Note: This paper was written during the fall semester of my third year and was selected for presentation at the Tech Econference from a competitive pool. All data analysis was conducted using Python, utilizing libraries such as NLTK, pandas, geopandas, regex, matplotlib, and MATLAB, etc.

Shamayla Durrin Islam
Department of Economics
University of Toronto
January 2023

The Power of Twitter Data: Investigating the Relationship Between Social Media Sentiment and Political Trends

Abstract:

In this paper, we primarily explore the relationship between voting outcomes and Twitter sentiment by conducting lexicon-based sentiment analysis on 1.72 million tweets regarding Joe Biden and Donald Trump during the 2020 US Presidential Election. Through exploratory data analysis, we found that there were more tweets, as well as more unique user IDs, tweeting about Biden, prompting us to account for this bias in our analysis. Additionally, the distribution of tweets across counties was highly uneven, so we focused our final analysis at the state level.

To predict election outcomes, we compared the mean sentiment scores for each candidate in each state, assuming that the candidate with the higher mean sentiment score would win. Using this approach, we successfully predicted the outcome in 32 out of 50 states and 8 out of 11 battleground states. However, due to the engagement bias skewed toward Biden, we employed net sentiment (the proportion of positive sentiment minus negative sentiment) for our final regression analysis, which revealed that net sentiment was statistically significant at the 0.05% level for both candidates. We found out that demographic information of states had association with sentiments in county level but were not significant at the state level. We acknowledge the limitations of this analysis due to the disproportionate use of Twitter, which could affect the generalizability of our findings.

Introduction

Twitter has become a prominent platform for political discourse, especially during elections, where a significant portion of public sentiment is shared and analyzed. Approximately one-third of US adults' tweets are political, and this engagement increases during elections.[1] This research explores whether Twitter data, despite its limitations, can serve as a reliable predictor of election outcomes, specifically focusing on the 2020 US Presidential Election. By analyzing sentiment in tweets, we aim to examine the relationship between online sentiment and actual vote shares, alongside other election-related variables such as voter turnout and demographic factors.

In recent years, much research has focused on using Twitter data for political forecasting. In recent years, there has been a significant amount of research conducted to analyze the sentiment of tweets. For instance, O'Connor et al. used tweets from the 2012 presidential election and conducted a sentiment analysis to compare their findings with polling data and forecast future polls.[2] Allcott and Gentzkow (2017) explored how fake news on social media, particularly during the 2016 U.S. election, influenced voting patterns, showing that the widespread dissemination of false information likely impacted voter behavior.[3] Tumasjan et al. predicted election results by comparing the share of attention the political parties receive on Twitter.[4] Brandon Joyce and Jing Deng used a lexicon and Naive Bayes algorithm to compare sentiments of the 2016 US election to polling data.[5] Building on this work, we focused on making predictions about the 2020 US election by using sentiment analysis on tweets to assess political sentiment across US states.

---

[1] Duggan, M., & Smith, A. (2022). Politics on Twitter: One-third of tweets from U.S. adults are political. Pew Research Center.
[2] O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series." *ICWSM*, vol. 11, no. 122-129, pp. 1–2, 2010.
[3] Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives, 31*(2), 211–236. https://doi.org/10.1257/jep.31.2.211
[4] Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. International AAAI Conference on Weblogs and Social Media (ICWSM).
[5] B. Joyce and J. Deng, "Sentiment analysis of tweets for the 2016 US presidential election," *2017 IEEE MIT Undergraduate Research Technology Conference (URTC)*, Cambridge, MA, USA, 2017, pp. 1-4, doi: 10.1109/URTC.2017.8284176.

In this paper, we applied sentiment analysis to over 1.7 million tweets mentioning Donald Trump and Joe Biden. Our analysis assessed whether sentiment scores could predict vote shares, with a focus on the mean sentiment scores and the proportion of positive and negative tweets for each candidate. By comparing these sentiment metrics, we achieved a 64% success rate in predicting state-level outcomes and a 67% success rate in battleground states.

Additionally, we explored the role of engagement bias, noting that there were more tweets and unique users tweeting about Biden. To address this, we introduced a proportional sentiment metric and a prediction score, which subtracts the negative tweet share from the positive tweet share. This proved to be a strong predictor of vote share in our regression models. Finally, we explored the relationship between sentiment, voter turnout, and demographic factors, controlling for income and race, to gain further insights into election dynamics.

2. Data

2.1. Data Source

The primary dataset is the US Election 2020 Tweets from Kaggle, collected via the Twitter API by Manchunhui. The dataset spans from September 13th to November 8th, 2020, covering the election period, and consists of approximately 1.72 million tweets related to Donald Trump and Joe Biden, tagged with relevant hashtags (#DonaldTrump, #JoeBiden). Each row represents a tweet about a candidate with associated metadata, such as tweet ID, timestamp, location (state, country), and text. We focused on US-based tweets to analyze political sentiment in the country. Election data, including vote counts and turnout rates, was sourced from the United States Elections Project and state median incomes were scraped from Wikipedia. Demographic data,

including the percentages of White, Black, and Hispanic populations, was sourced from US Census data.

