Analysis of @WeRateDogs Twitter Account (Udacity Project) – Report

By Manuel Cabrera

This report is a summary of the most relevant insights on the analysis performed on the @WeRateDogs Twitter account data.

# Introduction

@WeRateDogs is a Twitter account that makes humorous comments on pictures and videos of dogs using common internet lingo. He also gives a ranking to the dog based on how a "good boy" the dog is.

The purpose of this project is to gather, assess, clean, and analyse data obtained from Twitter user @WeRateDogs and provide any significant insights.

## Datasets and Gathering

There were three main data sets used:

* The Enhanced Twitter Archive – provided by Udacity, it has been enhanced by programmatically extracting variables from the body of the tweet (rating numerator, rating denominator, etc…)
* Twitter API (Likes/Favourites + Retweets) – extracted from Twitter’s API, mainly for the likes/favourites and retweets for each of the tweets.
* Image Prediction file – programmatically extracted from Udacity servers, this file is the result of using a neural network on the media (pictures/videos of a dog) from each tweet and identifying the dog breed for each tweet.

Assessment and Cleaning

Database had to be assessed and cleaned prior any further analysis. Main issues were found using programmatical and visual inspections, some of the issues were:

* Variables having an incorrect category.
* Invalid variables.
* Multiple variables combined into a single variable.
* Unnecessary information.
* Single variable split into multiple variables.
* Tables having different observational units.

Analysis and insights

The main questions raised from this project were:

* What is the user's most used platform/device to tweet from?
* What is the best time to tweet according to previous user engagement?
* Is there a difference between like and retweet engagement?
* Does the user have a preference to certain dog breeds (According to rating)?
* Does the user have a preference when using the internet dog names?
* Does the user continuously rate dogs throughout the year?

Using exploratory and explanatory visualisation, it was possible to answer the proposed questions for this project:

* What is the user's most used platform/device to tweet from?

@WeRateDogs user posts most of its tweets from the Twitter app for iPhone(94.3%), followed by vine (3.9%) (vine was a social app which as of 2019 was discontinued).

Chart

Description automatically generated

* What is the best time to tweet according to previous user engagement?

User engagement can be categorised as likes/favourites and retweets. Surprisingly, the preferred engagement time for these is slightly different. For better like/favourite engagement, it is recommended to tweet between 15:00 and 16:00. The engagement throughout the evening and late night seems to be consistent with a big drop at 01:00

From an initial inspection one could assume that tweeting at 6:00 would ensure the highest user interaction. However, when considering the error bar at 6:00am it is quite likely that the amount of observations (tweets) done at this time were quite low, hence the high error.

Therefore, it would seem that the best time to tweet would be at 15:00 or 16:00, mainly due to their number of likes and low error bar.

The months in which there is the most retweets would seemn to be starting June up until October, there seems to be a drastic drop of retweets in November.

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

* Is there a difference between like and retweet engagement?
* Does the user have a preference to certain dog breeds (According to rating)?
* Does the user have a preference when using the internet dog names?
* Does the user continuously rate dogs throughout the year?

# Timing of Tweets and User Interaction

Tweets made by @WeRateDogs seem to have a higher level of interaction (favorite and retweets) with other users at 00:00, 15:00 and 16:00.

* Tweets made at 06:00 were ignored due to their high error bar.

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

# User Preferred Method or App to perform tweets

Most of the tweets posted from the @WeRateDogs account were done via the iPhone Twitter App.

Chart, waterfall chart

Description automatically generated

# User Preferred Dog-Nickname

In general, the user has a preference for the nickname “pupper” when referring to dogs.

Chart, bar chart

Description automatically generated

However, when replying to tweets the user uses does not use “floofer” and increases the usage of “doggo”, but “pupper” would still be the preferred nickname.

Chart, bar chart

Description automatically generated

# Most common Breeds in tweets

Most of the tweets and interaction by the user are with Retrievers, mainly Golden and secondly Labradors. However, the breeds that rack up the highest user interaction are the French Bulldogs and the Cocker Spaniels.

Chart, bar chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

# Calculated Rating

The purpose of this account is to constantly provide ratings to images and/or videos of dogs. However, the rating system used by the user seems to be flawed since most of his scores are well beyond the calculated rating (numerator / denominator ratings) of 1.

The reason the user does this is to denote that all dogs are “good boys” regardless of what they do or look.

Chart, bar chart, histogram

Description automatically generated

However, it is important to note that it seems that the user gives even higher ratings when replying than when he normally tweets.

Chart, bar chart

Description automatically generated

# Pure breeds and User Interaction in replies

For this insight, several assumptions were made, and further investigation is required before taking any actions. One of these assumptions is that when the p1 confidence level is low, this is because the dog breed is not easily identifiable and could be a mixed breed.

In replies, there seems to be a high positive relationship between the calculated rating, retweets, favorites and the confidence level of P1, meaning that when the media of the dog can be easily identified by the AI this can generate higher user interaction.

This information can be seen when looking at p1\_conf against the favorite\_count, retweet\_count and Actual\_Rating.

