Comparison, Clustering, Classification, and Feature Analysis of Trump's Tweets in 2017 and 2020

Aaron Tsui

ABSTRACT

The central topic around my semester final project was President Donald J. Trump's tweets. I wanted to analyze how positive or negative his tweets were, what words are most and least distinctive of his tweets from year to year, and how the pandemic affected how he tweeted in 2020 as compared to how he tweeted in 2017 (and, in addition, what he tweeted about). I also wanted to cluster the data based on month and see which months would cluster near other months in terms of Euclidean distance.

The dataset is entirely composed of tweets– or 140 character maximum blurbs– written by President Trump through the Twitter platform for his millions of followers during his presidential term. The dataset has 56,572 total tweets with 9 separate variables each tweet has information on. These include tweet id (which is a unique identification number to be able to find a specific tweet with a specific unique tweet id), text (which is the exact tweet text of each of Trump's tweets, with a 140 character maximum), isRetweet (a boolean value that defines whether or not the tweet was a retweet), isDeleted (another boolean value that defines whether or not the tweet was deleted by President Trump), device (which device, platform, or application President Trump tweeted from), total favorites, total retweets, date (in the form of YYYY-MM-DD 00:00:00), and isFlagged (another boolean value which defines if the tweet was flagged).

Originally I also wanted to find out whether President Trump's deleted tweets had anything to do with being contrary to his political agenda, but through my analysis, many of the deleted tweets were extremely similar to other tweets in terms of Euclidean distance, likely meaning that the deleted tweets were simple typos or other spelling/grammar errors that would normally be meaningless in an in-depth analysis of President Trump's tweets. Thus this question was ultimately removed from the research question.

With these research questions in mind, I expected to find a somewhat significant difference between President Trump's tweets in 2020 as compared to those in 2017 due to the Coronavirus global pandemic being a massive effect on people worldwide. I decided to continue my college education through the pandemic and as a result it has caused me increased stress in its future consequences in terms of course availability.

Likewise, some of my family members lost jobs during the pandemic, like many Americans have. I also expected to see the months before November to be clustered near November because November is the main election month, as well as January to be relatively busy due to said typical presidential State of the Union address(es). Like many others, I also expected the sentiment of President Trump's tweets to be negative, due to how he is typically portrayed in public spaces and by many people in casual conversation. Coming into this project, I had low levels of initial expectation in terms of what words would be most or least distinctive of President Trump's tweets, but I did have a general idea around what words would be more likely to appear in either set.

RELATED WORK

These specific research questions have not been exactly explored before by other data scientists, but multiple other studies have analyzed President Trump's tweets' effect on various other measures of society.

One study, done by Heleen Brans and Bert Scholtens, was published on March 11th, 2020, studied the effect of President Donald Trump's Tweets had on the stock market[1]. It explored the effect President Donald Trump's Tweets had on the stock market over a period of 100 days, and involved three numerical measurements that showed general stock elements each day, over the period of 100 days[2]. These were defined as Alpha, Beta, and AAR, which measured the risk involved with each stock, the relationship between market and stock returns, and the average abnormal return as an effect of President Donald Trump's tweets, respectively[3].

Another study, done by Isobelle Clarke and Jack Grieve, was published on September 25, 2019, studied the stylistic variation and analyzed the linguistics of President Trump's tweets between 2009 and 2018[4]. It classified tweets multiple ways, one was through a series of 4 separate dimensions– conversational style, campaigning style, engaged style and advisory style– and also clustered by literary objects like the various tenses, interjections, use of common parts of speech, use of URL, etc[5]. The study also produced a structured heatmap to visually represent the classification

---

[1] Brans H, Scholtens B (2020) Under his thumb the effect of president Donald Trump's Twitter messages on the US stock market. PLoS ONE 15(3): e0229931. https://doi.org/10.1371/journal.pone.0229931
[2] Brans H, Scholtens B (2020) Under his thumb the effect of president Donald Trump's Twitter messages on the US stock market. PLoS ONE 15(3): e0229931. https://doi.org/10.1371/journal.pone.0229931
[3] Brans H, Scholtens B (2020) Under his thumb the effect of president Donald Trump's Twitter messages on the US stock market. PLoS ONE 15(3): e0229931. https://doi.org/10.1371/journal.pone.0229931
[4] Clarke I, Grieve J (2019) Stylistic variation on the Donald Trump Twitter account: A linguistic analysis of tweets posted between 2009 and 2018. PLoS ONE 14(9): e0222062. https://doi.org/10.1371/journal.pone.0222062
[5] Clarke I, Grieve J (2019) Stylistic variation on the Donald Trump Twitter account: A linguistic analysis of tweets posted between 2009 and 2018. PLoS ONE 14(9): e0222062. https://doi.org/10.1371/journal.pone.0222062

brackets[6]. They also measured stylistic change across a multi-year period using their own systems[7].

