TWITTER DATA

Source of data

Four data sources are used in this project.

1. The enhanced twitter archive data
2. Prediction data from a neural network, retrieved from the website
3. Twitter archive data retrieved directly from twitter bij API access
4. Popular dognames retrieved from the website by parsing the webpage

Accessing and cleaning the data

Quality issues

Dognames

The dognames are extracted from the www.puppyleaks.com/popular-dog-names page. With the accessing only the names are extracted in a clean fashion.

Enhanced twitter

This archive has a lot of issues

First off, all it contains retweets. We only want original tweets. The tweets will be removed. The retweets can be found in the text which contains RT @dog_rates and the retweet_id is not Null. All these rows are dropped.

There are several columns, particular used for the retweets which we will not use. These 5 columns are removed from the dataframe.

A number of columns have a datatype which is not the best type for the data in the field.

The Timestamp field is converted to a datetime datatype.
The source field contains URL’s this is changed in a categorical datatype and the URL is replaced with a more meaningful values. [iPhone, Vine, Web, TweetDek].

According to the twitter specification the tweet id is a 64 long integer best to presented as a string. Therefore the tweet_id is converted to a string. This is also done in the other 2 datasources.
The two rating fields should be a float instead of a integer. This is changed accordingly.

The extraction of the ratings is not always correct. The extraction is redone using a more advanced extraction string ['([0-9]+\.?[0-9]*)\/([0-9]+\.?[0-9]*] This extraction searches for a number an optional dot with numbers after the dot which will lead to a float numerator. After the slash [/] the same extraction is performed for the denumerator. This could remain an int64 but for future changes a floats is considered to be better.

The values for the dognames extracted from the text are not always correct names. All values not starting with a capital letter are assumed not to be real names. With an extraction [(^[a-z][a-z]*)] string all noncapital names are selected and removed from the column. The field name is also changed in the more meaningful name dogname.
From the assessment it is not clear if all dognames are extracted from the text. A further step could be to extract all words from the text and compare it against the popular dognames. However, this is considered too be out of scope for this project.

Prediction source

The prediction data has one flaw. The prediction field names could be more descriptive. All nine column names are given a more descriptive name.

Twitter archive website

The twitter data from the twitter archive with an API call extracted only will be used for the retweet and favorite counts. Therefore, no further investigation and/or cleaning is done on this data.

Tidiness issues

Two major problems are found in the data sources.

First: the stage of the dog [puppo, floofer, doggo, pupper] is stored in separate columns. These four columns are merged in one column were the ‘None’ data is replaced with an empty string. Because more columns in one row can contain data e.g. doggo and floofer filled the merging concatenates the values in the resulting field. Each value is separated with a comma [,].

Second: the dogdate is spread over the three datasoures. All three datasource are combined in one final data source with the following result.

```
['tweet_id',  'timestamp', 'source', 'text', 'expanded_urls', 'rating_numerator',
'rating_denominator', 'dogname', 'stage', 'retweet_count', 'favorite_count',
'jpg_url', 'img_num', 'First_breed', 'First_confidence', 'First_dog',
'Second_breed', 'Second_confidence', 'Second_dog', 'Third_breed',
'Third_confidence', 'Third_dog']
```

Some further enhancements can be made. The img_num could be removed or transformed to an int type. Depending if it will be used in getting insights from the data. For now it is left in the data.

Another enhancement could be to make the three prediction sets 3 columns instead of nine. For that a column prediction_number with values [1,2,3] should be created. The columns [First, Second,Third] prediction- and dog should be arranged accordingly.

CONCLUDING

The data sources are cleaned from several quality issues and tidiness issues such that it can be used for investigating the data