What Drives *Pupularity*?

What features and entities drive engagement among popular dog-themed twitter accounts?

🐶

Julie T. Do - Tucker Standlee - Eric Zelna

Computational Concepts in HCDE, Winter 2018
Human Centered Design & Engineering
University of Washington

Table of Contents

Introduction

The emergence of social media has created new, novel forms of communication that are rapidly evolving. Twitter is a prime example of social media-era communication. Posts from users are short (280 characters and under) and contain limited media (up to four images or a video). This constraint has allowed for new forms of creativity in communication. For marketers, influencers, and other communications experts, understanding how to leverage social media and its platform-specific communication styles is crucial in effectively engaging their audience.

In order to accurately study and understand what features of a tweet drive or encourage engagement, it is worth studying specific Twitter accounts that already possess high engagement amongst users in order to identify the attributes that draws in their audience.

An example of such popular twitter accounts selected for this study are dog-themed Twitter accounts, which exist with the purpose and intent of sharing pictures, stories, and messages about dogs. These types of accounts are interesting due to their high engagement in the form of millions of followers, and they are uniquely distinct from other popular twitter accounts as they are not associated with widely recognized brands, celebrities, or distinguished identities. They do not rely on previously existing fame or recognition to drive their engagement, and have created their popularity (or *pupularity*) on their own.

To those who wish to understand what drives Twitter engagement, it is highly valuable to study these dog accounts and understand exactly how one might captivate others and reach an extensive audience purely through the “art” of creating original tweets.

Literature Review

Both academic (Hung, et. al. 2015) and business entities have conducted research in order to understand the qualities of an engaging tweet. In general, tweets with a high level of engagement tend to have the following properties (Rogers, 2014; Lee, 2014):

- includes photos and video
- includes links
- uses relevant hashtags
- is short - about 100 characters.

In a 2015 FastCo article, Kevan Lee cites research from Buddy Media (now Salesforce Marketing Cloud) claiming that 100 characters is the ideal length of a tweet. Simon Rogers writes in a 2014 blog post that, according to twitter's internal research, adding media (photos and videos) and links all "result in an impressive boost in the number of retweets."

From their studies, Twitter found that tweets containing photos boosted retweets by 35%, videos produced a 28% boost, quotes a 19% boost, the inclusion of numbers a 17% boost, and hashtags a 16% boost (Rogers, 2014). Both of these posts were written from a business standpoint and were published with the intent in helping businesses increase engagement.

As stated earlier, dog-themed twitter accounts are especially popular, and are interesting due to their high engagement despite not being associated with widely recognized brands, celebrities or distinguished identities. Dog-related communities have formed on several social media platforms, such as Dogspotting on Facebook, and members in these communities have even formed their own lingo (gutterstats.com, 2016; Boddy, 2017). Many posts from dog-themed twitter accounts have become viral (Ohlheiser, 2017).

While there has been some speculation on the driving force behind the popularity of these dog-related communities (Valdez, 2018), there has not been extensive research conducted to study the attributes that make these communities and accounts so popular. Therefore, our study will aim to analyze some of the features present in popular dog-themed twitter accounts.

Research question

"What features and entities drive engagement among popular dog-themed twitter accounts?"

Hypothesis

We hypothesize that in the context of dog-themed twitter accounts, one or more of following qualities will be the most highly correlated with tweet engagement:

- Inclusion of media
- Shorter tweet length
- Unique "doggo" lingo
- Trending hashtags
- Positive sentiment

Data Collection

To investigate our research question, tweets were collected from the timelines of two popular dog-themed twitter accounts: WeRateDogs (@dog_rates) and Thoughts of Dog (@dog_feelings). Both of these accounts are written by the same creator, post original text content with credited media sources, and have large followings (5.99M and 822K followers, respectively, as of March, 2018). WeRateDogs shares stories about dogs submitted by users in their community, and is primarily media-based, while Thoughts of Dog is written from the point of view of a dog and is primarily text-based.

The HCDE tweet timeline module was used to collect tweets from both timelines. Tweet replies and retweets were excluded in order to focus the analysis on original content. A total of 732 tweets were collected for @dog_rates and 625 tweets for @dog_feelings.

Formatting Data for Analysis

In order to simplify and expedite the analysis of our tweets, an abbreviated dictionary of tweets was created with only the features intended to be studied. Those tweet features include the following:

- **ID** (id) - the original 18 character twitter ID value for the tweet.
- **Short ID** (shortid) - a counter variable representing the sequence of the tweet in the account history in chronological order, from oldest to newest
- **Text Length** (textlength) - the number of characters in the tweet's "full_text" variable.
- **Retweet Count** (retweet_count) - the total number of retweets for the individual tweet.
- **Favorite Count** (favorite_count) - the total number of favorites for the individual tweet.
- **Hashtag Count** (hashtagcount) - the total number of hashtags found within the tweet text.
- **URL Count** (URLcount) - the total number of URL entities in the tweet.
- **Media Count** (mediacount) - the number of media entities attached to the tweet. This turned out to be a binary variable, as there is a maximum of one media entity per tweet.
- **User Mentions Count** (usermentionscount) - the number of user mentions included in the tweet.
- **Day** (day) - the day of the week on which the tweet was created.
- **Rating** (rating) - The subjective rating (0-15 out of 10) submitted by the author of the @dog_rates twitter account. Rating was not present in the @dog_feelings account.

The abbreviated dictionary also included the sentiment ratings for negative, neutral and positive (sneg, sneu, and spos respectively) and a binary value for the inclusion of top

terms for each account (“dogs”, “good”, “meet”, “pup”, “boy”, “hckin”, “human”, “fren”). The collection and identification of these top terms is detailed later in the report.
This dictionary was then formatted for export to CSV since the majority of our data visualization and correlation analysis would be performed in Microsoft Excel 2016.

Formatting the data for import into the Windows version of Microsoft Excel 2016 presented two additional challenges in the form of Excel’s max number sizes and UTF encoding for windows.

The 18 character id value from twitter proved too large for excel to process as an integer without using scientific notation, which effectively truncated the last three characters off of the ID. This is because excel conforms to the IEEE 754 standard (Microsoft, 2017) which limits the length of integers to facilitate more efficient calculations. To adapt to this limitation, twitter’s ID value was stored as a text value for reference, and we created our own “shortid” value which is a simple incrementing variable.

The UTF encoding posed a challenge for Windows machines. With the default settings, the tweet text contained many errors regarding the representation of apostrophes, new lines and emojiis.


Say hello to Bruin. Heâ€™s currently lost in exestential wintry bliss. 13/10 would not disturb https://t.co/XAhCpqeziL

Figure 1. Example of misformatted tweet text prior to UTF-8 encoding

These issues were resolved by first encoding the text with the builtin function:

```
.encode("utf-8")
```

And then cleaning the text by replacing newline characters “\n” with empty spaces:

```
text = dic['full_text'].replace('"', '""').replace('\n', ' ').replace(',', ' ')
```

Then after exporting the text as a comma-delimited .csv file, the data was imported into excel using a File Origin setting of 65001 : Unicode UTF-8. This combination of steps enabled all tweets to come through in a format that could be parsed by excel, even with tweets that included emoticons, as shown in Figure 2.


this 👏 is 👏 what 👏 progress 👏 looks 👏 like 👏 1310 httpstcoxozehvtk1x

Figure 2. Example of correctly formatted tweet text after UTF-8 encoding

Data Analysis

Our data analysis was focused on analyzing the top terms, determining the sentiment of the tweets, and then incorporating the top terms and sentiment scores into our dataset to evaluate the correlation against the chosen dependent variable: *pupularity* (favorite count).

Top Terms

To identify the most popular terms within the accounts, a script was created that would parse through the cleaned timelines and identify the words that occured the most often. In order to ensure that the most relevant terms were gathered that were the truest representations of the data, punctuation and capitalization were eliminated from the tweet text, in addition to any stopwords based on the existing NLTK collection of stopwords.

For the analysis of the terms, it was necessary to customize the list of stopwords to ensure it was taking into account the context of the twitter accounts and what was relevant to them. For instance, it was necessary to ensure that the actual names of the accounts (@dog_rates and @dog_feelings) were added to the stop words list as they naturally appeared frequently but were not necessarily accurate representations of the most common terms that users were specifically choosing to use.

There was one stopword in particular that existed within the NLTK collection that needed to be removed: good. This term, though accurately classified as a stopword in other situations, was actually quite important to these two accounts as such phrases as “good dogs” or “good boy/girl” were very common within the accounts and were important reflections of the user’s feelings towards the accounts’ content of dogs. One of the accounts’ most popular jokes and most quoted messages reads, “they’re good dogs brent.” Exclusion of stop words would reduce this message down to “dogs brent” (Ohlheiser, 2017). As shown in Figure 3, the exclusion of the word good would not be a true and accurate representation of the data, and it was important that it was included in the collection of popular terms. In the end, “good” actually appeared in the list of the top ten most frequent words and proved its importance in being excluded from the stopword list.

Figure 3. A screenshot of a tweet exchange that went viral. With the default stopwords provided in the HCDE module provided, the viral tweet would have read, "dogs brent."

The identified top terms from each account were then graphed to display their frequency within the two accounts. The top terms for @dog_rates can be seen in Figure 4 and the top terms for @dog_feelings can be seen in Figure 5.

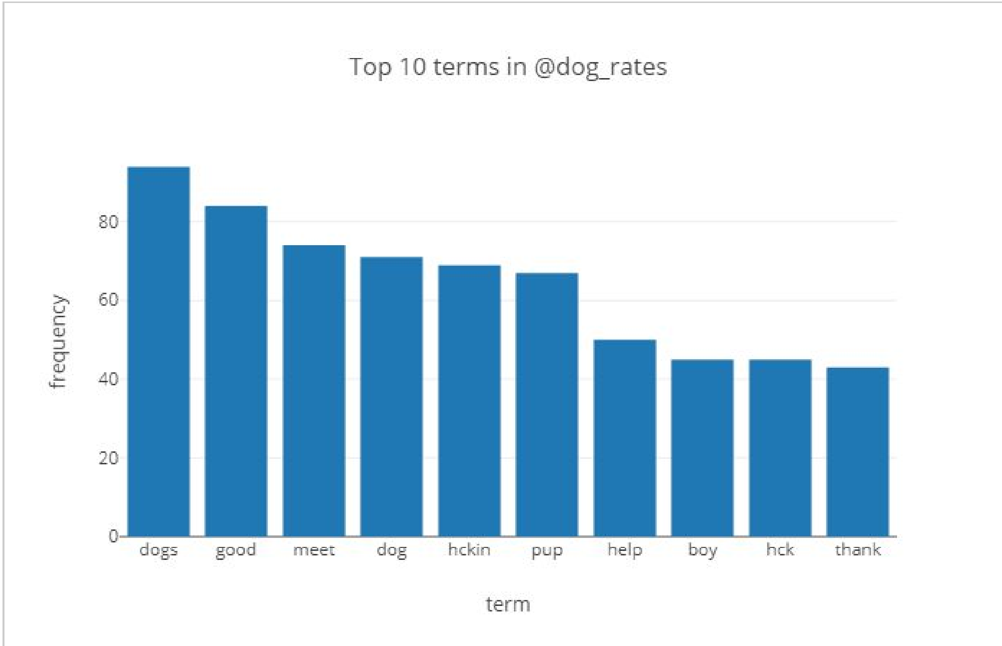


Figure 4. Most frequently occurring terms in @dog_rates.

