A Quantitative Observation on the Spread of 2020 Election Misinformation and Political Book Discussions on Twitter

1st Pamela Pan
M Sci. Connective Media
Cornell Tech
New York City, United States
pp452@cornell.edu

2nd Andrea Wan
MBA
Cornell Tech
New York City, United States
jw2282@cornell.edu

3rd Courtney Beckham
M Eng. Computer Science
Cornell Tech
New York City, United States
cjb374@cornell.edu

***Abstract*—Despite the prevalence of bite-sized digital content, books continue to be a significant source of information for many people. We found a correlation between low mentions of political books in the US on Twitter and the spread of disinformation related to specific narratives and listed some suggestions for Twitter and the government to implement in the future.**

Introduction

In January 2021, following the defeat of Former President Donald Trump in the 2020 Presidential Election, an armed mob shockingly attacked the U.S. Capitol in the belief that the election results were fraudulent. The attack manifested months of online mis- and disinformation that eroded Americans' trust in the government. Observing the spread of misinformation in the 2020 Election, existing research tends to focus on investigating the online space of social media. However, less is known about what contributes to the spread of election-related misinformation in the real world.

To fill in the research gap, this project examines the potential correlation between the low discourse around political books and the spread of misinformation in various U.S. regions. Despite the prevalence of bite-sized digital content, books remain a significant source of information for many people in the United States. According to a 2021 survey distributed by the Pew Research Center, 75% of U.S. adults said they had read at least one book in the past year in any format, and Americans read an average of approximately 15 books during the past year [1]. Furthermore, research suggests that increased book readership positively influences the level of information literacy in readers [2].

This project aims to accomplish two goals: 1. Observe Tweets from areas where discussions about political books are happening versus where election-related misinformation narratives are spreading; 2. Draw insights from data analysis to help relevant stakeholders understand how to foster a healthier online information ecosystem. The timeframe of observation and analysis is the 10 days directly following the election date (November 4, 2020, to November 14, 2020).

Main Features

Our analysis has two stages: qualitative analysis and quantitative analysis. The qualitative analysis decided our approach to the misinformation narratives and then we performed data analysis on the data we gathered.

A. Qualitative Analysis

The qualitative methodology in this study is divided into two parts: 1. Obtaining a set of misinformation narratives from relevant credible sources; 2. Acquiring a list of books with politically-related content on the day after the 2020 Elections. These qualitative methods set out to operationalize Twitter data collection in further quantitative steps.

a) List of Misinformation Narratives:

To obtain a set of misinformation narratives from the 2020 Election, this project draws insights from a 2021 report called "The Long Fuse: Misinformation and the 2020 Election" by the Election Integrity Partnership (EIP) [3]. The EIP comprises organizations that specialize in the investigation of information dynamics in the U.S., including the Center for an Informed Public, Digital Forensics Lab, Graphika, and Stanford Internet Observatory.

By forming a collaborative, multi-stakeholder partnership, the EIP identified misleading and false claims that went viral, and information platforms that enabled these narratives to disseminate. In summary, the EIP outlines several kinds of misinformation narratives, including 1) Election-Theft Narratives; 2) Ballot-Related Narratives. From each category, researchers in this project then picked a couple of prominent trending Twitter hashtags – **Stop the Steal, Ballot Harvesting, Civil War** – during the 10 days directly following the election date (November 4, 2020, to November 14, 2020). These narratives are then used in the quantitative research as keywords to pull tweet data from.

b) List of Books with Political Content:

To learn about the books of popularity around the 2020 Election period, researchers utilized the New York Times Books API to get a list of 15 bestselling books on the day after the 2020 election (November 4, 2020). Based on

the content description of books, researchers then finalized a set of books that directly involve or indirectly allude to a range of issues such as political philosophy, history, and international relations. Below is the list of political books and their respective information:

