An Analysis of Trump's Tweets: What Drives Engagement?

Osvaldo Valentino Aceves
505564001
oaceves@ucla.edu

Maximus Ong
705587695
maxong3@email

Wyatt Stone
005593807
wyattstone77@g.ucla.edu

Joseph Crosby
805497294
josephcc713@gmail.com

September 19, 2024

Abstract

Donald Trump has been an inflammatory figure for much of his time in the public sphere. This phenomenon was exacerbated by his presence on Twitter, only growing during his presidency from 2017 to 2021 until his banning from the platform in January of 2021. We are interested in whether the nature of his comments – as defined by the sentiments of anger, anticipation, disgust, fear, joy, sadness, surprise or trust – had any correlation with the amount of engagement his posts on Twitter received. We anticipated that joy and trust would be the least influential in terms of driving engagement. By the end of our investigation, we determined that joy did in fact have the least impact on engagement, and overall sentiment could be used to predict high or low engagement about 13% more accurately than flipping a coin.

1 Introduction

Throughout his tenure in the public eye, Donald Trump's presence on Twitter served as a focal point for both adoration and controversy. His prolific use of the platform, particularly during his presidency from 2017 to 2021, amplified his reach and influence, sparking intense debate and scrutiny. However, this era came to an abrupt end in January of 2021 when Trump was banned from Twitter, further fueling discussions surrounding online discourse and the impact of social media on public dialogue. In light of these events, our study aims to delve into the intriguing intersection of Trump's Twitter activity and audience engagement, particularly focusing on the relationship between the sentiment expressed in his tweets and the level of engagement they garnered.

With the aid of sentiment analysis techniques, we set out to explore whether the emotional tenor of Trump's tweets — ranging from anger and disgust to joy and trust — had any discernible relationship with the engagement they elicited from Twitter users. Our investigation was guided by the hypothesis that sentiments such as joy and trust would exert little influence on engagement compared to more polarizing emotions like anger or fear. Through sentiment analysis and statistical modeling, we look to uncover patterns and trends within Trump's Twitter career.

2 Preprocessing

The full dataset came as a .csv file with 56,571 observations and the following components:

1. Tweet: A tweet posted by the @realdonaldtrump Twitter account.
2. Retweets: Number of Retweets for respective tweet.
3. Favorites: Number of Favorites for respective tweet.
4. isRetweet: True if the tweet is a retweet. False otherwise.
5. isFlagged: True if the tweet was "flagged" by Twitter. False otherwise.
6. Date: Date and time the tweet was posted.

2.1 Data Cleaning

In preparing our dataset for analysis, several steps were taken to clean or preprocess the data. Firstly, we created a new column titled "engagements" which sums the counts of retweets and likes. By combining these variables into a single response variable, we focus on capturing the overall engagement of Trump's tweets. We then began the text cleaning process which involved removing irrelevant characters such as Twitter handles, image URLs, and miscellaneous text patterns that could potentially introduce noise in the data. Furthermore, empty observations resulting from the text cleaning were removed. Retweets were filtered out from the dataset, narrowing our analysis to only include original tweets by Donald Trump. The final step was to streamlined the dataset by retaining only the columns for relevant to our research objectives, namely "Tweet" and "Engagement." Ultimately, these data cleaning procedures culminated in a refined dataset of 45,379 observations. By meticulously cleaning the dataset, we remove potential noise and bias and lay a foundation for our analysis.