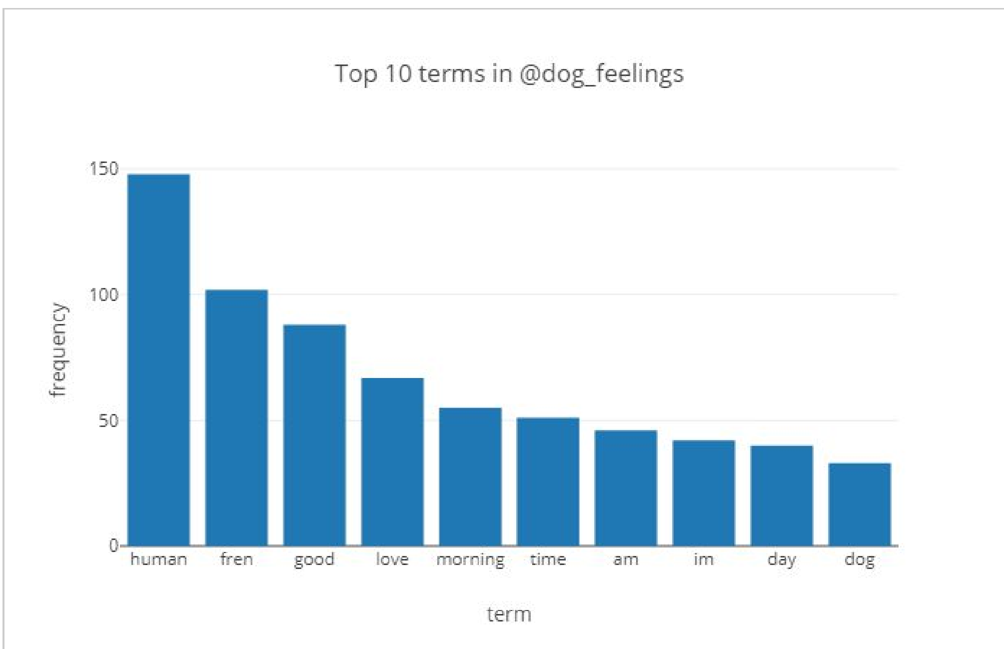


Figure 5. Most frequently occurring terms in @dog_feelings

After determining the top terms, the correlation of their usage with favorite counts was calculated. The results of these correlations will be detailed later in the report.

Top Hashtags

Another script was written to parse through the tweet timelines and extract the most common hashtags in both accounts.

After identifying the top hashtags, we discovered that @dog_rates had enough hashtags to actually result in a top ten list, while @dog_feelings only used a grand total of one tweet within the the entire timeline of tweets. It is our suspicion that the stylistic voice of the tweets in this account being from the perspective of a dog specifically resulted in the author's intended omission of the use of hashtags as it seems that the author does not feel that dogs think in hashtags. The top ten hashtags in @dog_rates are shown in Figure 6, and the lone hashtag in @dog_feelings is shown in Figure 7.

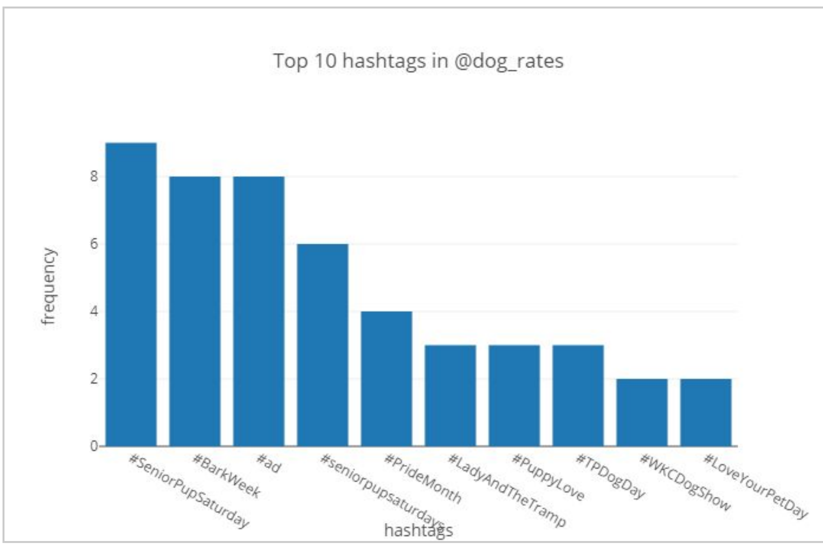


Figure 6. The most frequently occuring hashtags in @dog_rates

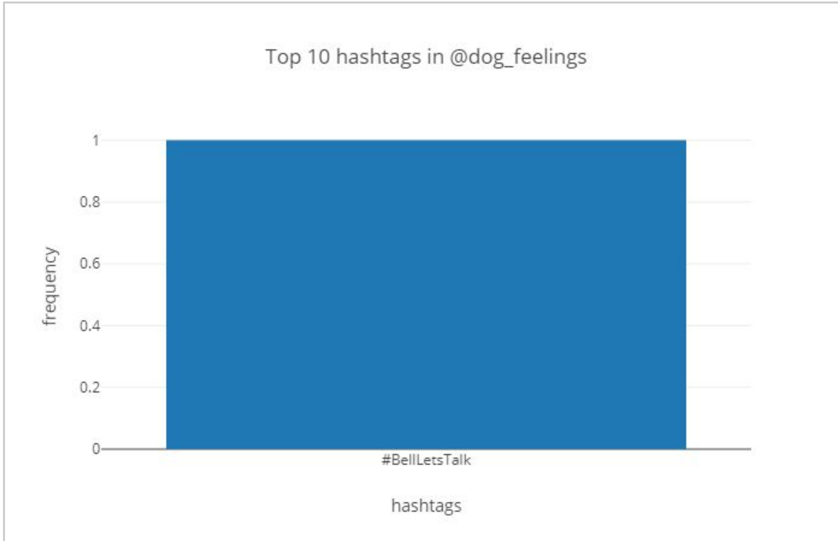


Figure 7. The lone hashtag utilized in @dog_feelings

It is evident from these graphs that hashtags are not highly utilized within either account, with @dog_rates' most popular hashtag only occurring a grand total of eight times and @dog_feelings only using the one hashtag in its entire history.

Since hashtags were so uncommon in both accounts, we decided to calculate the pupularity with the correlation between the presence of any hashtag, rather than a specific hashtag. The results of these correlations will be detailed later in the report.

Sentiment analysis

A script was written using the open-source NLTK VADER Sentiment Analysis toolkit (Hutto & Gilbert, 2014) in order to calculate the positive, neutral, negative, and compound sentiment scores for the text of each tweet.. Unlike the top terms and hashtags, the text did not need to be cleaned for the sentiment analysis. Examples of sentiment scores are shown in Figures 8 and 9.

Tweet text:
This is Benedict. He wants to thank you for this delightful urban walk. Hopes you know he loves you. 13/10 super duper good boy https://t.co/26BXueUgbs

Sentiment Score:

- Negative: 0.0
- Neutral: 0.483
- Positive: 0.517
- *Compound: 0.9652*

Figure 8. An example of a highly positive tweet.

Tweet text:
This is Oliver. You're witnessing one of his many brutal attacks. Seems to be playing with his victim. 13/10 fr*ckin frightening #BarkWeek https://t.co/WpHvrQedPb

Sentiment Score:

- Negative: 0.383
- Neutral: 0.561
- Positive: 0.056
- *Compound: -0.8885*

Figure 9. An example of a highly negative tweet.

We used the compound score for our preliminary comparisons. Using the Dash by Plotly python framework, we plotted scatterplots of the sentiment scores and favorites count for both accounts, shown in Figures 10 and 11. As expected, the tweets of both accounts seem to be generally positive.

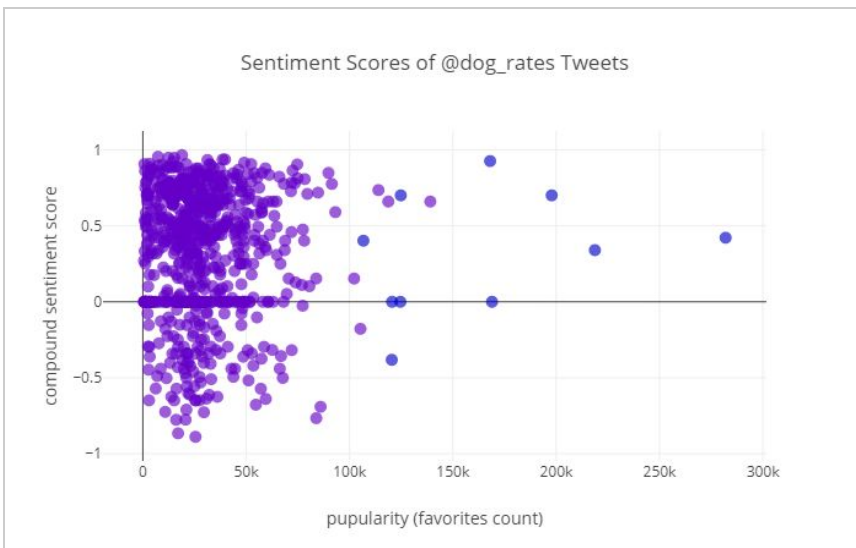


Figure 10. A scatterplot of compound sentiment scores and favorites count for @dog_rates.

