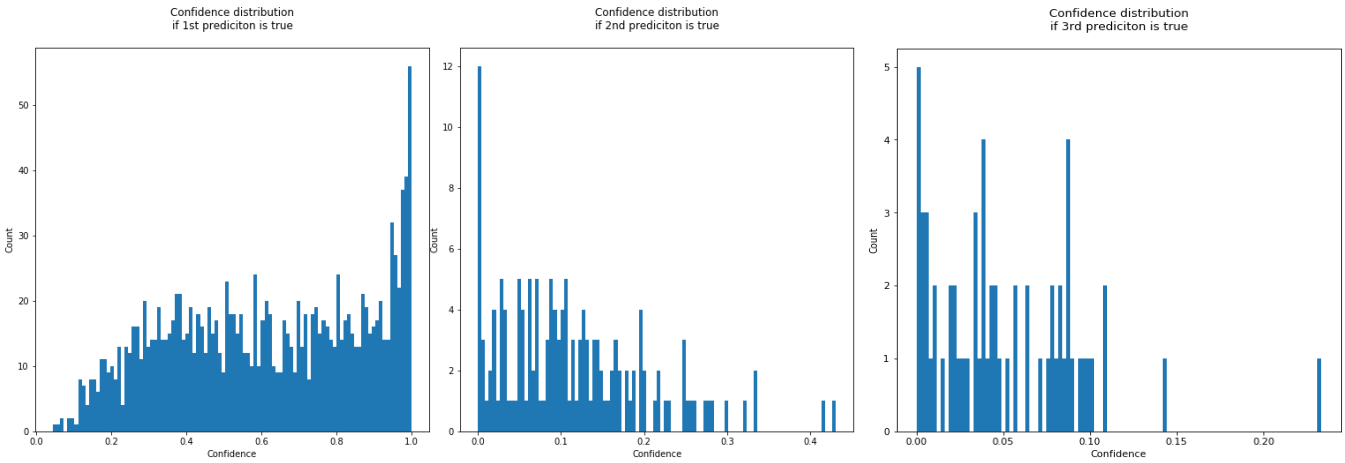
In general, neural network results are good, but isn't reliable in all cases. Research is needed in order to find the reason for some failure. Failures might happen due to bad quality of pictures. For further research, data set update with latest tweets together with neural network results is needed. Another area of research – text column. This column might have additional keywords that might be useful to extract in separate column for better data filtering.



1.Picture. Dog characters and tweets being faovrited.



2.Picture. Prediction rate on each attempt of predicting.



3.Picture. Distribution of confidence values on each true attempt
(assuming that previous one was wrong).