Project 1 Establish Data & Analysis Plan
Group members: Reese Quillian, Lauren Smith, Ann Sofo
Leader: Ann Sofo
DS 4002
Tuesday Feb 14

I. Restate Hypothesis/Research Question/Model Approach

Hypothesis
We believe that the presence of six words/phrases in Donald Trump's tweets [*"Wall", "fake news", "media", "Democrat" , "great", "proud boys"*] signifies whether the tweet was posted before, during or after the 2016 election period.

Research Question
Did Donald Trump's Twitter language change over the course of his campaign/election year (before, during, after his election)?

Modeling Approach
We will be using a clustering algorithm to define groups within our data. Clustering algorithms take in unlabeled data (in this case, Donald Trump's tweets between 2009 and 2021) and look for meaningful patterns based on similarities of features. We expect to see clusters defined based on the presence of certain words and time period that the tweet was posted. However, we're most interested in learning from the model's ability to detect groupings in our dataset that we would not be able to easily detect ourselves. The fact that clustering is an unsupervised learning method works to our benefit; we're able to quickly identify the similarities and differences between our data points and determine which features are significant.

II. Executive summary

We modified an existing dataset to include additional features denoting the presence of the six words and phrases included in our hypothesis. Initial exploration revealed that Trump used at least one of the words of interest in about 24% of his tweets. After some further research, we decided to build a model using a classification algorithm to predict if the presence of these particular words or phrases indicate whether the tweet was sent before, during or after the 2016 election.

III. Dataset establishment details

Dataset Summary
The original dataset was downloaded from Kaggle. Published in 2021, the csv file provides daily tweet data from Donal Trump's twitter account from May 2009 (when it was created) until January 8 2021 (when his account was blocked). The dataset provided the text, number of favorites and retweets, whether it was deleted, and the date. In total we added 8 features, with 7 to identify whether specific words and phrases from our hypothesis were present, and one to identify the election period in which the tweet was posted.

To create the election_period column, we ensured that the date column was being stored as a date and then used case_when() and simple inequalities to categorize each tweet under their correct time period. For the contains_word column, str_detect() was used in conjunction with case_when() to assign a 1 when any of the phrases were found. Missing values were replaced with a 0 after. The individual word features (wall, fake_news, etc.) were assigned in a similar manner, but with six if / else statements to create six different columns. The full description of each feature in our established dataset can be found in the Data Dictionary below.

Data Dictionary

Variable Name	**Description**
text	The content of the tweet made
favorites	Number of likes tweet has
retweets	Number of times tweet was retweeted
date	Date of the tweet
isDeleted	Binary variable showing whether a tweet has been deleted (0 = No, 1 = Yes)
election_period	Labels whether tweet was posted pre,post, or during the 2020 election *Pre: Before June 16th, 2015* *During: June 16th, 2015 - November 8th, 2016* *After: After November 8th, 2016*
wall, fake_news, media, democrat, great, proud_boys	Binary variables labeling whether these words are present in the tweet (0 = No, 1 = Yes)
contains_word	Binary variable labeling whether any of the targeted words are mentioned in the tweet (0 = No, 1 = Yes)

**highlighted variables are those which we created*

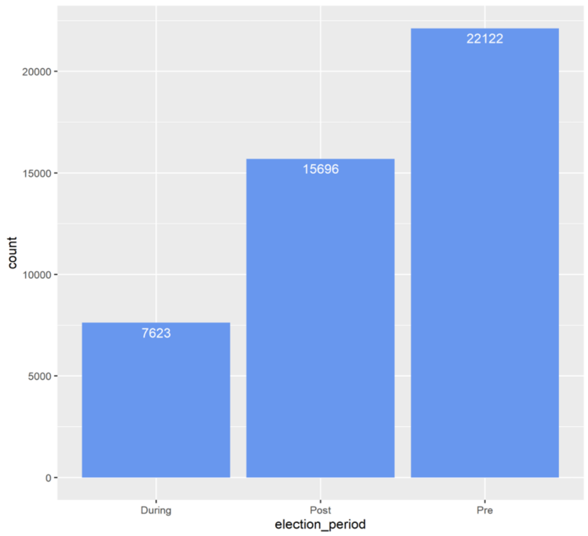
The full dataset and code resources used to create it can be accessed through Github: https://github.com/smithlauren785/DS4002-Project1/tree/main

Questions Explored

For the initial exploration of the content of our data, we chose three questions to visualize:

1. In what time period did Donald Trump send the most tweets? (Figures 1 and 2)
2. Which words appeared most frequently in Trump's tweets? (Figure 3)
3. How often did Trump tweet each of the words of interest? (Figure 4)

Figure 1 shows that the most tweets were posted in the pre-election period, followed by post-election, and the least amount posted during the election period. While this is helpful in terms of understanding the distribution of our data, we need to account for the length of time each period encompasses (pre-election covers approximately six years of tweets, during the election only covers one year, and post-election covers five years). To normalize this we calculated the average number of tweets per month for each election period. So even though Donald Trump sent the most tweets prior to his election, it was during the election and after where he posted the most frequently.



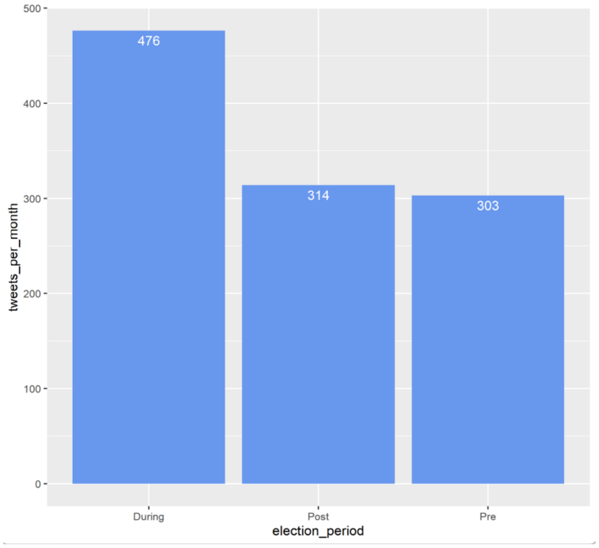
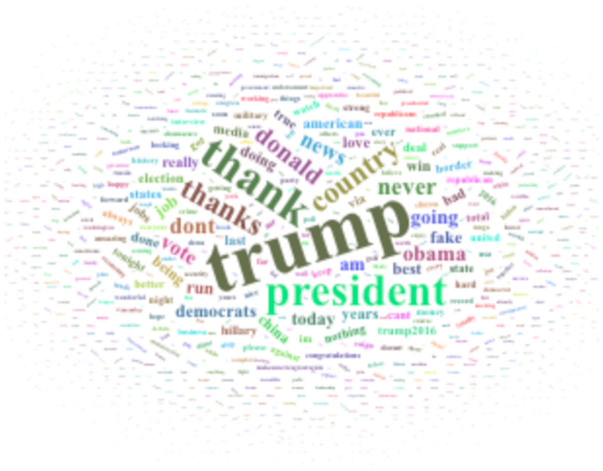


Figure 1. Number of Tweets posted before, during, and after the 2016 election

Figure 2. Number of Tweets per month posted before, during, and after the 2016 election

Figures 3 and 4 explore the content of the Tweets being posted. Before looking at the six words and phrases relevant to our hypothesis, we created a word cloud (Figure 3) with all words present in the text data. "Trump" is noticeably the most frequently appearing word, followed by "president" and "thank". It can also be observed that two of the words in our hypothesis appear with a high frequency ("democrat" and "fake").



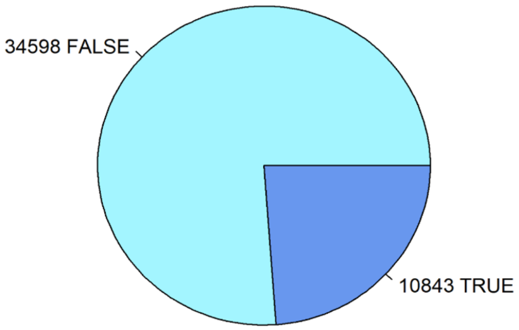


Figure 3. Words appearing most frequently in Donald Trump's tweets

Figure 4. Proportion of Tweets containing at least 1 word of interest

Figure 4 shows that approximately 24% (almost 11,000 total) of tweets posted by Donald Trump between May 2009 and January 2021 contain at least one of the six words or phrases in our hypothesis. However, some of these words appear much more frequently than others, as shown in Table 1.