2.2 Sentiment Application

Using R's `syuzhet` library, we are able to get the NRC sentiment of the each tweet. This creates a new matrix with the same 45,379 rows but with ten columns being the sentiments for anger, anticipation, disgust, fear, joy, sadness, surprise, trust, negative, positive. Each column is a numerical vector that measures the valence of each emotion for each text input – in this case, each tweet. By quantifying the sentiment in the data, we effectively transform our dataset to reveal the emotional content conveyed in each tweet. Moreover, by column-binding the engagements column to the sentiment analysis results, we create a comprehensive dataset ready for data analysis.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0      72    1724   40764   67534 2278572
```

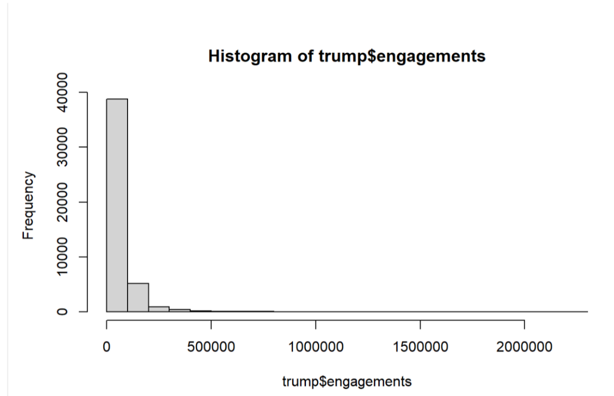

Figure 1: There are large outliers in the amount of engagements with the maximum engagement being the tweet where Trump announced his positive COVID-19 status.

3 Exploratory Data Analysis

3.1 Summary Statistics

Upon conducting exploratory data analysis, it became apparent that the dataset was right skewed in terms of engagement. This skewness indicated that a significant portion of tweets received low levels of engagement, while a smaller subset of tweets attracted substantially higher engagement. In fact, there was one outlier tweet – the one Trump posted when he tested positive for COVID-19 – that had nearly one million more engagements than the next most-engaged with tweet. Despite this distribution, we made a decision not to balance the dataset. Balancing the dataset to increase the representation of tweets with higher engagement, could potentially lead to overfitting. Overfitting occurs when a model over relies on patterns specific to the training data, decreasing its accuracy to new, unseen data. By keeping the original distribution of engagement, we aimed to maintain the integrity of the dataset and ensure that our analyses would capture the true variability and nuances in the data.

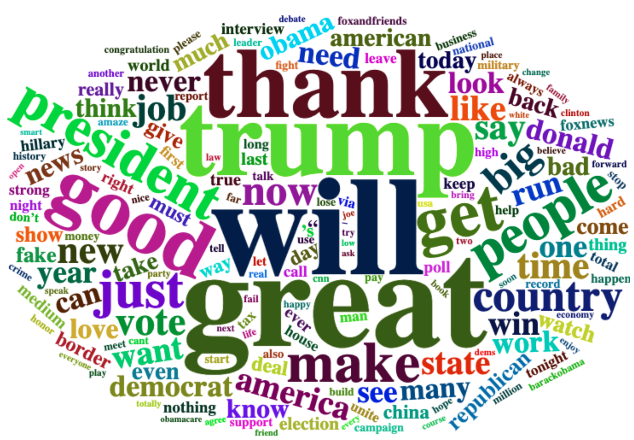

Figure 2: Most Common Words used in Trump's Tweets

3.2 Text Analysis

To understand the general textual content of tweets by Donald Trump, we performed basic text analysis to uncover the most frequently used words within his

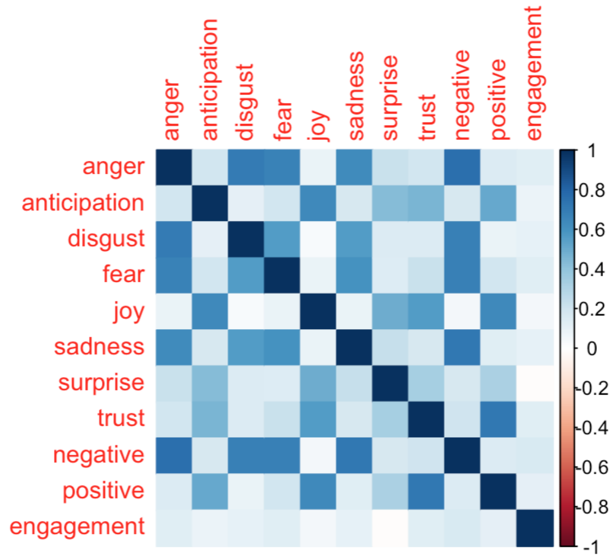

Figure 3: Correlation Matrix of NRC Sentiments and Engagement

tweets. We first preprocessed the tweet data by removing numerical digits, punctuation marks, and common stop words that typically do not carry significant semantic meaning. Subsequently, we employed lemmatization techniques to standardize the remaining words to their base or root forms to reduce redundancy. Using the `wordcloud2` library in R, we generated a word cloud visualizing the frequency of words within the tweets. The word cloud provided a visually compelling snapshot of the most prevalent words in Donald Trump's tweets. Notice that the most common words convey predominantly positive or neutral sentiment. In fact, many of these words were directly in the "Joy" lexicon dictionary used in our sentiment application of the dataset.

In addition to our explorations in basic text analysis, we constructed a correlation matrix using the NRC sentiment dataset. This provided a comprehensive overview of the interplay between various sentiments and engagement levels, offering valuable insights into our research question regarding the influence of sentiment on engagement.

The resulting correlation matrix showed higher correlation coefficients strictly between sentiments. For instance, "anger" had a strong correlation "disgust", "fear", and "sadness". The matrix suggests potential overlaps of emotions. Furthermore, our analysis revealed that engagement had essentially no correlation with sentiments as no sentiment demonstrated any significant correlation. This challenged our assumptions regarding the influence of sentiment on engagement and foreshadowed what would come to be later in the analysis. Overall, the correlation matrix was a crucial step in the analysis, laying the groundwork for subsequent stages of our analysis.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    53128.8      814.0  65.265  < 2e-16 ***
anger           3139.4     1072.9   2.926 0.003435 **
anticipation    4522.0      847.0   5.339 9.44e-08 ***
disgust        -3841.9     1124.7  -3.416 0.000636 ***
fear            2579.4      913.3   2.824 0.004744 **
joy            -1858.8     1070.0  -1.737 0.082357 .
sadness        -2961.5     1033.9  -2.864 0.004182 **
surprise      -12396.3      976.9 -12.689  < 2e-16 ***
trust           8031.7      694.4  11.566  < 2e-16 ***
negative       10002.4      754.6  13.255  < 2e-16 ***
positive        1863.6      612.1   3.045 0.002331 **
```

