Data and Programming for Public Policy II: Final Project Report

Celia Zhu
xiling@uchicago.edu
The University of Chicago
Chicago, IL 60615

1 Introduction

This project explores tweets that “hashtagged" Donald Trump and those “hashtagged" Joe Biden during the 2020 US Presidential Election. I analyzed the setiments of those tweets, estimated the effect of campaign events on candidate’s presence on Twitter, and visualized the geographical distribution of popular tweets along with their contents.

2 Data

I used three datasets for tweets and campaign event history.

- Tweets hashtagged Donal Trump

  File name: `hashtag_donaldtrump.csv`

- Tweets hastagging Joe Biden

  File name: `hashtag_joebiden.csv`

- 2020 Presidential Candidate General Election Events Tracker

  File name: `2020 Presidential Candidate General Election Events Tracker(maintained by FairVote, Nov Version).xlsx`

In the project report, I refer to the first two datasets as tweets, and the third as campagin events.

Tweets were collected by Kaggle user Manch Hui. The campaign events were maintained by FairVote

3 Research Questions

I intended to explore three questions:

1. Overall, whether the tweets hashtagged Trump and those hashtagged Biden expressed different sentiments?

2. Does campaign events promote candidates’ presence on Twitter?

3. What did the most popular tweets say about the two candidates, and how those tweets distributed geographically?

4 Approach and Coding

4.1 Data Wrangling

Tweets data were in two separate `.csv` files. I merged them into one tidy dataset to perform event study, and extracted tweets into `.txt` files for sentiment analysis. To remove tweets from potential twitter “bots", I dropped tweets from users joining later than 2020-01-01 AND with no more than 5 followers.

Campagin events were organized in a wide format. One row of observation corresponded to two events: one is by Biden/Harris, the other by Trump/Pence. I cleaned the campaign events into a tidy data format.

The script for this step is `step1_data_wrangling.R`.

4.2 Sentiment Analysis

I tokenized and stemmed the tweets that hashtagged either Trump or Biden. [1] I used Bing sentiment lexicon and NRC emotion lexicon.

The script for this step is `step2_text_analysis.R`.

Results of sentiment analysis Overall, compared with the tweets that hashtagged Trump, those hashtagged Biden expressed more positive sentiments, such as “Anticipation", “Joy", and “Trust". I presented the results in figure 1

4.3 Event Study

My second research question intends to explore if campaign events promote candidates' presence on Twitter. In a simple DID setting, where treatment[2] starts at the same time for all treated units[3], it is straightforward to investigate the pre-trend. But candidates visit different states at different times. To investigate the pre-trend, we need to aggregate over multiple campaign visits and define event time k in the following way:

Suppose a candidate visited state i at date t_i^*. Let the date t be given, we define the event time $k = t - t_i^*$.

I define the window for each event to be 2. That is, we will consider the tweets two days prior to the campaign visit, and two days after the visit. This is meant to obtain a balanced panel and minimize the overlap between different events.

To estimate the average effect of campaign visits on the number of tweets hashtagged the candidates from the visited state, I fitted the model

$$tweets_{it} = \alpha_i + \sum_{k=-2}^{2} \phi_k D_{it}^k + u_{it} \tag{1}$$

where $tweets_{it}$ is the number of tweets hashtagged the visiting candidate from a state i on a date t, α_i is the state fixed effects so we can preserve the daily variation in the coefficients, ϕ_k is the average treatment effect on the treated states, D_{it}^k is the dummy variable for event time k in state i on date t.

To simplify the event study analysis, I removed those VP candidates' visits because I didn't collect tweets that specifically hashtagged Harris or Pence. I also imposed the homogeneous treatment effects across all visited states, and separately fitted the model for those tweets hashtagged Trump and those hashtaging Biden.

The script for this step is `step3_analysis.R`.

[1] For simplicity, I did not extract or analyze tweets that hashtagged both candidates.

[2] In our context, the treatment is candidates' campaign visits.

[3] I used states as units.

Effect of Campaign Events on Twitter Presence The number of tweets hashtagged Trump changes volatilely with his campaign event time. From the figure 2, the number of tweets hashtagged Trump increased a lot on the days of his campaign visits. While the number of tweets hashtagged Biden are not very correlated with his event time – the number of tweets hashtagged Biden even dropped on the days of his campaign visits.

Our analysis concludes that campaign events do not have significant impact on candidates' presence on twitter: most of the effects shown in figure 2 and figure 3 are null effects.

4.4 Interactive Plot

I built an interactive map showing the contents of tweets hashtagged Biden or Trump in the election year, the "like" counts, and the locations.

The script for this step is `step4_interactive_plot/app.R`. The published web application is here.