1) *A Republic Under Assault* by Tom Fitton. The book argues that there is a "deep state" conspiracy against President Donald Trump and the American people. Fitton claims that the intelligence community, the media, and the political establishment are colluding to undermine the Trump presidency and the rule of law.
2) *How to be Anti-Racist* by Ibram X. Kendi. The book is a guide to becoming an anti-racist, which Kendi defines as actively opposing racism and taking measures to promote racial equity. The book argues that racism is not just an individual problem, but a systemic one, and that anti-racism requires both personal reflection and social action.
3) *Killing Crazy Horse* by Bill O'Reilly and Martin Dugard. It is a historical account of the conflicts between Native American tribes and the United States government in the 19th century, focusing on the final days of the Sioux Wars and the Battle of Little Bighorn. The book explores the political and social factors that contributed to these conflicts, as well as the actions taken by political leaders and policymakers during this period.
4) *One Vote Away* by Ted Cruz. The book argues that the Supreme Court is the most important battleground in American politics where the crucial political issues of our time are being decided with one vote. The book explores a number of recent Supreme Court cases and argues that the outcome of these cases could have significant consequences for the future of the country.
5) *Battle of Brothers* by Robert Lacey. The book is a biography of Prince William and Prince Harry, the two sons of Prince Charles and Princess Diana. It further explores the relationship between the two brothers and the tensions that have arisen between them over the years. The book also delves into the history of the British monarchy and the challenges faced by the younger generation of royals.

It is necessary to note that the book Caste by Isabel Wilkerson was not included despite being a political book. The book's name yielded ambiguous search results due to a substantial amount of references to the caste system in India.

B. Quantitative Analysis

a) Data Gathering and Preprocessing:

Our team utilized the Twitter API with Academic access to pull data for tweets between November 4, 2020, and November 14, 2020, with 20,752 users tweeting about our scoped political books, and 26,689 users tweeting about the three misinformation narratives. The following information fields were gathered from Tweets:

- author id
- context annotations
- conversation id
- entities
- geo
- id
- in reply to user id
- lang
- public metrics
- referenced tweets
- reply settings
- source

Utilizing the author ID, we queried the Twitter API for the following user information:

- created at
- description
- entities
- id
- location
- name
- public metrics
- username

The final dataset after dropping tweets that had no resolved location information had tweet counts as follows:

	Book Title or Misinformation Category	Count of Users
Books		
	A Republic Under Assault	137
	Battle of Brothers	21
	Killing Crazy Horse	52
	One Vote Away	20272
	How To Be An Antiracist	270
Misinformation		
	Civil War	1693
	Stop The Steal	2687
	Ballot Harvesting	22309

For additional analysis later on in the paper, we utilized a dataset from Kaggle about the winning political candidate in each county in the US for the 2020 presidential election as well as a dataset of polygons of US states.

C. Key Technical Challenges

The Twitter User API does not return any latitudinal or longitudinal information from a user's profile unless the user is the person making the API call. Due to this limitation, our team instead used the **location** field in the Twitter User API. The location field is a free text field where a user can input their location. This free text field has varying entries, with some locations being fictional and others being real places. However, for our analysis, we had to have geographic information about the user to analyze the volume of tweets and misinformation narratives in certain locations. To overcome this hurdle, our team utilized the Google Maps Geocoding API, which takes in a free text string and returns a list of candidates of potential locations. For our purposes, we kept the first candidate location as the location of the user and used it for further analysis.

Additionally, it is important to note that for pulling tweets with specific narratives, our team utilized string matching. However, this does not guarantee that the tweet is specifically talking about the misinformation narrative or the book title. It does mean that the full string is located in the tweet text, but the exact contents of the tweet cannot be assumed. Further deep work would need to be performed in regards to data preprocessing and sentiment analysis to weed out any extraneous tweets, but that was outside of the scope of this investigative process.

D. Data Analysis Results

Utilizing the Geopandas package, our team was able to join user locations to their respective state and to the dataset of 2020 presidential candidates. Using this cleaned dataset, for every state we were able to calculate the percentage of users in that state that were tweeting about the misinformation narratives versus the political books. We also were able to calculate the percentage of counties in the state that voted for the 2020 Democratic Presidential Candidate Joe Biden versus the percentage of counties in the state that voted for the 2020 Republican Presidential Candidate Donald Trump.

