MEMO: DATA STRUCTURE CHANGES

TO: Jane Watson, VP Social Media
FROM: Rodrigo Tiscareno, Data Analytics
SUBJECT: Data Structural Changes

Dear Ms. Watson,

I am writing to give you an update as to how we collect, organize, and structure our data effectively as of the next quarter. From the previous data analysis quarter, our data team have observed and realized some issues with the structure of the data currently being collected that we'd like to share.

From the Excel Sheet you continuously update, we have decided to make the dog stages a single column to represent a single variable. Previously, we had a column for each dog stage in which the social media team would write the stage down for each field it corresponds to and "None" for every other stage the dog did not correspond to. From now, having a single column to describe the dog stages is preferred to having a single categorically-split column. In addition, for the names of the dog, there were some values that did not contain a realistic name. We found values such as "a", "an", or "the." This may be an error in how the data from the tweet is being processed from our end. However, we do advise to please leave the name and dog_stage fields empty if they are not provided instead of writing "None," as text. Our software will account for the null values automatically.

There may be an issue with the machine learning algorithm when it comes to extracting the rating off of a tweet. We saw some unrealistic ratings and unintentional ratings. For instance, our software picked up the expression 24/7 by rating the dog image as 3.4. I would highly advise checking if the system processed each rating if it's not explicitly clear in the tweet itself to avoid outliers in the final dataset. We are also converting the rating into a single column in order to have the variables represented by a single column - not two.

There were other, minor data wrangling efforts to keep a note of. In terms of type conversions we have set the tweet_id and timestamp to the string and datetime data types respectively. This is also reflected in the default data type in the master Excel. Also in the Excel file, I've added John from Data Science's machine learning results as well as a live favorite and retweet counter to keep all the data in one place. No action is required for these structural changes. The favorite and retweet counter that is automatically extracted from an API is defaulty collected as a string, but not to worry - this has been corrected by our data team as well.

Thank you for your efforts in keeping the data we collect as clean as possible!

Regards,

Rodrigo Tiscareno