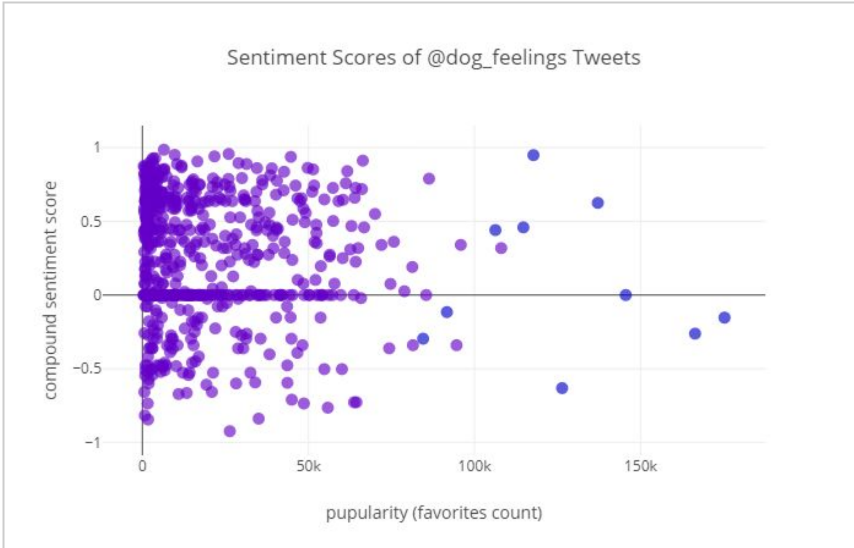
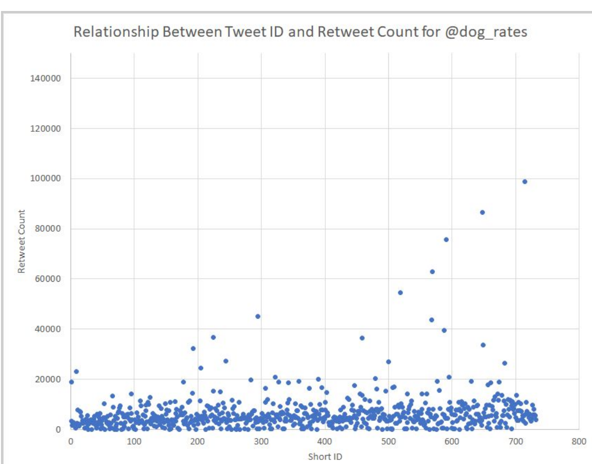
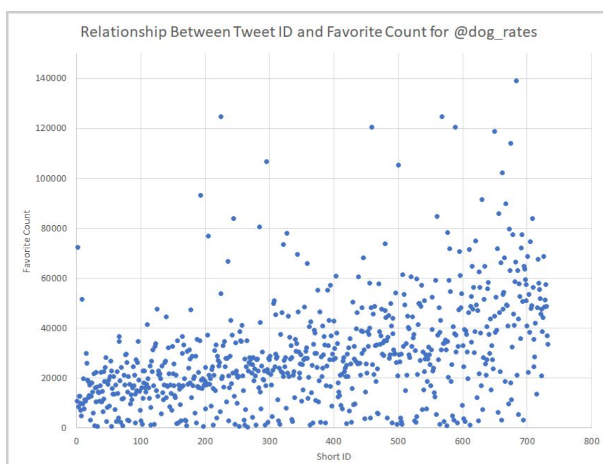


Figure 11. A scatterplot of the compound sentiment scores and favorites count for @dog_feelings.

Measuring *Pupularity*: Identifying the Dependent Variable

In order to measure the effect of different tweet attributes on engagement levels, a dependent variable needed to be identified for our correlation tests. Much of the available research (Dan, 2011) and advice from marketers (Lee, 2014) focuses on retweets as a primary measure and driver of engagement. This makes sense intuitively, since a retweet is often the most direct way of increasing the likelihood that others will see the tweet. During our preliminary analysis, we found that the strength of the relationship between retweets and the other factors we evaluated was weaker than the relationship between favorites and those same factors. You can see an example of this below in Figure 12 and Figure 13 where short ID has a visibly stronger correlation with favorite count (r=0.42) than it does with retweet count (r=0.18).





Figures 12-13. Scatterplot of favorite count over time and retweet count over time for @dog_rates

This is a bit of a surprise given that the correlation between the retweet count and the favorite count is quite strong (r=0.9) as seen below.

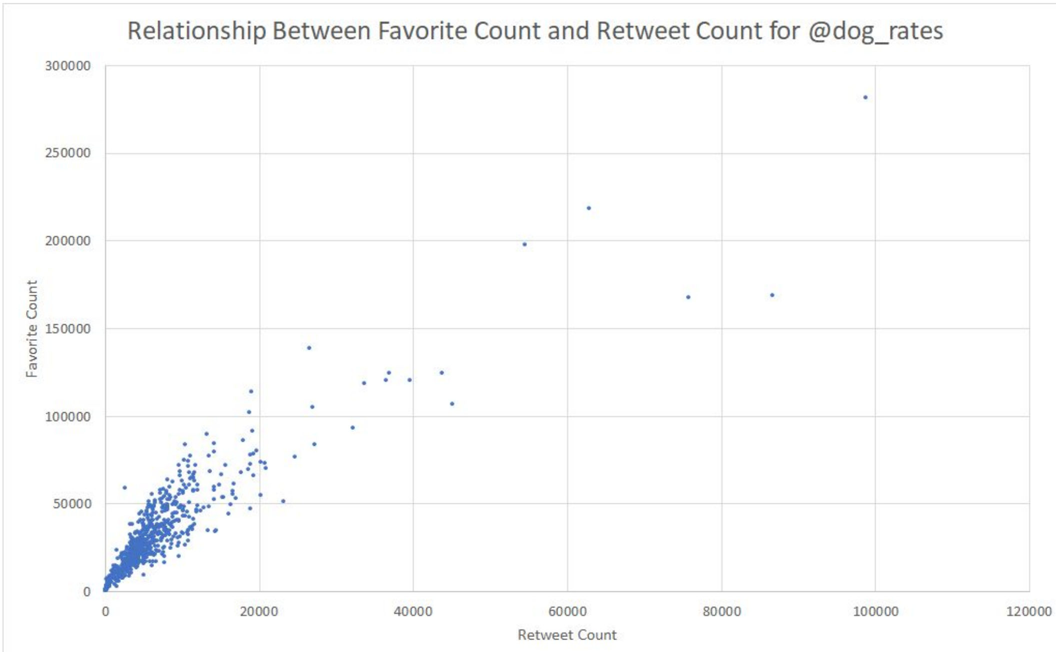


Figure 14. Scatterplot depicting the relationship between favorites and retweets for @dog_rates

Identifying the reason for this disparity is outside of the scope of this paper, however, because there was effectively no meaningful relationship between retweet count and any of the features we studied. It therefore became necessary to focus on favorite count as our metric for *pupularity,* as much stronger relationships could be observed, with the strongest correlation having a value of r=0.24.

@dog_rates	shortid	textleng	favorite	hashtag	urlco	media	userm	day	rating	sneg	sneu	spos	scomp	dogs	good	meet	pup	boy	hckin	human	fren
retweet_count	0.18	0.02	0.90	0.08	0.24	0.06	0.11	0.03	0.20	0.06	0.06	0.03	0.00	0.03	0.01	0.01	0.01	0.01	0.01	0.00	0.09
favorite_count	0.42	0.07	1.00	0.08	0.36	0.17	0.14	0.07	0.20	0.06	0.08	0.04	0.02	0.01	0.01	0.01	0.02	0.06	0.01	0.02	0.06

Table 1. comparison of correlation scores across all features for retweet count and favorite count.

For the purposes of this study, we define *pupularity* as our key metric for tweet engagement, as measured by favorite count.

Calculating Correlation

Having identified favorite count as the dependent variable for our analysis, we took an exploratory approach, and calculated the correlation coefficient of all variables against all other variables in a matrix to identify interesting relationships to investigate further.

For continuous independent variables, we used a Pearson correlation coefficient, as implemented by Excel's built-in CORREL function for our continuous variables (i.e., shortid, textlength, favorite, hashtagcount, urlcount, mediacount, usermentions, day, rating, sneg, sneu, spos, scomp).

Since the Pearson correlation coefficient is a poor model for binary variables, we used the biserial correlation coefficient for our binary independent variables (i.e., dogs, good, meet, pup, boy, hckin, human, fren) as implemented by the BCORREL Function found in the excel Realstats add-in package. Changing from the CORREL function to the BCORREL function for our binary variables produced correlation coefficients that were considerably higher.

After creating the matrix for both accounts (Appendix I), it was then easy to identify the variables which were most strongly correlated with favorite count, and then plot the raw data for those combinations to visually investigate any interesting relationships. Where those relationships exist, we then manually reviewed individual tweets to identify the potential underlying causes for those relationships.

Identifying Outliers

Initially we identified outliers by using the common method of identifying data points that were 2.5 times the interquartile range of favorite count. However when this data was plotted, it became apparent that the pronounced trend in the data meant that this method of identifying outliers was omitting several data points that deviated from the trend, and included other data points that were close to the trend as the trend increased in favorite count.

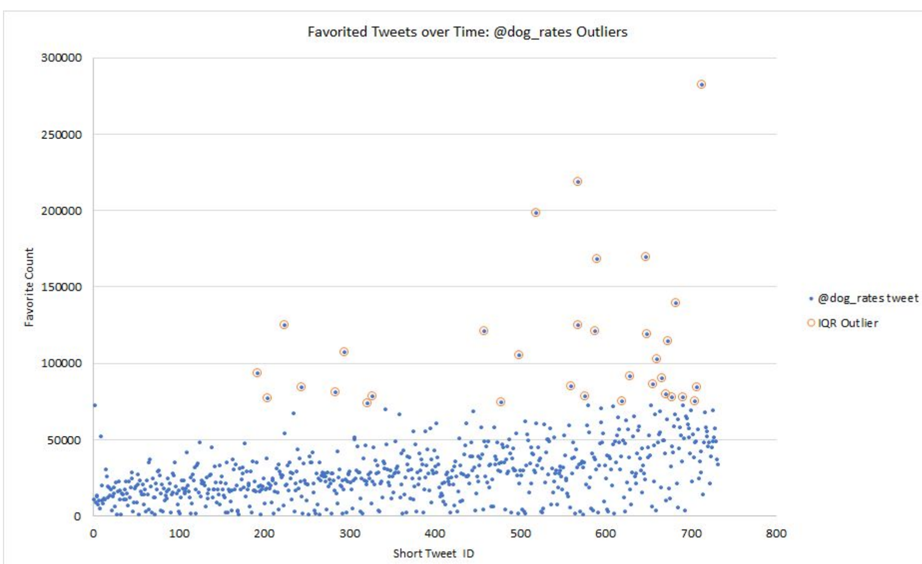


Figure 15. Outliers in the @dog_rates account identified by the interquartile range and highlighted in red.

