wrangle_report

September 25, 2023

0.1 Reporting: wragle_report

This document summarizes the data wrangling process in a rather coarse-grained fashion. For a more detailed documentation of the wrangling process performed, refer to the actual project notebook. As part of the data assessment, we have identified and slved the following issues with the data provided:

1. Most of the ID columns (in the twitter archive as well as the prediction table) were in a wrong format, which was not suitable for analysis and further data wrangling. Hence, we converted all ID related columns to a into a common and consistent format which uses string objects instead of floating point numbers and integers.

2. We found that the time data of the tweets in the archive table was given as strings, which actually is not useful for analyzing time or dates. Hence, we converted these data into pandas datetime format.

3. We found that 181 retweets in the archive table, which are irrelevant for our analysis and would make the analysis more complicated if kept. Hence, we removed all of these retweets since our aim is to only analyze original tweets.

4. The provided twitter archive contains 2356 tweets but only 2297 which of them contain images. This was found by investigating the expanded_urls column which - if empty - indicates the abscence of an image in the tweet. Hence, about 59 tweets are irrelevant since they do not contain images and might introduce false conclusions in the analysis. These tweets were removed from the archive table.

5. Our investigation showed that the number of data in the predictions table is less than in the other tables - even when subtracting retweets and non-image containing tweets. This suggests that some of the images were discarded or overlooked by our machine learning colleagues. From this it can be concluded that only tweets present in all three tables should be considered in the analysis since otherwise the analysis becomes difficult, error prone and maybe even wrong. We tackled this issue by performing an inner join based on the tweet ID between these two tables (archive and prediction tables) to ensure that we have a common baseline of tweets.

6. The vast majority of tweets do not contain a categorization into any of the dog type categories (only 380 rows have a categorization into doggo, pupper, etc.) but only states “None” which is a validity issue. Furthermore, this set of data also presents a tidiness issue since each property should only form one column. We resolved those issues by melting the redundant columns into a single categorical column, i.e. also strings were casted into categories in the melting process.

7. We have identified further tidiness issues based on the division of the common data into multiple tables. Hence, we merged all three tables provided into a single master table to ease the analysis and to avoid repetitive join commands which might even become computationally expensive once our database grows.

[ ]: