Abstract

A person's social media presence and activity can yield a lot of information about their personality, supporters, and daily habits–particularly in the case of a prominent public figure. In this paper, we use sentiment analysis techniques in R to uncover the trends and patterns lurking behind Donald Trump's 57,000 tweets from 2009 to 2021 while on the social media platform called X, formerly known as Twitter. We converted the textual data of Trump's tweets to numeric values based on the overall sentiment of the tweet, assigning negative tweets to negative values and positive tweets to positive values. Subsequently, we fit models and created visualizations to understand how the time of day affects tweet sentiment, whether or not certain tweets had a higher deletion frequency than others, and how Trump's original tweets differ from his retweets. Through our analysis of the data, we found that time of day could accurately predict the sentiment of a tweet, the sentiment of a deleted tweet typically skewed negative, and the average sentiment of a retweet is significantly different from that of a tweet written by Trump himself.

Introduction

Donald Trump, the 45th President of the United States, is known for expressing controversial opinions on Twitter (aka X). The Trump tweets data set contains 56,571 observations, each corresponding to one of his tweets from 2009 until his ban from the app in 2021. Trump's Twitter handle had over 88.9 million followers at the time of his ban, which occurred during his final days as President [1]. His extensive audience and use of social media as a form of political communication highlight the importance of analyzing his tweets, which often express polarizing opinions.

With the rise of social media in the political realm, there has been an extensive effort to understand how politicians should use social media to benefit their campaigns, gain more supporters, and improve their public perception. X, or Twitter, has been of particular interest due to its plethora of political activity and popularity among Americans all over the country. This project aims to investigate the specific ways Donald Trump used Twitter to reach his voters by analyzing trends relating to the sentiment of his tweets. If aspects of his social media use- such as the time of day the tweet was created or the activity he garnered with each tweet- can be broken down and better understood, we may gain valuable insight into the relatively new phenomenon of politicians incorporating social media into their campaigns.

Questions

After examining the data and considering the important context above, we decided to focus our analysis on answering questions we felt would illuminate Trump's behavior and patterns the most:

- Does the time of day that the tweet was sent out affect the resulting sentiment?
- Do tweets containing content that have a high level of negative/polarizing sentiment more likely to trigger action against it (i.e. tweet is deleted)?
- Is there a significant difference in sentiment between tweets written by Trump and his retweets?

Tweets Dataset

Data Structure

In addition to the tweet itself, the dataset we worked with contains the following information about whether the attention it received, the time it was posted, and more, storing everything under a mix of logical (true or false), numeric, and categorical variables. More specifically, the logical variables fall into the following categories: whether or not it was a retweet (isRetweet), deleted (isDeleted), or flagged (isFlagged). The numeric tweets include favorites, retweets, and the date of the tweet, down to the very second. There is only one categorical variable (device), which specifies the device the tweet came from.

Data Cleaning

Before any formal analysis could be conducted, we first ensured the tweets were in a form that would permit an accurate conversion of words to sentiment values. This meant removing characters that cannot be analyzed through sentiment analysis, such as links, numbers, whitespace, and symbols. Additionally, we also converted all the characters to lowercase and eliminated the punctuation.

Sentiment Creation

After cleaning the tweets, we utilized several packages in R to conduct the first part of the analysis (the main packages we used for this step include `tidytext`, `bing`, `word2vec`, and `stringr`) [1]. In general, words can be assigned a value based on meaning or connotation. We dissolved the collection of tweets into a (very long) string of individual words and phrases, assigning negative, neutral, or positive values to them based on their definition. For instance, 'egregious' may have a negative value, while 'book' may have a neutral value. Based on the

sentiment value for each word or phrase, we compiled the string back into each tweet, creating an overall score for each one and naming the resulting value 'net_sentiment' in our dataset. The 'net_sentiment' variable measures the sentiment of each tweet on a scale of -10 to 10, with most values falling between -3 and 3. Lower values indicate a more negative sentiment, while higher values indicate a positive sentiment. A score of zero or around zero signifies a neutral tweet overall. Note that a score of zero does not necessarily mean that the whole tweet was neutral-toned; it could also mean that there was roughly an even amount of positivity and negativity in the tweet. Now that our metric of interest has been derived, we can begin to address the above questions.

References:

Hicks, Stephanie. (2022, October 13). Working with Text: Sentiment Analysis. Retrieved from https://www.stephaniehicks.com/jhustatcomputing2022/posts/2022-10-13-working-with-text-sentiment-analysis/