From this observation it became apparent that we needed to use a method of identifying outliers which took into consideration the fact that our average was trending upward over time, and our dataset spanned months of data. To address this issue, we changed our method of identifying outliers from the interquartile range to be based on the residuals instead. This produced a set of outliers that extended back to the beginning of the time series graph. You can see the difference in the two methods in Figures 16 and 17.

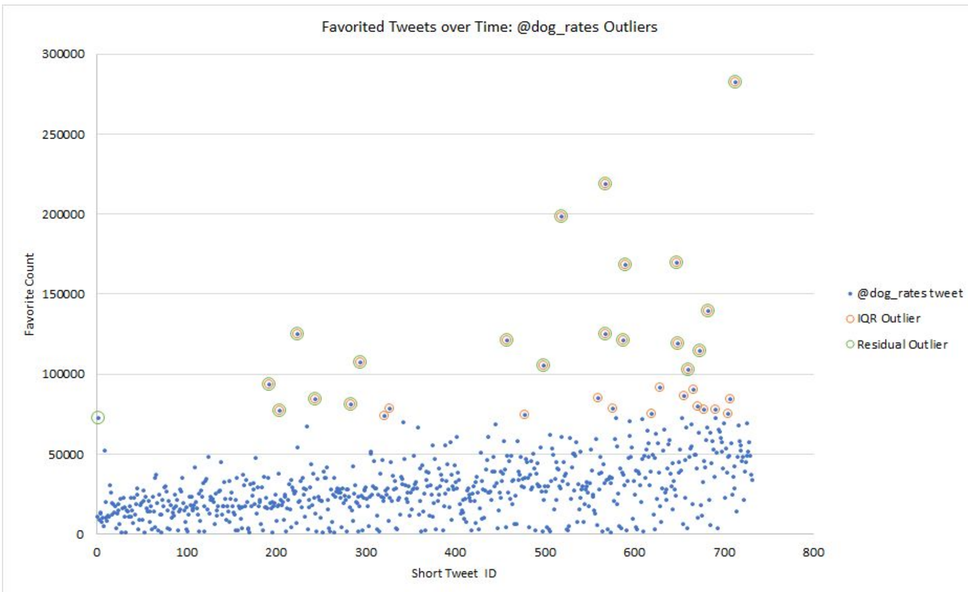


Figure 16. The change in outliers in the @dog_rates account when using residuals

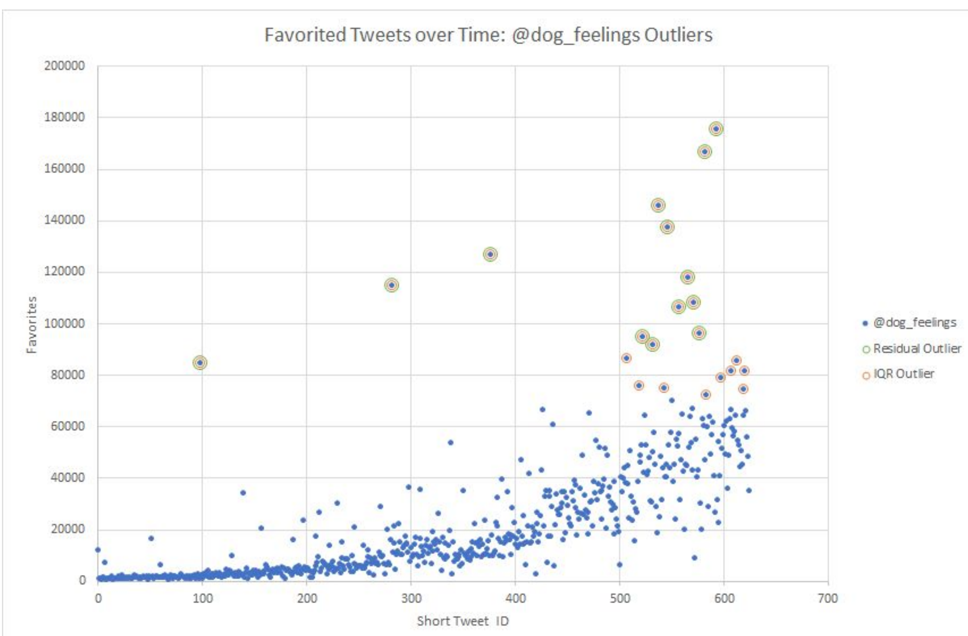


Figure 17. Outliers in @dog_feelings when using residuals

These outliers all rank in the top 10 tweets in terms of pupularity. Reference Appendix II for examples of exceptionally good dogs.

Results

By ranking the correlation values from each account, we can quickly see which attributes have the strongest correlation with favorite count. Both accounts show the ID value as having the strongest correlation, with with a correlation of r=0.42 for @dog_rates and r=0.73 for @dog_feelings respectively. For @dog_rates, the only features which had a correlation above 0.3 were “shortid”, “URLcount” and “fren”, whereas for @dog_feelings only “shortid”, “textlength”, “human” and “fren” had a correlation of greater than 0.3, with “hckin” coming closely behind at 0.21.

@dog_rates	favorite_count Multiple R	significance ($\alpha = 0.05$)		@dog_feelings	favorite_count Multiple R	significance ($\alpha = 0.05$)
shortid	0.42	p < 0.05		shortid	0.73	p < 0.05
urlcount	0.36	p < 0.05		textlength	0.54	p < 0.05
fren	0.31	p < 0.05		human	0.52	p < 0.05
rating	0.20	p < 0.05		fren	0.36	p < 0.05
mediacount	0.17	p < 0.05		hckin	0.21	p < 0.05
usermentionscount	0.14	p < 0.05		sneu	0.20	p < 0.05
boy	0.11	p < 0.05		dogs	0.19	p < 0.05
hashtagcount	0.08	p < 0.05		spos	0.18	p < 0.05
sneu	0.08	p < 0.05		urlcount	0.16	p < 0.05
textlength	0.07	p < 0.05		mediacount	0.15	p < 0.05
day	0.07	p > 0.05		good	0.12	p < 0.05
sneg	0.06	p > 0.05		pup	0.10	p < 0.05
human	0.06	p > 0.05		boy	0.10	p < 0.05
spos	0.04	p > 0.05		meet	0.07	p < 0.05
meet	0.03	p > 0.05		scomp	0.07	p > 0.05
pup	0.02	p > 0.05		sneg	0.04	p > 0.05
scomp	0.02	p > 0.05		hashtagcount	0.00	p > 0.05
dogs	0.02	p > 0.05		day	0.00	p > 0.05
hckin	0.01	p > 0.05		rating	NA	NA
good	0.01	p > 0.05		usermentionscount	NA	NA

Table 2. Independently ranked correlations of features with favorite count for both accounts

The correlation values for all features on @dog_feelings are much stronger than those on @dog_rates, which may be due to how much more tightly grouped the trends are for that account. By presenting the same data in a heatmap we can easily see how few values had any correlation at all. On the left we see each list of features ranked by their own r value, whereas on the right we can easily make comparisons by sorting the features so they match for both accounts. Rating and usermentionscount are of course both excluded from @dog_feelings because those features are unique to the @dog_rates account. Features whose correlation was statistically significant (p<0.05) are outlined with a bold black box.

By limiting our analysis to those correlations which were found to be statistically significant (p>0.05), and aligning the R values by feature across both accounts we can see that certain patterns emerge.

	@dog_rates		@dog_feelings	
	favorite_count Multiple R	significance (α = 0.05)	favorite_count Multiple R	significance (α = 0.05)
shortid	0.42	p < 0.05	0.73	p < 0.05
urlcount	0.36	p < 0.05	0.16	p < 0.05
fren	0.31	p < 0.05	0.36	p < 0.05
rating	0.20	p < 0.05	NA	NA
mediacount	0.17	p < 0.05	0.15	p < 0.05
usermentionscount	0.14	p < 0.05	NA	NA
boy	0.11	p < 0.05	0.10	p < 0.05
sneu	0.08	p < 0.05	0.20	p < 0.05
textlength	0.07	p < 0.05	0.54	p < 0.05

Table 3. Statistically significant correlated features for both accounts, sorted by @dog_rates

Of note are the very similar R values for “fren” on both accounts, and the much stronger R value for “textlength” and “human” on the dog_feelings account. By plotting the data for “textlength”, we can see this relationship emerge visually.

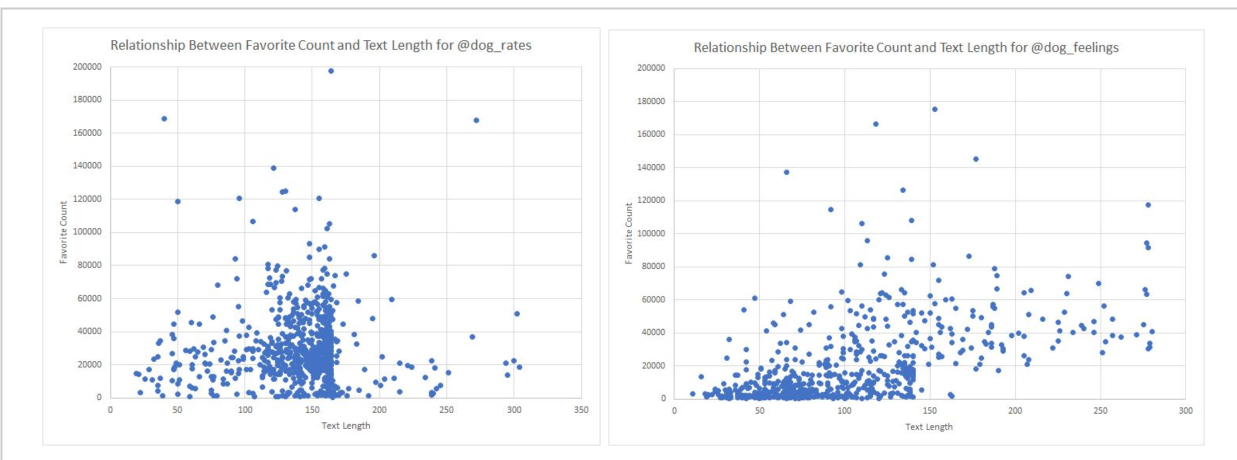


Figure 18. The relationship between favorite count and text length for @dog_rates and @dog_feelings

Despite our initial expectations that shorter tweets would drive higher engagement, it is not entirely surprising that the purely text-based @dog_feelings account would see a stronger relationship between text length and favorite count than the more media-based @dog_rates account. Also affecting this analysis is the fact that during the time span in question, the 140 character tweet length limit was raised, which creates an artificial shelf of tweets at the previous limit, which may affect the strength of the correlation values.

URL Correlation

By plotting the raw data for favorite count and URL count we can see that the majority of the tweets featuring URLs only featured 1 URL, but the data points which featured 2 URLs had an extremely low favorite count.

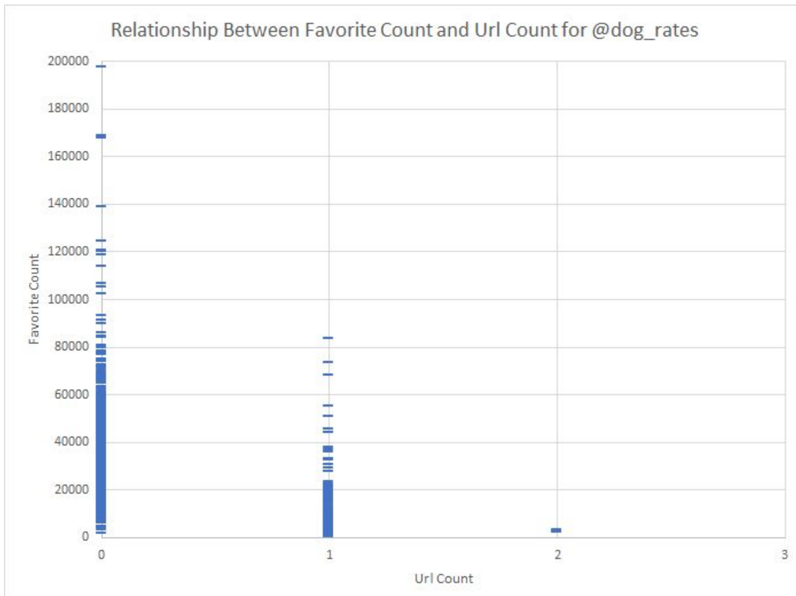


Figure 19. The relationship between favorite count and URL counts@dog_rates

The average favorite count for @dog_rates tweets with two URLs was only 2,858. Meanwhile, the average favorite count for @dog_rates tweets with one URL was 12,455, and the average favorite count for @dog_rates tweets with zero URLs was 35,421. To understand why tweets with no URLs would have a 180% higher favorite count than tweets with a URL, we only have to look at the content of the tweets that comprise these groups. The zero URL tweets are the staple of the account, and typically feature a dog picture, a story and a rating, like the two examples below which received 49,000 and 69,000 favorites respectively.

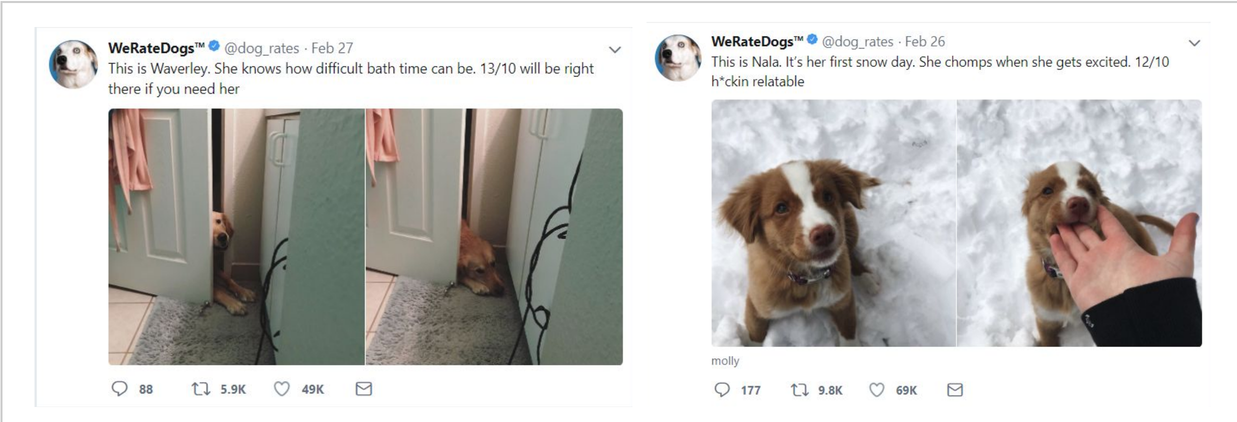


Figure 20. Examples of the most common tweet format within @dog_rates, which does not contain URLs.

Meanwhile the tweets which feature one or more URLs are typically promoting merchandise sales or otherwise promoting the site shop, like the two examples below, which received 3,200 and 3,100 favorites respectively.

Figure 21. Examples of tweets within @dog_rates that contains URLs.

Following this trend we also see much lower favorite counts for tweets which feature URLs on the @dog_feelings account, with an average 8,236 favorites for tweets with a URL and a an average 21,437 favorites for tweets with no URL.

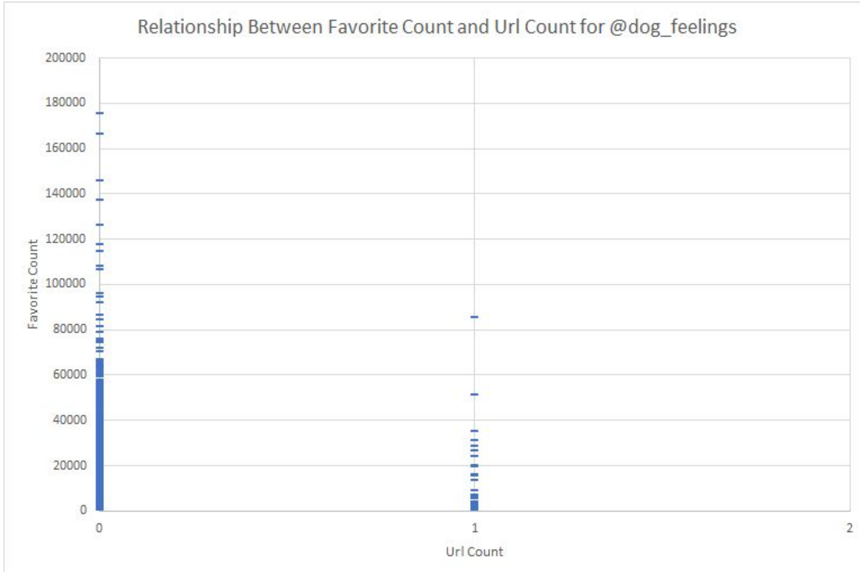


Figure 22. Relationship between favorite count and the usage of URLs in @dog_feelings

Term Usage: "Fren" and "Human" Correlations

The relationship between the usage of the word "fren" and pupularity must be interpreted with some caution, as the term only appears in the @dog_rates twitter account three times in total. Despite the 54,386 average favorites those tweets had, that is not enough data to draw definitive conclusions for the @dog_rates account. However, it is interesting that the rate of favorites for tweets using the word "fren" was almost the same across both accounts. Focusing instead on the @dog_feelings account which had 93 uses of the word "fren", the admittedly weak relationship doesn't create a stark visual contrast when plotted.

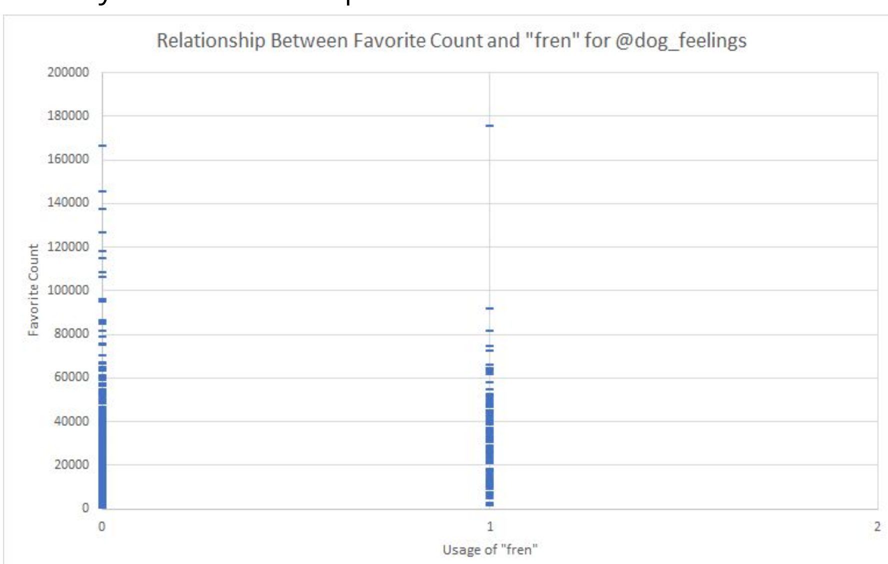


Figure 23. Relationship between favorite count and the use of the term "fren" in @dog_feelings

Despite the weak relationship, the tweets featuring the word “fren” did have a consistently higher average favorite count of 34,249 as compared to the average favorite count of 17,733 for tweets which did not feature the word “fren”. It is particularly interesting that the average would still be higher for tweets using the word “fren” when you can see above that so many of the high-favorite outlier tweets fall into the group of tweets which do not feature the word “fren”.

We see this phenomenon again below when the 135 tweets which used the term “human” had an average favorites of 37,691, whereas the 490 tweets which did not use the term “human“ had an average favorites of 15,369. This seems to suggest that the specialized language used has some impact on the engagement of users.