2.2. Data Cleaning and Pre-Processing

The dataset was filtered to include only tweets from October 15th to November 2nd, 2020. We removed entries without location data, standardized US state names, and excluded tweets from outside the United States. Duplicate tweets were dropped based on tweet IDs, and the tweet text was cleaned by removing URLs, punctuation, emojis, and stopwords using TextBlob. After cleaning, tweets without content were excluded from the analysis.

2.3. Sentiment Analysis

Sentiment analysis was performed using **TextBlob**, which assigns a sentiment polarity score ranging from -1 (negative) to +1 (positive) to each tweet. These sentiment scores were aggregated at the state level to evaluate the overall sentiment for each candidate, and the results were compared to actual voting outcomes in the 2020 election.

2.4. Fixing Geographic Level of Analysis

To determine the geographic level of analysis, we examined the distribution of tweet counts at both the state and county levels. County-level data was highly uneven, with observations from only 876 out of 3,143 US counties, and 50% of counties had fewer than 15 tweets. In contrast, the state-level distribution was more balanced, with a median of 1,768 tweets per state. Given the limited Twitter engagement in many counties and the more even tweet distribution at the state level, we chose states as the geographic focus for our analysis. The following plot illustrates the

tweet count distribution for counties and states. The median of tweets counts in counties is close to 0.

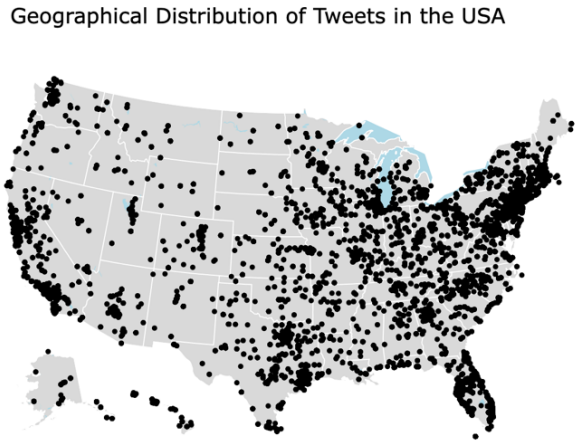

Figure 1: Map Showing the Location of Tweets in USA

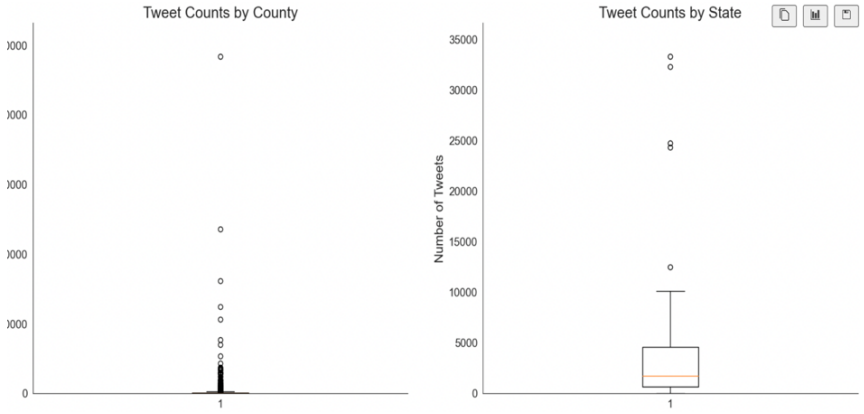

Figure 2: Boxplot of Tweet Counts in US Counties and States.

*Note: The map shows the geographical distribution of tweets from the dataset across the USA, illustrating a disproportionate spread with many counties reporting zero tweets. The boxplot of tweet counts by county reveals that the median is close to zero, indicating that most counties have very few tweets, while a few outliers have significantly higher counts. The boxplot by state presents a more balanced distribution, with certain states having notably higher tweet counts.

3. Simple Prediction of Election Result Using Mean Sentiment Scores of Tweets

After calculating sentiment scores for each tweet, we assessed whether the mean sentiment score within individual states could serve as a reliable predictor of election results. We plotted the mean sentiment scores of Trump and Biden across US states and compared them with their respective vote shares. Our analysis revealed a clear positive correlation between Trump's mean sentiment score and his vote share, with the correlation being more pronounced for Trump than for Biden. This observation is critical as it suggests that mean sentiment scores may be useful for predicting election outcomes.