DATASET

The entirety of the dataset that I used for this project, code, and report is from Twitter, mainly tweets taken from President Trump's twitter account using a web scraper. It represents a majority of the social media presence that President Donald Trump had during his presidential term. It also represents the main source of contact he had with the general public for both his presidential campaign as well as term. Its limits definitely include the fact that it is only restricted to Twitter, as in data from other social media platforms (like Facebook, TikTok, Youtube, etc.) where President Trump may have had or do still have influence are not counted as part of this dataset. Some key collection statistics are as follows:

Donald Trump's Followers	Current Data
Twitter Followers (despite suspension)	88,783,411
Facebook Followers	**34,844,444**
TikTok Followers	**413,300**

METHODOLOGY

The techniques I used to analyze the data are comparison by Jaccard similarity, clustering by Euclidean distance, classification via sentiment and political partisanity analysis, and feature analysis via dunning G scores. I used quite a few python libraries and packages. Namely, I used Pandas, Numpy, OS, RE, and csv. In addition, I used Counter from collections, various functions from sci-kit-learn and scipy. I used VSCode and Jupyter Notebook in terms of existing software. VSCode was used as a second notebook to test out and modify functions to fit my semester project's and analysis' needs. Jupyter Notebook was used to house the entire codebase for this project. Google Sheets was also used to inspect and manipulate the dataset to fit the needs of this project. Each of my methods answers a separate part of my research question, as there are various research questions. Comparison by Jaccard similarity shows how President Trump's tweets changed after the emergence of the pandemic. Clustering by

[6] Clarke I, Grieve J (2019) Stylistic variation on the Donald Trump Twitter account: A linguistic analysis of tweets posted between 2009 and 2018. PLoS ONE 14(9): e0222062. https://doi.org/10.1371/journal.pone.0222062

[7] Clarke I, Grieve J (2019) Stylistic variation on the Donald Trump Twitter account: A linguistic analysis of tweets posted between 2009 and 2018. PLoS ONE 14(9): e0222062. https://doi.org/10.1371/journal.pone.0222062

Euclidean distance shows which months are closer together in terms of topic and content. Classification via sentiment analysis shows us how positive or negative Trump's tweets are, and classification via political partisanity analysis shows us whether or not President Trump talks more about the Republican Party or the Democratic Party. Feature analysis via dunning G scores shows us what words are most and least distinctive amongst the dataset.

ANALYSIS

COMPARISON

We tried to compare tweets on specific days in 2017 vs 2020, but Jaccard similarity resulted in 0.0, meaning there was no similarity between the respective pairs of tweets. This didn't give us any constructive information to analyze so we chose a different set of tweets set in different time periods.

We compared the first 2460 tweets Trump tweeted after he was sworn in to the first 2460 tweets Trump tweeted after his first mention of the pandemic. This resulted in a Jaccard similarity of 1.0, meaning the two datasets are incredibly similar, which may suggest that Trump's way of tweeting didn't change after the existence of the pandemic from that of his pre-pandemic tweets.

```
#-----------------------------------------------------------------------
def jaccard(vector_a, vector_b):
    # Construct sets of the non-zero entries of each vector
    set_a = set(np.flatnonzero(vector_a))
    set_b = set(np.flatnonzero(vector_b))
    # Compute the intersection and union of the two sets
    overlap = set_a & set_b
    overall = set_a | set_b
    return len(overlap) / len(overall)
```

```
id_index = {data['id']:i for i, data in enumerate(tweet_data)}
```

```
jaccard(tweet_data[id_index['2017']], tweet_data[id_index['2020']])
```

```
1.0
```

CLUSTERING

These are very expected results, October and November are typically voting/election months. February and December have relatively low political activity and thus are clustered with September, which is sandwiched between the 2 election months after it and the months of Cluster 3.

- Cluster 0: Jan, Apr, Jun
- Cluster 1: Feb, Sep, Dec
- Cluster 2: May
- Cluster 3: Jul, Aug
- Cluster 4: Oct, Nov
- Cluster 5: Mar

CLASSIFICATION

Creating a Lexicon for Sentiment Analysis: We found the Top 500 most common words in the dataset as well as their respective counts and we will preprocess that result to create the lexicon we will use for SA. We used the syuzhet lexicon as a basis for some of the scores in our lexicon.

This political lexicon measures whether he tweets about his own Party (the Republican Party) or the Democratic Party (and its members) more.

In the first 5 months he tweets a lot about the Democratic Party, whereas he is relatively neutral in June. However for the rest of the year Trump tweets mostly about the Republican Party and ironically doesn't tweet about his party as much in November (the main election month) as compared to September, October, or December. The likely cause of this is that a vast majority of elections take place in early November and the rest of the month isn't nearly enough time to know complete and definitive results as compared to December when polled results have a higher chance to be accurate to the whole population. Mid-to-late November is merely speculation time for most elections as edge cases such as mail-in or absentee votes have likely not been counted yet as part of the final election total.

Another possible reason is that the Electoral College typically votes on the Monday after the second Wednesday in December, which takes place after mid-to-late November.

```
for item in bymonth_data:
    print(score_counts(item['counts'], pl))
```

```
-0.9000000000000004
-1.9
-3.0
-8.200000000000001
-3.2
6.9
23.300000000000004
44.800000000000004
50.2
26.700000000000006
14.299999999999997
29.3
```

This lexicon measures how positive or negative Trump's tweets are by month.