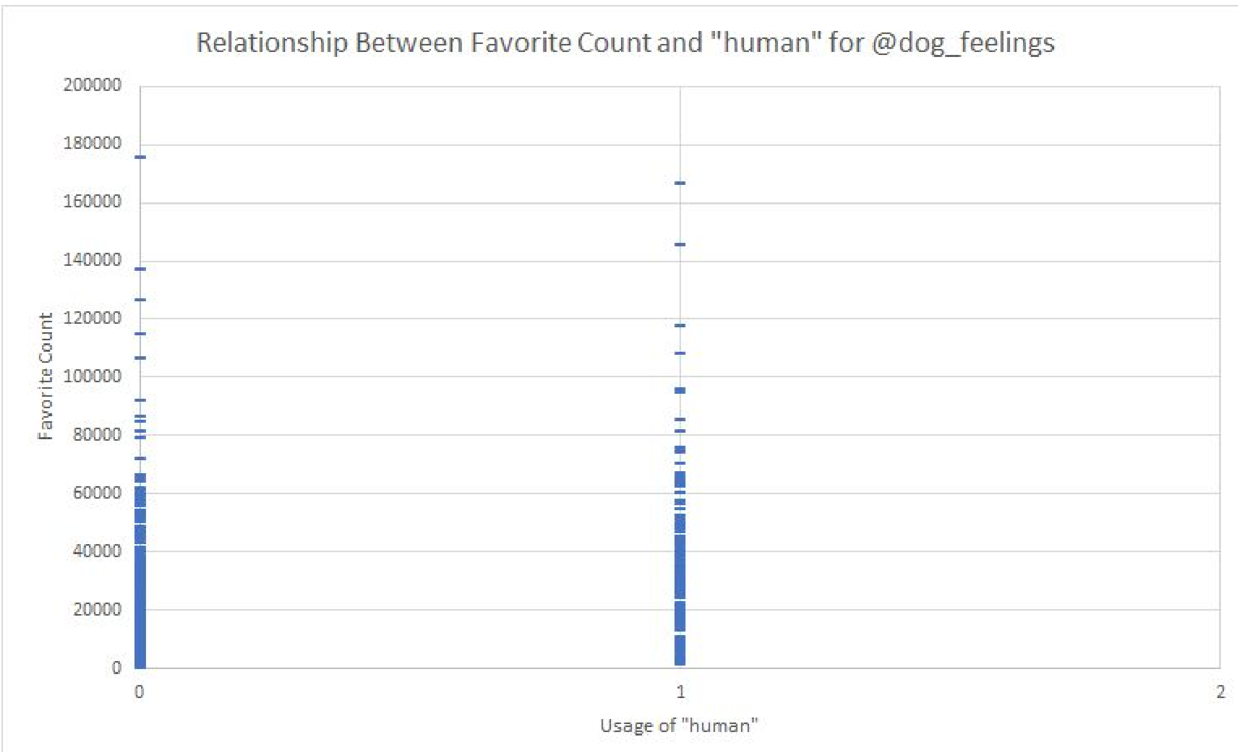


Figure 24. Relationship between favorite count and the use of the term “human” in @dog_feelings

Dog Rating Correlation

Although the correlation value was somewhat low at r=0.20, it is interesting to observe that there does appear to be some relationship between the ratings given by the @dog_rates author and the number of favorites the tweet receives, with the two 15/10 tweets receiving well above the average number of favorites.

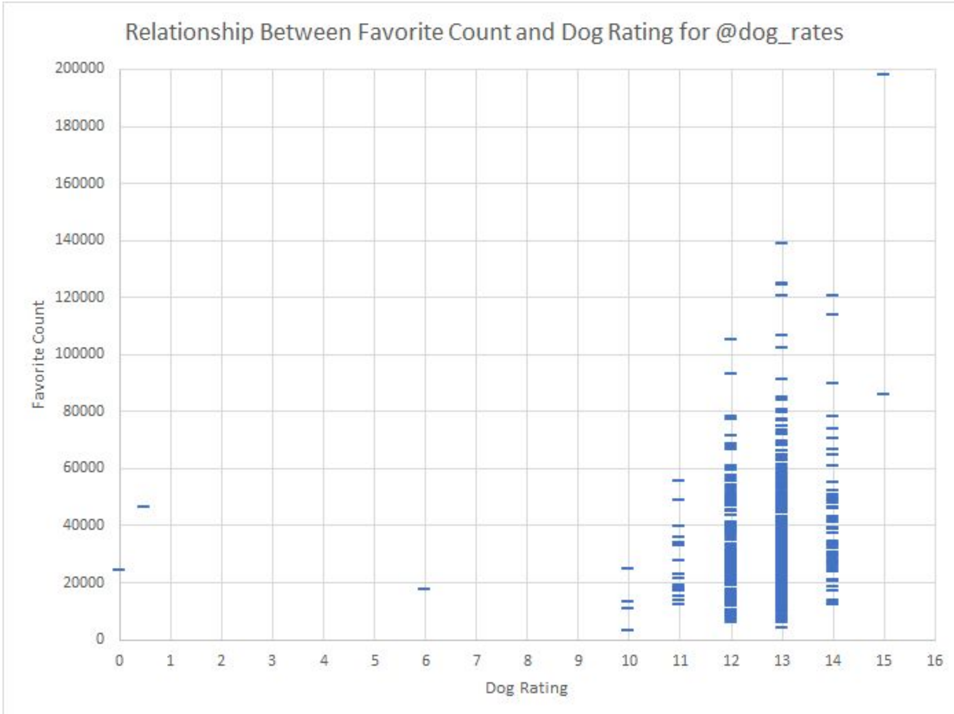


Figure 25. Relationship between favorite count and dog ratings within @dog_rates

This relationship may be weakened by the fact that so many of the tweets receive the default score of 13/10, as shown in Table 4.

Rating	Number	Percentage
10/10	5	1%
11/10	23	4%
12/10	181	31%
13/10	316	55%
14/10	50	9%
15/10	3	1%
Total	578	

Table 4. Dog ratings within @dog_rates

Despite this being a purely subjective ranking, it suggests that there is some link between the account author's assessment of the dog, and the viewer's own assessment of the dog. To ascertain whether that relationship has a causal element would require a controlled study, and is left for future research.

Media Correlation

Since the @dog_feelings account does not make use of media entities, with the exception of 2 tweets, we are instead focusing on the @dog_rates account for which 625 out of 732 tweets incorporated media entities, which you can see below in Figure 26. While we initially expected that there would be a strong correlation between the use of media entities and favorite count, the pearson correlation coefficient was only 0.17 while the biserial correlation coefficient was only 0.26, making it lower than ID, urlcount and even fren.

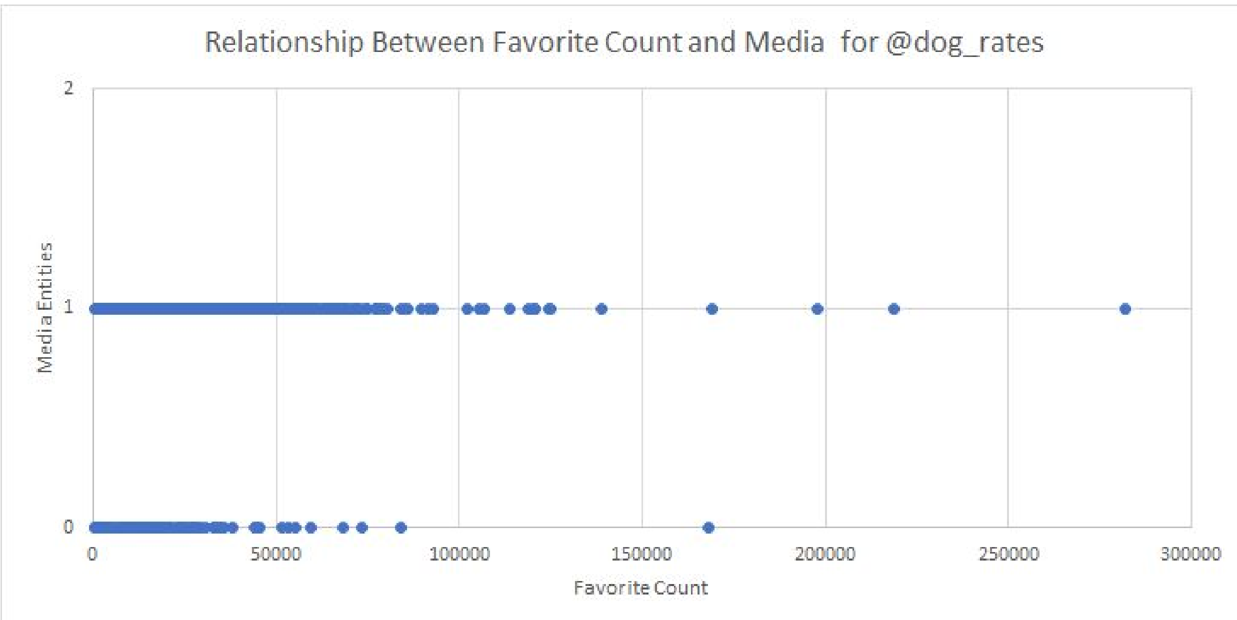


Figure 26. Media entities and favorite count for @dog_rates

Conclusions

Contrary to our initial expectations, most of the attributes we anticipated to be the highest drivers of user engagement as measured by pupularity (favorite count) did not have as strong of a correlation as we predicted.

The usage of media did not have as strong of a correlation as expected. Opposite of our initial expectations, longer tweets were more strongly correlated with engagement than shorter tweets. Hashtags, which our research asserted as a driver of engagement (Rogers, 2014; Lee, 2014), did not have a significant impact. Positive sentiment also did not have a strong correlation with engagement. However, though words like “fren” and “human” only had a weak correlation, it did validate our hypothesis that the usage of doggo lingo could help drive engagement.

Strongest Correlations

Unexpectedly, the most strongly correlated driver of engagement ended up being the unique tweet IDs that increased with account age. This suggests that the two accounts are becoming more popular over time and that account maturity is an important factor for engagement. It is worth noting that the second most strongly correlated feature was tweet length, but this was exclusive to @dog_feelings. Tweet length had a much lower correlation for @dog_rates. One additional attribute which had an unexpected correlation was URL, which was negatively correlated with favorite count, in contrast to the findings of our initial research (Rogers, 2014).

Future Work

Based on our observations of the content of these tweets, we speculate that there are many qualitative aspects of the tweets that resonate with their followers and that inspires their engagement. One example is emotional appeal. Consider the top three tweets from the account @dog_rates as seen in Figure 27 below:

Figure 27. The three most popular tweets from Twitter account @dog_rates

In these three tweets, we see content that is of vastly varying sentiment. One tells a heartwarming tale of a canine companion who helps comfort those who have suffered through a traumatic experience. One displays a cute little puppy who did not receive his Halloween pajamas. The last is a tragedy about the passing of a dog named Smiley. All three of these tweets draw an emotional response from the reader, which is not easily captured in the sentiment analysis. A deeper qualitative study of tweet contents could help determine the impact these emotional appeals have on a tweet's engagement.

While we did not find a strong correlation between the sentiment of a tweet and its pupularity, it is important to note that our sentiment analysis was fairly simple and did not account for context. For a further analysis of sentiment, we would look more closely at both the positive and negative sentiment scores and look for variance that may suggest more extreme emotional impact.