Utilizing the Python **corr()** function with the Pearson standard correlation coefficient, we analyzed correlations between political information and tweet information. As seen in Figure 1, there is a moderately positive correlation between users tweeting about misinformation and counties that voted Republican in 2020. There is also a negative correlation between users tweeting about books and counties that voted Republican in 2020. The inverse is true for Democrats, with a moderate correlation between users tweeting about books and counties that voted Democrat, and a negative correlation between users tweeting about misinformation and counties that voted Democrat. This suggests that misinformation is more likely to spread in areas that voted Republican in the 2020 election, while more book discourse is likely to spread in areas that voted Democrat. However, it is important to note that correlation does not equal causation and that many other factors and variables were excluded from the scope of this project that could be causing this correlation. Taking a deeper look into the relationship between reading habits and voting, you can see in Figure 2 that as reading habits increase, the number of Democratic wins in a state increases, and vice versa for Republicans. Then, taking a more geographic approach, you can see the correlation more clearly in the plots in Figure 3 that display the percentages across each state. In the map shown in Figures 4 & 5, the green circle represents discussions about books, while the red triangle represents the misinformation narratives. A drop-down menu is implemented for users to select the dataset of interest. Then, users can hover over the individual data points to learn about what specific narrative or book the user was tweeting about.

For deeper insight, we produced a functioning prototype of the data visualization website. The website features a Google map containing the geocoded information from the Google Maps API view centering the United States. It displays

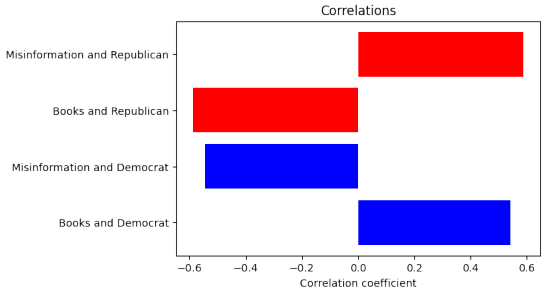

Fig. 1: Correlation between partisan bias, books, and misinformation

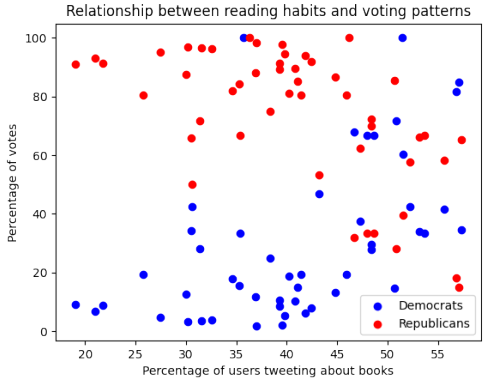

Fig. 2: Relationship between reading habits and voting patterns

the areas where people were tweeting about political books and misinformation narratives. For a speed-maximizing user experience, the website prototype only imported 15% of the tweets data, which were randomly and respectively drawn from two datasets (books and misinformation narratives).

Policy Suggestions

Throughout the semester, we have discussed changes that policymakers should implement to combat misinformation on social media platforms. Here are some steps we would encourage Twitter, other major social media platforms, and the government to implement in conducting research and promoting readership and education:

- Twitter should invest in a research and development initiative to prevent widespread misinformation. This initiative will:
  - Conduct further research on the major misinformation events. They could use existing data on the 2020 election to see when and where the misinformation about the election started to escalate.
  - Based on these findings, the company should dedicate a team to implement preventative measures like blocking scam and clickbait accounts, banning certain keywords when they became trending with misinformation, and enforcing content moderation.
  - Consider pouring funding into public education initiatives that increase the diversity of reading material