Most people would think of Trump as a hateful figure, spreading hate towards LGBTQ+ people and multiple other minority groups. Because of this, you would expect the general sentiment of Trump's tweets to be overwhelmingly negative. But this simply is not the case. The vast majority of his tweets are quite neutral, and aren't mostly negative.

It really puts how he was portrayed by some media outlets into a different perspective. Granted, this analysis consists only from Trump's *tweets,* which means he may say (or may have said) more negative things in interviews or other sources that have shifted people's opinions over time.

Separate from analysis, simple observation and observational thinking shows us:

- Media outlets are **companies**, which need to make money to stay in business
- To make money, they need readers/users
- Articles with high absolute value sentiment measurements tend to invite more emotion[8]

[8] Tyng, Chai M et al. “The Influences of Emotion on Learning and Memory.” Frontiers in psychology vol. 8 1454. 24 Aug. 2017, doi:10.3389/fpsyg.2017.01454

```
for item in bymonth_data:
    print(score_counts(item['counts'], tsl))
```

```
16.599999999999998
29.649999999999988
46.35
50.20000000000001
46.05
60.75
68.69999999999999
70.95000000000003
119.59999999999998
51.05000000000001
92.60000000000002
63.49999999999999
```

FEATURE ANALYSIS

Top 10 Most and Least distinctive words used in Trump Tweets in 2017 and 2020.

The results are pretty reasonable and expected. Many of the most distinctive words are about topics that set the tone and general idea of the tweet. You can't really 'terrorists' in the same tweet as 'violations' because they both deal with separate distinct topics that generally don't have overlap. Many of the least distinctive words are general ideas like twitter or hurricaneirma; Twitter is the platform Trump is tweeting on and Hurricane Irma was a natural disaster that affected a lot of people.

```
print("Most distinctive words used by Trump in 2017 and 2020:")
print_extremes(vocabulary, scores, 10, True)
```

```
Most distinctive words used by Trump in 2017 and 2020:
1. violations
2. uranium
3. reasons
4. dream
5. notice
6. study
7. terrorists
8. europe
9. vetting
10. extreme
```

```
print("Least distinctive words used by Trump in 2017 and 2020:")
print_extremes(vocabulary, scores, 10, False)
```

```
Least distinctive words used by Trump in 2017 and 2020:
1. the
2. ricardo
3. pro-growth
4. careful
5. muslim
6. hurricaneirma
7. shelters
8. twitter
9. destructive
10. flgovscott
```

CASE STUDY

The comparison between Trump's tweets in 2017 to his tweets in 2020 is very interesting. Before this study, I had thought that most people's lives were massively influenced by COVID-19 and the Coronavirus pandemic, but that isn't applicable to President Trump. I definitely expected much of his tweets to have some visible effect from the pandemic but it's almost entirely negligible. I think while we may see presidents in general as more "the people's representative" and see them as such when they are elected, we also have to take into account that almost every president, former or in term, is worth tens of millions of dollars. And if not at that level, they are billionaires. In that sense, we (as a nation) should stop and realize that while they may say they have our best interest at heart, that typically is not *entirely* the case. If someone even has a 1% chance of becoming president, they are already far beyond the salary and financial class of the average American.

The feature analysis, especially the most distinctive words for President Trump's tweets, is what I also find interesting. Uranium is the second most distinctive word, which is strange considering this was before Trump visited the North Korean dictator Kim Jong Un in June 2019, and far before the current challenging political landscape when it comes to Russia's invasion of Ukraine. Not to mention how 'terrorists' is relatively low on the top 10 most distinctive words. In hindsight, it feels strange considering the threat of ISIS was likely one of the biggest foreign affairs problems in 2017. This is possibly another point on realizing how much of an effect news media has on people's opinions of current events.

CONCLUSION

In conclusion, President Trump's tweets were not greatly affected by the pandemic. President Trump tended to tweet more about the Democratic Party during the first half of the year and more about the Republican Party during the latter half.

I learned a lot from this process. I had multiple errors and bugs that I had to resolve in order to preprocess the dataset successfully. In terms of follow up questions, I'd like to know whether some of the same conclusions reached in this study are applicable to other presidents, namely President Obama or President Biden. If I had unlimited time and resources, I would do the same analysis for every president that has had a presence on Twitter since Twitter's creation. I would also expand this analysis to other presidential candidates or other famous figures with a strong and wide social media presence.

Bibliography

1. Brans H, Scholtens B (2020) Under his thumb the effect of president Donald Trump's Twitter messages on the US stock market. PLoS ONE 15(3): e0229931. https://doi.org/10.1371/journal.pone.0229931
2. Clarke I, Grieve J (2019) Stylistic variation on the Donald Trump Twitter account: A linguistic analysis of tweets posted between 2009 and 2018. PLoS ONE 14(9): e0222062. https://doi.org/10.1371/journal.pone.0222062
3. Tyng, Chai M et al. "The Influences of Emotion on Learning and Memory." Frontiers in psychology vol. 8 1454. 24 Aug. 2017, doi:10.3389/fpsyg.2017.01454