Table 1: Number of Tweets containing words of interest

	wall	fake_news	media	democrat	great	proud_boys
total	521	882	1131	1649	7784	0

Current Unknowns

The initial exploration outlined provides context about the time period of tweets posted and their content, but separately. What still remains unknown is whether the presence of certain words and phrases correlates with when the tweet was posted. This will be explored when we run classification algorithms on our data, as we look at the ability of the content-related features we created to predict the time period in which the tweet was posted. Another question that is still unanswered is whether there are certain features/words that are more associated with different time periods compared to others. Again, this is something that will become clear after testing our model, where we can observe the changes in accuracy with the addition or removal of certain features.

Refinement of hypothesis/model

After further exploration, we discovered that Trump did not tweet the phrase “proud boys” and we decided that our research would be more fascinating if we predicted the time period that a word he frequently used was tweeted leading us to replace “proud boys” with “trump” in our hypothesis. As we could see from the world cloud, “trump” has been his most frequently tweeted word. Additionally, we revised our method of analysis from performing clustering to classification as we recognized that our data was labeled. Each tweet was categorized as containing one of the words or phrases of interest or not containing it. Therefore, we felt it would be more straightforward and insightful to perform classification to see if we could predict the time period one of Trump’s tweets were sent based on the presence of six key words and phrases. With this new modeling approach, we plan to run multiple classification models on our data to attempt to predict when a tweet was sent. After analyzing the accuracy of each model, the model most accurate at prediction will be chosen as the final model and further conclusions will be made based on our hypothesis.

Revised hypothesis: We believe that the presence of six words/phrases in Donald Trump’s tweets [*“Wall”, “fake news”, “media”, “Democrat” , “great”, “trump”*] signifies whether the tweet was posted before, during or after the 2016 election period.

With our finalized hypothesis in mind, we wanted to obtain a better understanding of how the presence of these words varied within each period. We asked two questions: “How did the proportion of tweets containing the word ‘Trump’ change within each period?”, and “How did the proportion of tweets containing any of the words in our hypothesis change within each period?” (Figures 5 and 6). It can be observed that the proportion of tweets containing certain words changes between election periods. Specifically, Figure 5 shows that there is a stronger presence of the following words/phrases: [*“Wall”,*