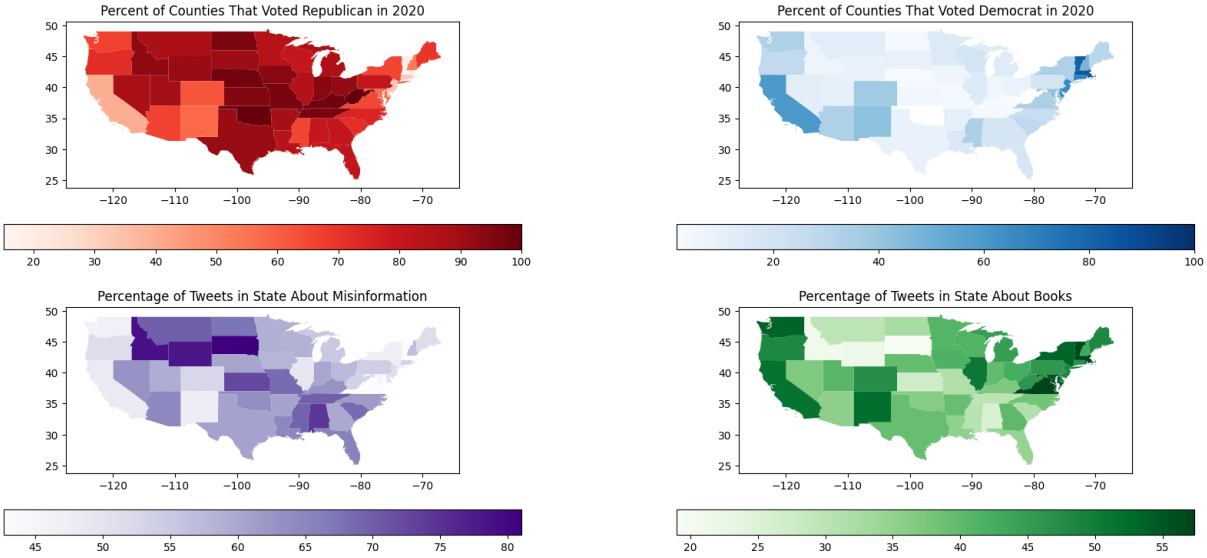

Fig. 3: Demographics of bias, books, and misinformation in the United States

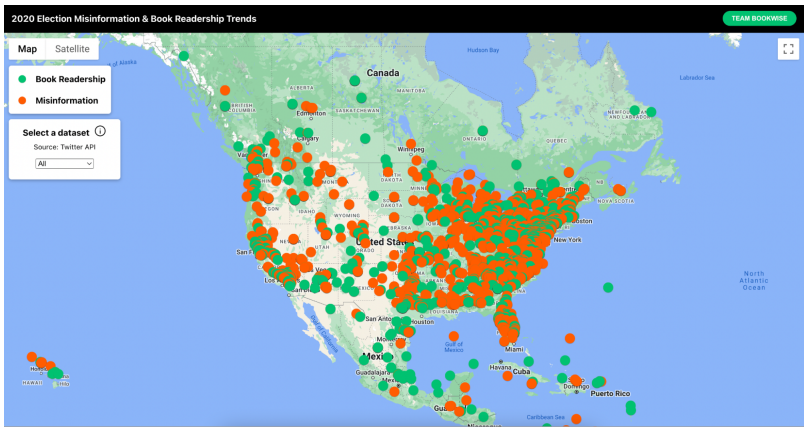

Fig. 4: Where the tweets are located in the United States

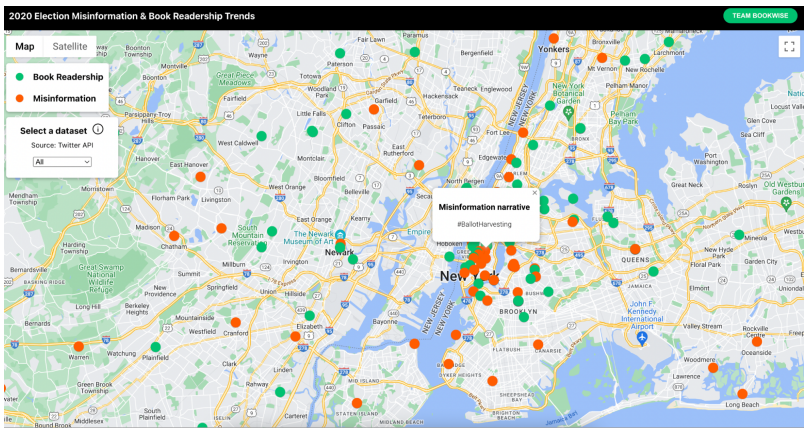

Fig. 5: Where the tweets are located in New York City

in areas that have been prone to misinformation spread.