5 Discussion

5.1 Selection Bias

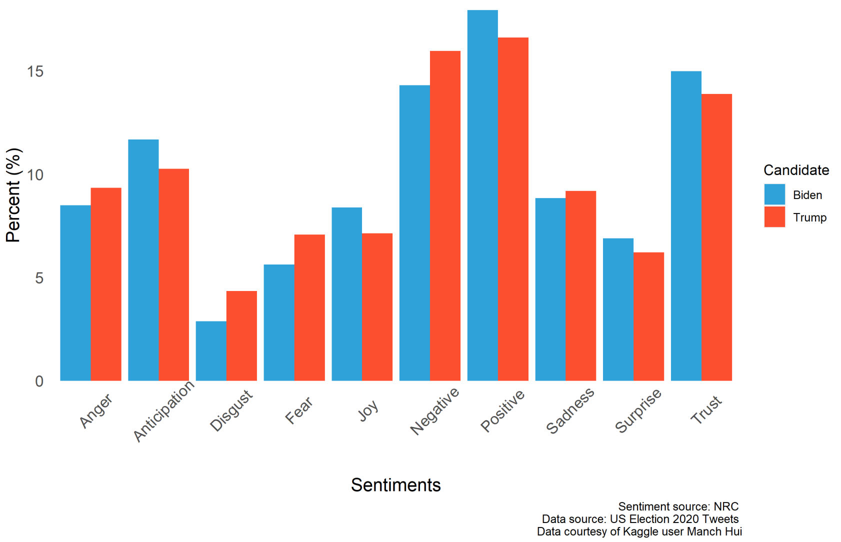
Twitter users are self-selected. According to an analysis by Pew Research, Twitter users are younger, more educated, and more likely to be Democrats than general public.

The tweets in my sample expressed more negative sentiments towards Trump. But this does not represent the general public's view. For the same reason, the tweets sentiment analysis does not enable us to predict the election results.

5.2 Future Research

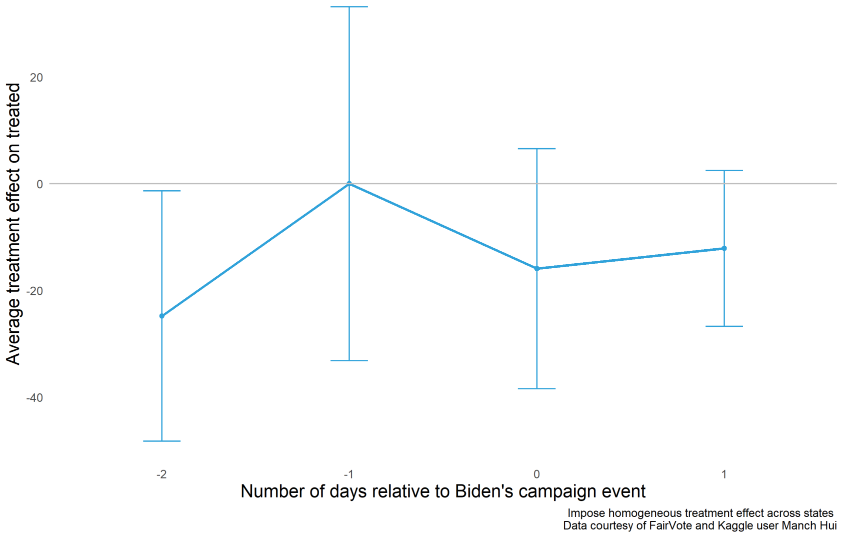
I'm interested in using network analysis to find out the key Twitter influencers in the 2020 Presidential Election, analyze the sentiments of their tweets related to the election, and visualize their geographic distribution and the network centered around them.

6 Figures



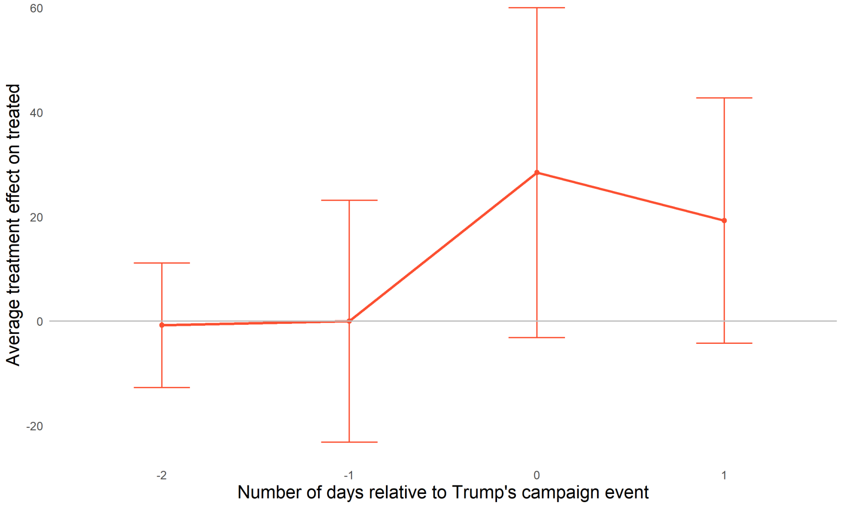
Sentiment source: NRC
Data source: US Election 2020 Tweets
Data courtesy of Kaggle user Manch Hui

Figure 1: Sentiment analysis using NRC emotion lexicon



Impose homogeneous treatment effect across states
Data courtesy of FairVote and Kaggle user Manch Hui

Figure 2: Average treatment effect over event time for Biden's twitter presence (impose homogeneous treatment effect across states)



Impose homogeneous treatment effect across states
Data courtesy of FairVote and Kaggle user Manch Hui

Figure 3: Average treatment effect over event time for Trump's twitter presence (impose homogeneous treatment effect across states)