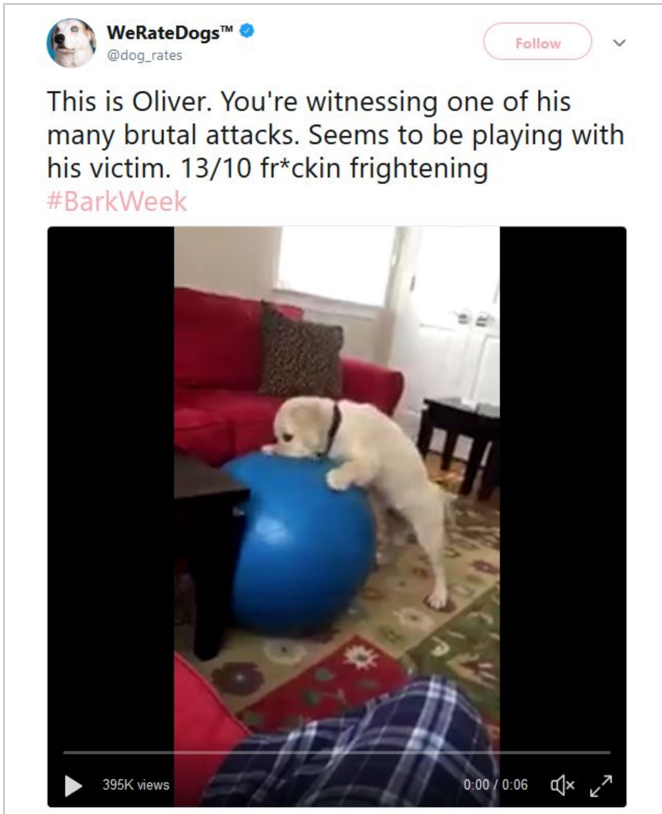


Figure 28. A screenshot of a tweet with a highly negative sentiment rating

In order to understand sentiment more deeply, we would likely need to create our own classifier that accounts for the doggo-lingo that is specific to dog-interest communities. The classifier would need to be able to "understand" the subtleties of humor and context. While the tweets with highly positive sentiment scores are fairly straightforward, the tweets with highly negative sentiment scores are trickier to analyze. As shown in Figure 28, tweets with a highly negative sentiment score have words that are seen as negative, such as "brutal" and "frightening." However, the tweet itself contains a video of a puppy playing, and the juxtaposition adds to a humor that resonates with the audience. Much like sarcasm, these subtleties in communication are difficult to pick up with a basic sentiment analysis.

Another aspect that we would like to explore is image and video analysis. Within the tweets that do have pictures or video, we are interested to see if the breed of dog or age of the dog has any impact on engagement. In addition, we are also curious to see how tweet engagement is impacted by major events. Finally, based on our findings we are curious to explore why favorites are more strongly correlated with tweet features than retweets.

Project Reflections

In addition to the findings that we collected from our research and analysis, there was much that we learned over the course of this project. Specifically, there were many aspects about the project process itself that we learned lessons about mostly as a result of obstacles faced and adjustments made.

Challenges

One of the most difficult challenges that we encountered through the course of the project occurred towards the beginning and came in the form of narrowing down our research and determining a specific question or problem to investigate. The initial question we had was what can make a tweet popular, and this question proved to be far too broad and there were countless ways to approach answering this question. In addition, identifying specific research for this question resulted in far too many resources to investigate that made discovering truly useful and applicable sources difficult. It was only after narrowing down our question and study to the specific aspect of text-based features and entities of a tweet, that we were really able to gain traction and have a focused path to pursue for the duration of the project.

Another challenge that we faced early on was being limited by our own unfamiliarity with python and its capabilities. At the start of our project, it was difficult for us to understand what was feasible in the scope of this course, and as a result, it was difficult to decide on the research methods and data analysis we would conduct on the tweets. It was only through trial and error that we learned just what was possible with the tool and therefore what methods we could utilize with our study.

During the analysis phase, we also encountered numerous challenges. Because of the broad scope of our project, we would occasionally discover a new feature to include, which required us to redo parts of our analysis. Dictionaries would have to be updated, variables named and passed into other functions, and then excel spreadsheets would be recreated, many times from scratch. Some examples of issues that drove this rework include:

1. Changing from twitter's ID value to our own short ID value.
2. Discovering that retweets have zero favorites and excluding them from the dataset.
3. Changing our method of identifying outliers from the interquartile range to residuals.
4. Incorporating sentiment score and term frequency into the correlation analysis.
5. Changing correlation values from pearson to biserial for the binary variables.

While we were not able to definitively parse out the attributes that make these dog-themed twitter accounts so appealing, we can be sure that dogs truly are good.

References

1. Hung, Y. H., Hwu, D. S., Arkenson, C., & Lee, Y. C. (2015). Designing for Retweets–A Study on Twitter Interface Design Focusing on Retweetability. *Procedia Manufacturing, 3, 5496-5503.* Retrieved March 11, 2018, from https://www.sciencedirect.com/science/article/pii/S2351978915007003
2. About your Twitter timeline. (n.d.). Retrieved March 11, 2018, from https://help.twitter.com/en/using-twitter/twitter-timeline
3. Boddy, J. (2017, April 23). Dogs Are Doggos: An Internet Language Built Around Love For The Puppers. Retrieved March 11, 2018, from https://www.npr.org/sections/alltechconsidered/2017/04/23/524514526/dogs-are-doggos-an-internet-language-built-around-love-for-the-puppers
4. Dan, O., Davison, B. D., & Hong, L. (2011, March 28). *Predicting popular messages in Twitter* [Scholarly project]. In *WWW '11 Proceedings of the 20th international conference companion on World wide web*. Retrieved March 11, 2018, from http://wwwconference.org/www2011/
5. Faulkner, J. (2016, July 24). "WE ARE NOT A CULT": Analysing the lifestyle of Dogspotting through data. Retrieved March 11, 2018, from http://gutterstats.com/2016/07/24/dogspotting/
6. Floating-point arithmetic may give inaccurate results in Excel. (2017, January 7). Retrieved March 11, 2018, from https://support.microsoft.com/en-us/help/78113/floating-point-arithmetic-may-give-inaccurate-results-in-excel
7. Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.
8. Lee, K. (2014, April 7). Https://www.fastcompany.com/3028656/the-proven-ideal-length-of-every-tweet-facebook-post-and-headline-online [Web log post]. Retrieved March 11, 2018, from https://www.fastcompany.com/3028656/the-proven-ideal-length-of-every-tweet-facebook-post-and-headline-online
9. Lewis, C. (2017, August 22). 5 Twitter content formats that influence purchase behavior [Web log post]. Retrieved March 11, 2018, from https://business.twitter.com/en/blog/5-twitter-content-formats-that-influence-purchase-behavior.html
10. Ohlheiser, A. (2017, January 5). How an online fight about dogs became one of 2016's greatest memes. Retrieved March 11, 2018, from https://www.washingtonpost.com/news/the-intersect/wp/2017/01/05/how-an-online-fight-about-dogs-became-one-of-2016s-greatest-memes/
11. Seiter, C. (2014, May 13). Twitter 101: How To Get More Clicks, Retweets, And Reach [Web log post]. Retrieved from

https://www.fastcompany.com/3030462/twitter-101-how-to-get-more-clicks-retweets-and-reach
12. Valdez, A. (2018, January 10). Meme's Best Friend: The Rise of 'Doggo'. Retrieved March 11, 2018, from https://www.wired.com/story/rise-of-doggo/
13. What fuels a Tweet's engagement? [Web log post]. (2014, March 10). Retrieved March 11, 2018, from https://blog.twitter.com/official/en_us/a/2014/what-fuels-a-tweets-engagement.html
14. Zaiontz, C. (n.d.). Real Statistics Using Excel. Retrieved March 11, 2018, from http://www.real-statistics.com/free-download/real-statistics-resource-pack/
15. Microsoft (2017, January 7). Floating-point arithmetic may give inaccurate results in Excel. Retrieved March 11, 2018 from: https://support.microsoft.com/en-us/help/78113/floating-point-arithmetic-may-give-inaccurate-results-in-excel

Appendix

Appendix I: Full correlation matrices for @dog_rates and @dog_feelings

@dog_rates	shortid	textlength	retweet_count	favorite_count	hashtagcount	urlcount	mediacount	usermentions	day	rating	sneg	sneu	spos	scomp	dogs	good	meet	pup	boy	hckin	human	fren
shortid	1.00																					
textlength	0.33	1.00																				
retweet_count	0.18	0.02	1.00																			
favorite_count	0.42	0.07	0.90	1.00																		
hashtagcount	0.09	0.09	0.08	0.08	1.00																	
urlcount	0.06	0.02	0.24	0.36	0.03	1.00																
mediacount	0.18	0.44	0.06	0.17	0.04	0.41	1.00															
usermentionsco	0.02	0.07	0.11	0.14	0.13	0.11	0.07	1.00														
day	0.10	0.01	0.03	0.07	0.11	0.07	0.03	0.08	1.00													
rating	0.20	0.07	0.20	0.20	0.08	0.02	0.02	0.06	0.01	1.00												
sneg	0.03	0.05	0.06	0.06	0.05	0.04	0.04	0.06	0.04	0.05	1.00											
sneu	0.02	0.02	0.06	0.08	0.02	0.02	0.09	0.00	0.05	0.02	0.33	1.00										
spos	0.01	0.04	0.03	0.04	0.05	0.00	0.12	0.04	0.08	0.01	0.25	0.83	1.00									
scomp	0.06	0.12	0.00	0.02	0.07	0.01	0.03	0.05	0.06	0.03	0.62	0.43	0.80	1.00								
dogs	0.13	0.11	0.05	0.02	0.09	0.04	0.23	0.18	0.02	0.23	0.05	0.39	0.37	0.20	1.00							
good	0.14	0.18	0.01	0.01	0.05	0.11	0.12	0.01	0.09	0.06	0.06	0.35	0.40	0.31	0.16	1.00						
meet	0.17	0.30	0.06	0.03	0.09	0.03	0.14	0.06	0.24	0.01	0.08	0.08	0.04	0.04	0.02	0.03	1.00					
pup	0.10	0.19	0.02	0.02	0.04	0.18	0.14	0.06	0.05	0.06	0.00	0.11	0.11	0.04	0.13	0.06	0.00	1.00				
boy	0.07	0.00	0.03	0.11	0.07	0.20	0.15	0.08	0.05	0.15	0.12	0.17	0.24	0.26	0.08	0.44	0.06	0.01	1.00			
hckin	0.04	0.02	0.02	0.01	0.01	0.14	0.16	0.02	0.17	0.05	0.13	0.05	0.02	0.12	0.04	0.00	0.02	0.05	0.02	1.00		
human	0.24	0.34	0.01	0.06	0.01	0.05	0.17	0.07	0.14	0.14	0.02	0.07	0.08	0.15	0.00	0.01	0.01	0.02	0.05	0.00	1.00	
fren	0.13	0.13	0.49	0.31	0.23	0.10	0.18	0.06	0.35	0.04	0.05	0.07	0.11	0.13	0.02	0.05	0.00	0.01	0.07	0.02	0.01	1.00