- Like Twitter, other social media platforms have also been places where misinformation thrives. It is important to consider that there have been some changes in social media usage dynamics since 2020, especially after the

change in executive leadership at Twitter last November. Therefore, we recommend major or trending social media platforms to also invest in research initiatives that target the spread of misinformation. They could take the following steps:
  - Monitor platform integrity by analyzing data about elections or politically-related topics that are prone to widespread misinformation.
  - If sources of large-scale misinformation are identified, the platform should create a task force to address the integrity challenge by more actively implementing algorithmic measures, reducing the reach of false information, and educating users.
- On the other hand, the government should consider the following aspects:
  - Clarify leadership roles and publish bills to coordinate responses in the event of political incentives and misinformation spreading.
  - Dedicate a team in the justice system to investigate and follow up with major misinformation events and relevant stakeholders.
  - Local officials should be educated about past misinformation events and the consequences of such information on social media. They should work with the central government to approach greater integrity in future elections.
  - Start a public education initiative to tackle misinformation at its roots – information literacy. One such education initiative could target teaching K-12 students ways to identify suspicious information and clickbait on social media. The curriculum could also include materials for parents to protect their children from online misinformation, explicit language, or cyberbullying.

Theory of Change & Conclusion

Observing the results from this research, it is clear that the moderate correlation between political parties and online discourse requires further investigation and analysis. Our team recommends that Twitter further analyze potential causes of this correlation, and use the findings to develop policies around platform design in certain areas to discourage misinformation spread and encourage wider readership. As the U.S. social media landscape has been changing, the team also recommends other major social media platforms take steps to investigate platform integrity with regard to political misinformation and more actively address these issues. Lastly, the team suggests that the government should look into ways to enhance local and central policymakers to prevent the future spread of misinformation with political events. They might also look into initiatives that can fund communities to bolster public education and reading, such as investing in public libraries, book clubs, or reading initiatives. As the U.S. leans on the usage of social media platforms like Twitter, it is important that everyone on a platform is able to be part of the democratic activity and have a voice in public discourse. Utilizing insights from academic research efforts like ours, platforms can dive deeper into how much their platform contributes to belief in misinformation in low readership areas. In this way, platforms can start contributing to public education and take a more proactive approach to combating misinformation, rather than a reactive one.

Collaborators & Work Distribution

Most of the work is quite evenly distributed as we discussed the idea and approach together.

- Courtney Beckham:
  - Pulled all tweet posts for the misinformation narratives and NYT bestsellers list.
  - Pulled all tweet author information for the misinformation narratives and NYT bestsellers list.
  - Connected to and utilized the Google Maps API to resolve string locations to latitude/longitudes.
  - Utilized Geopandas to perform geospatial analysis of authors
  - Utilized Python correlation package to perform statistical analysis of authors
- Pamela Pan:
  - Conducted literature review on the topic of misinformation, book readership, and information literacy.
  - Sampled and cleaned geospatial data as preparation for visualization.
  - Created interactive data visualization web application using React.JS.
- Andrea Wan:
  - Discovered the political books angle and found data from New York Time developer API and filtered the top 10 books on the day after the 2020 election
  - Collected and pre-processed 2020 Election Candidate data from Kaggle
  - Formatted the paper and presentations and brainstormed policy suggestions

References

[1] Faverio, Michelle, and Andrew Perrin. "Three-In-Ten Americans Now Read E-Books." Pew Research Center, 6 Jan. 2022

[2] Karadeniz, Abdulkerim, and Remzi Can. "A Research on Book Reading Habits and Media Literacy of Students at the Faculty of Education." Procedia - Social and Behavioral Sciences, vol. 174, Feb. 2015, pp. 4058–4067.

[3] Observatory, Stanford Internet, et al. "The Long Fuse: Misinformation and the 2020 Election." FSI.

[4] Github Link to Twitter Data Analysis

[5] Github Link to Data Visualization