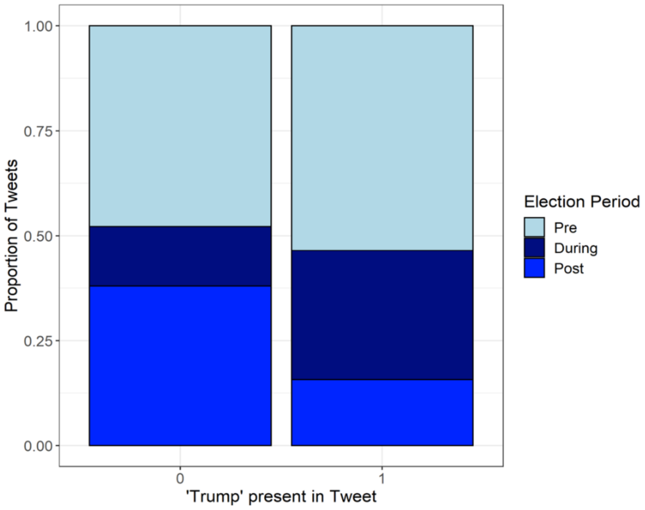


Figure 5. Proportion of tweets containing the word ‘Trump’ by election period

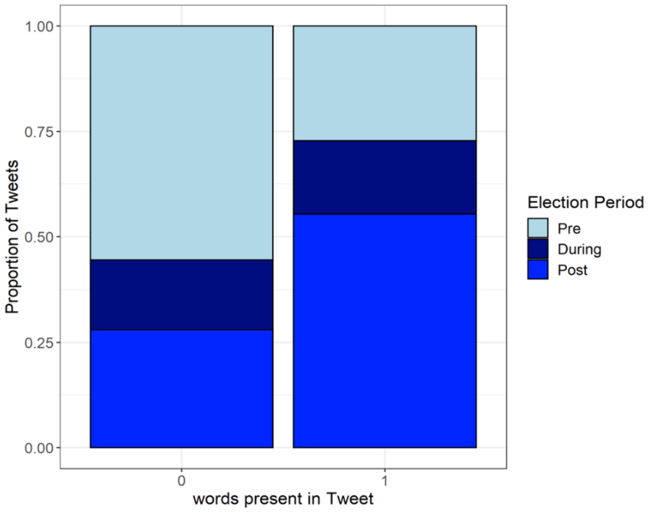
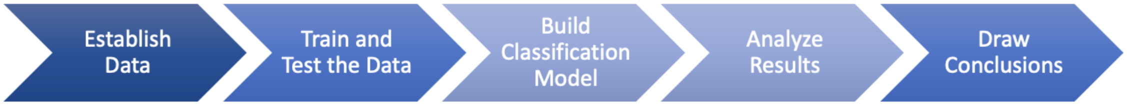


Figure 6. Proportion of Tweets containing at least 1 word of interest by election period

“fake news”, “media”, “Democrat” , *“great”, “trump”*] after the 2016 election. This supports our hypothesis and will be examined further in our later analysis.

IV. Analysis Plan



Paragraph Summary of each Step

1. Establish Data

In this step, we will solidify our target words and features we want to use and verify that each word is meaningful for our overall research question and motivation. In addition to establishing words, we will finish cleaning our datasets so that it is completely ready to run multiple classification algorithms on. Depending on the algorithm, we may have to alter the dataset slightly to adhere to that particular procedure.

2. Train and Test the Data

With supervised learning, we will need to create training and testing sets to prepare for model building. We will do this by splitting the data into 70% training and 30% testing, and then ensuring that our classes (pre/post/during the election) are not too unbalanced so we can prevent overfitting. If we do find that we need to make the classes more even, we plan on grouping the ‘during’ and ‘pre’ classes together or doing some kind of change to the category splits we initially did based on date.

3. Build Classification Model

After creating training and testing sets, we will then be able to deploy a classification model. We plan on creating multiple models to see which will lend the best results; while decision trees tend to produce quick, clear results, a random forest model may give us less overfitted results of the classification of our data – these are some key components we will keep in mind during the model building process. Furthermore, we will test different features and methods – such as bagging or boosting to correct overfitting or underfitting – to later compare whether the decision tree or random forest methods will produce the best results.

4. Analyze Results

A variety of evaluation metrics will be used to determine how successful our model is; for example, accuracy, precision, recall, variable importance, ROC curve, and AUC value. The correlation between our words of interest and the predicted time period of a tweet is also a number we'll find. On top of these metrics, we will analyze results by seeing how the predicted classes compare with the true values based on outside context that is not necessarily quantifiable. There is a possibility that the model will catch something that we did not in the data and we may address that by changing some of our previous steps, or maybe even re-addressing our research question.

5. Draw Conclusions

Using the evaluation metrics and results produced, we'll determine whether our quantifiable goal is achieved. The ranking of variable importance is also a metric that could help us draw conclusions besides accuracy, precision, etc., as it relates to our hypothesis that certain words (our features) will play important roles when predicting Trump's tweets' time periods. Coming up with conclusive answers to our preliminary questions and validating our hypothesis will also be a part of this step; bringing it all together and trying to find explanations for why certain metrics and results came to be will be part of our conclusion(s) as well.

Quantifiable Goal for Analysis

Ideally, we are looking to achieve an accuracy rate of 95% or above with our model. However, we will also be accounting for the other evaluation metrics mentioned above. Perhaps even finding at least 6 predictors/ words with the most significance that may be different than the ones in our hypothesis could be an added goal.

V. References

Classification | Machine Learning. (n.d.). Google Developers. Retrieved February 14, 2023, from [*https://developers.google.com/machine-learning/crash-course/classification/video-lecture*](https://developers.google.com/machine-learning/crash-course/classification/video-lecture)

Shantanu Roy. (2021). *Donald Trump Tweets Dataset*. https://www.kaggle.com/datasets/codebreaker619/donald-trump-tweets-dataset

Tuychiev, B. (2022, December 9). *Comprehensive Guide to Multiclass Classification With Sklearn*. Medium.https://towardsdatascience.com/comprehensive-guide-to-multiclass-classification-with-sklearn-127cc500f362