@dog_rates	shortid	textlength	retweet_count	favorite_count	hashtagcount	urlcount	mediacount	usermentions	day	rating	sneg	sneu	spos	scomp	dogs	good	meet	pup	boy	hckin	human	fren
shortid	1.00	0.33	0.18	0.42																		
textlength	0.33	1.00	0.02	0.07																		
retweet_count	0.18	0.02	1.00	0.90																		
favorite_count	0.42	0.07	0.90	1.00																		
hashtagcount	0.09	0.09	0.08	0.08	1.00																	
urlcount	0.06	0.02	0.24	0.36	0.03	1.00																
mediacount	0.18	0.44	0.06	0.26	0.04	0.41	1.00															
usermentionsco	0.02	0.07	0.11	0.14	0.13	0.11	0.07	1.00														
day	0.10	0.01	0.03	0.07	0.11	0.07	0.03	0.08	1.00													
rating	0.20	0.07	0.20	0.20	0.08	0.02	0.02	0.06	0.01	1.00												
sneg	0.03	0.05	0.06	0.06	0.05	0.04	0.04	0.06	0.04	0.05	1.00											
sneu	0.02	0.02	0.06	0.08	0.02	0.02	0.09	0.00	0.05	0.02	0.33	1.00										
spos	0.01	0.04	0.03	0.04	0.05	0.00	0.12	0.04	0.08	0.01	0.25	0.83	1.00									
scomp	0.06	0.12	0.00	0.02	0.07	0.01	0.03	0.05	0.06	0.03	0.62	0.43	0.80	1.00								
dogs	0.13	0.11	0.05	0.02	0.09	0.04	0.23	0.18	0.02	0.23	0.05	0.39	0.37	0.20	1.00							
good	0.14	0.18	0.01	0.01	0.05	0.11	0.12	0.01	0.09	0.06	0.06	0.35	0.40	0.31	0.16	1.00						
meet	0.17	0.30	0.06	0.03	0.09	0.03	0.14	0.06	0.24	0.01	0.08	0.08	0.04	0.04	0.02	0.03	1.00					
pup	0.10	0.19	0.02	0.02	0.04	0.18	0.14	0.06	0.05	0.06	0.00	0.11	0.11	0.04	0.13	0.06	0.00	1.00				
boy	0.07	0.00	0.03	0.11	0.07	0.20	0.15	0.08	0.05	0.15	0.12	0.17	0.24	0.26	0.08	0.44	0.06	0.01	1.00			
hckin	0.04	0.02	0.02	0.01	0.01	0.14	0.16	0.02	0.17	0.05	0.13	0.05	0.02	0.12	0.04	0.00	0.02	0.05	0.02	1.00		
human	0.24	0.34	0.01	0.06	0.01	0.05	0.17	0.07	0.14	0.14	0.02	0.07	0.08	0.15	0.00	0.01	0.01	0.02	0.05	0.00	1.00	
fren	0.13	0.13	0.49	0.31	0.23	0.10	0.18	0.06	0.35	0.04	0.05	0.07	0.11	0.13	0.02	0.05	0.00	0.01	0.07	0.02	0.01	1.00

Appendix II: Most Pupular Tweets

@dog_rates


WeRateDogs™
@dog_rates
Follow

This is Jacob. In June of 2016, he comforted those affected by the Pulse nightclub shooting in Orlando. Four months ago he flew to Vegas to help families of 58 people grieve. Today, he is in Parkland, Florida hugging students who lost classmates yesterday. He is our third 15/10

1:30 PM - 15 Feb 2018

99,035 Retweets 283,557 Likes

LCC K9 Comfort Dogs

1.6K 99K 284K


WeRateDogs™
@dog_rates
Follow

This is Samson. His Halloween pajamas got lost in the mail. Didn't arrive til today. Rather pupset. 13/10 guess we're celebrating again

5:11 PM - 8 Nov 2017

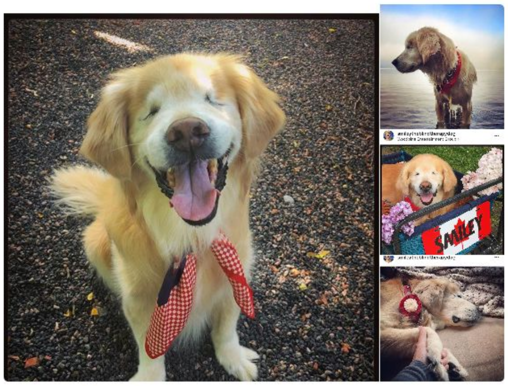
62,716 Retweets 218,620 Likes

1.2K 63K 219K

WeRateDogs™
@dog_rates

Follow

This is Smiley. He dedicated his life to making others smile. Today, his battle with cancer ends. He is our first 15/10. Rest easy Smiley ❤️



10:36 AM - 14 Oct 2017

54,435 Retweets 197,756 Likes

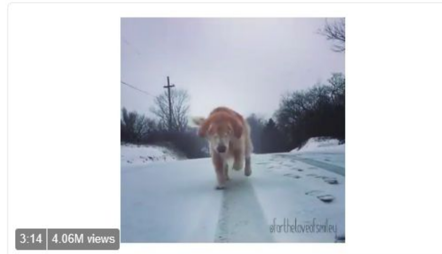
Smiley

2.7K 54K 198K

WeRateDogs™
@dog_rates

Follow

The Dogs of 2017



11:15 AM - 29 Dec 2017

86,584 Retweets 168,971 Likes

1.9K 87K 169K

WeRateDogs™
@dog_rates

Follow

I'm thankful for big dogs and small dogs, young dogs and old dogs, dogs with great floof, and dogs with no floof, dogs who protecc, dogs who attacc, and dogs who snacc, dogs with wheels and dogs who heel, dogs who save lives and dogs who high five... I'm thankful for dogs

8:54 AM - 23 Nov 2017

75,696 Retweets 167,964 Likes

716 76K 168K

WeRateDogs™
@dog_rates

Follow

Say hello to Ollie. This is his first snow experience. I think he's a fan. 13/10 would hug softly

4:58 PM - 25 Jan 2018

26,469 Retweets 139,058 Likes

Lauren

644 26K 139K

WeRateDogs™
@dog_rates
Follow

This is Jamesy. He gives a kiss to every other pupper he sees on his walk. 13/10 such passion, much tender

5:28 PM - 21 May 2017

36,802 Retweets **124,790** Likes

Phillip McCarthy

585 37K 125K

WeRateDogs™
@dog_rates
Follow

Here's a puppo who forgot he cloned himself. Quite the surpuprise. 13/10 for both (IG: bosunthefrenchie)

6:45 AM - 8 Nov 2017

43,673 Retweets **124,578** Likes

1.0K 44K 125K

WeRateDogs™
@dog_rates
Follow

Here is a puppo with a paw print on her nose. 13/10 would boop/high five

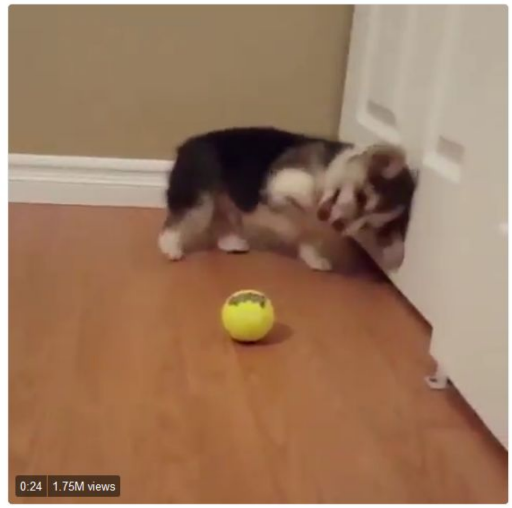
8:58 AM - 11 Sep 2017

36,414 Retweets **120,497** Likes

533 36K 120K

WeRateDogs™
@dog_rates
Follow

This is Gustaff. He's never seen a tennis ball put up such a fight before. 14/10 would assist in the conquering
(IG: itsthegustaff)



9:08 AM - 21 Nov 2017

39,453 Retweets **120,406** Likes

1.3K 39K 120K

@dog_feelings Most Pupular Tweets

Thoughts of Dog @dog_feelings

i had a long talk. with my fren. about how to spot. a fake ball throw. the optimal strategy. is to follow the ball. with your eyes. instead of your heart

2:51 PM - 27 Jan 2018

38,443 Retweets 178,053 Likes

556 38K 178K

Thoughts of Dog @dog_feelings

sometimes. the human gets sad. and i don't know why. but if i jump. in their lap. and stay there. i can usually fix it

6:41 PM - 14 Jan 2018

33,454 Retweets 166,634 Likes

613 33K 167K

Thoughts of Dog @dog_feelings

the human says. there are two options.. inside or outside. but if they would. elevate their mindset. they would uncover. a third option: stand. in the doorway. and sniff the air

8:54 AM - 7 Dec 2017

33,856 Retweets 145,521 Likes

631 34K 146K

Thoughts of Dog @dog_feelings

please find. your entertainment. elsewhere

12:13 PM - 14 Dec 2017

32,411 Retweets 137,053 Likes

684 32K 137K

Thoughts of Dog @dog_feelings

i drank. all the water. in my bowl. earlier. but just now. i returned. to the same bowl. and it was. full again.. the bowl. is haunted

7:07 PM - 5 Sep 2017

32,413 Retweets 126,259 Likes

382 32K 126K

Thoughts of Dog @dog_feelings

i heard. something about. making resolutions. and although i'm perfect. i have a few:

4. more peanut butter
b. increase. zoom efficiency
3. improve my sits
1. convince the human. i'm invisible
4. longer snoozles
f. sticks
2. more peanut butter
7. further turn up. the love dial

5:05 PM - 31 Dec 2017

28,559 Retweets 117,787 Likes

472 29K 118K

Thoughts of Dog
@dog_feelings
Follow

human. is sad again. two licks. on the hand. only a single pat. in return. i. will do better

8:12 PM - 26 Jul 2017

29,004 Retweets **114,678** Likes

486 29K 115K

Thoughts of Dog
@dog_feelings
Follow

the human. is hosting a get together. this evening. and my job. is to greet the guests. without jumping on them.. a nearly. impossible task

2:18 PM - 7 Jan 2018

15,753 Retweets **108,092** Likes

268 16K 108K

Thoughts of Dog
@dog_feelings
Follow

gooooob morning. the house tree. brought gifts. and now. i must remind everyone. that i. am the real gift here

6:17 AM - 25 Dec 2017

24,954 Retweets **106,296** Likes

205 25K 106K

Thoughts of Dog
@dog_feelings
Follow

my ear. is inside out. and the human. is not home to fix it. i have put the household. on alert level. dark grey.

11:04 AM - 11 Jan 2018

16,340 Retweets **95,983** Likes

288 16K 96K