Figure 4: MLR table of dataset with less than 500 engagements

3.3 Multiple Linear Regression

The application of multiple linear regression allowed us to assess the significance of sentiment variables in predicting engagement levels. Our regression analyses were conducted across different subsets of the dataset to account for variations in engagement levels. We performed regressions on the full dataset, as well as on subsets excluding tweets with fewer than 100 and 500 engagements, respectively, to explore how engagement thresholds might impact the observed relationships.

We observed similar results across different subsets of the dataset with respect to the significance of sentiment variables. Some minor differences were in the regression performed on the subset excluding tweets with less than 500 engagements, which did not have a significant "joy" variable. Despite these nuances, the regression models had low R-squared values ranging from 0.048 to 0.092, indicating little explanatory power. This further pushes the notion emerging from our study that sentiment and engagement may not exhibit a straightforward relationship.

4 Random Forest

The final step in our analysis was to see what kind of predictive power the sentiments of a tweet may or may not possess. We decided to use a random forest analysis as our predictive method due to its classification abilities. Before we could employ the forest, though, we had to decide exactly how we wanted to classify the data. We landed on a binary method of `high` or `low` engagement. A tweet was classified as `high` engagement if it had more than the median number of engagements, and `low` if it had fewer than the median. Because we split the dataset into three different subsets based on engagement, we had three different medians of increasing value. That is, the median for all tweets was lower than the median of tweets with at least 500 engagements.

In the pursuit of thorough analysis, we conducted three different random forest analyses, one on each dataset. For each forest, the training dataset was made of 10,000 of the tweets, well the testing dataset was all the remaining tweets.

All three forests returned extremely similar results around 62-63%.

	Above	**Below**
Above	4162	4003
Below	2039	6016

Table 1: Table of Random Forest Results, tweets with over 500 engagements

As seen from the above table, the random forest for tweets with over 500 engagements yields an accuracy of 62.75%. This was the best result of any of the forests, but neither of the other two forests fell below 62%. This means that based on the sentiment of a trump tweet, we could predict if it had high or low engagement about 13% more accurately than flipping a coin to determine high or low engagement would be. While this is not the level of predictive accuracy we might want from a random forest, it does show us that there is enough of a relationship between engagement and sentiment to carry some level of significance.

5 Conclusion

5.1 Overall Conclusions

From our analysis, we discovered that Joy was the emotion with the least significant relationship to engagement and Trust performed better than we initially expected, comparing similarly to the other sentiments on the basis of engagement. However, the fact that there is not much variation in engagement levels with tweet sentiment shows that tweet sentiment was not as useful in predicting engagement as we expected; there were no sentiments that stood out in particular for having a significant influence on engagement.

Overall, the type of sentiment in a Trump tweet appears to be largely irrelevant in predicting engagement, being only slightly more reliable (13%) than flipping a coin.

5.2 Areas of Future Analysis

Although our initial investigation is a good starting point that did yield interesting results, there is always more to be analyzed. In our analysis, we did not account for several variables including whether the tweet was deleted, flagged for misinformation, or when the tweet was posted. Thus, future investigations could look into incorporating these variables by exploring the sentiments of deleted or flagged tweets or the average engagement of flagged tweets for example. Moreover, by using the data showing when tweets were posted, we can also consider if there is a relationship between the time a tweet is published and its sentiment or engagement. Finally, we could have explored if specific words or hashtags used by Trump had any correlation with engagement. We could further extend this by categorizing tweets by content (e.g. foreign policy, immigration, defense) and see if any particular tweet content has a strong relationship with engagement.