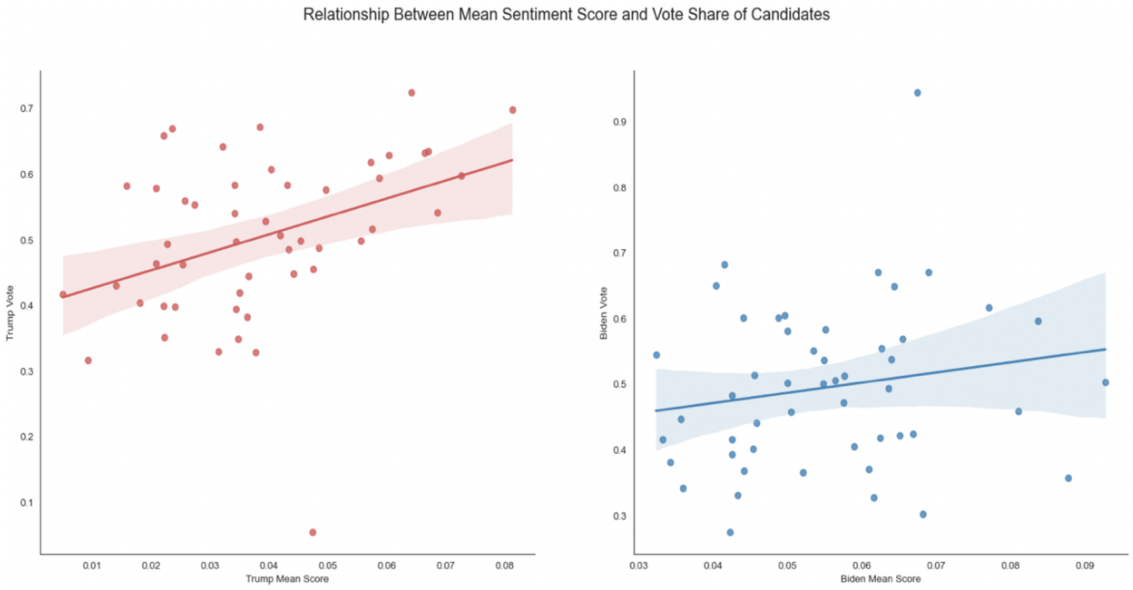

Figure 3: Scatterplot of Mean Sentiment Score for Candidates and Vote Share

*Note: The above scatterplots show the positive association between tweeter sentiment about a candidate and their vote share in that state.

The following plot demonstrates the distribution of mean sentiment scores for both candidates across states. The median value of Biden's mean sentiment score is notably higher than Trump's, and the symmetry of the box plots indicates that the mean sentiment scores are close to their respective medians. This suggests that, on average, public sentiment was more favorable toward Biden across all states. Additionally, the greater spread of Trump's mean sentiment scores, compared to Biden's, may explain the stronger correlation observed between Trump's sentiment scores and vote share.

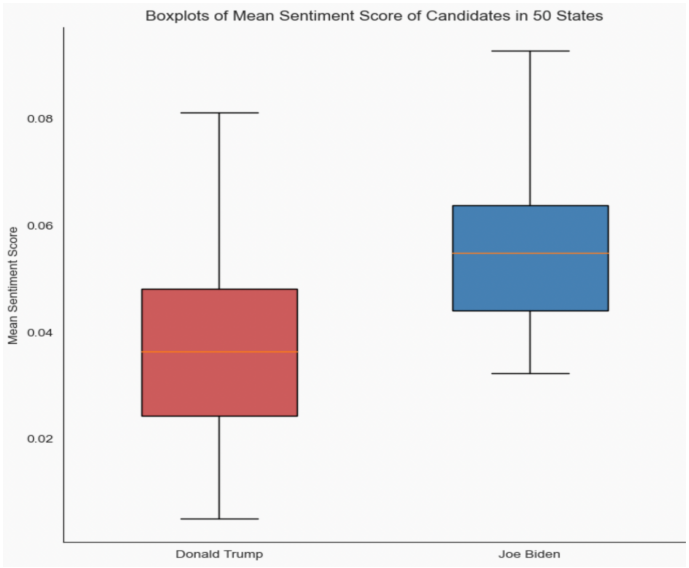

Figure 4: Boxplot of Distribution of Mean Sentiment Score for Candidates in US states.

Given this, we hypothesized that states where a candidate had a higher mean sentiment score would likely reflect stronger public support and, therefore, result in a higher vote share for that candidate. Using this hypothesis, we successfully predicted the election outcome in **32 out of 50 states** and **8 out of 12 battleground states**. The term "battleground state" refers to states where the outcome of an election is uncertain, often referred to as "swing states." In the states where our model failed to predict the outcome, many had very narrow margins of victory, indicating that the model's failure may not be entirely attributable to its predictive ability.

The following plot shows a comparison of mean sentiment scores for candidates in battleground states. Our analysis shows that, with the exception of **Arizona**, **North Carolina**, **Ohio**, and **Texas**, the candidate with the higher mean sentiment score won in most battleground states. This led to a 67% success rate in predicting outcomes in battleground states.

In **Arizona**, Biden won by a narrow margin of 0.3% (49.4% vs. Trump's 49.1%), and this slim margin is reflected in the similarly close sentiment scores. In **Texas**, Trump won by a margin of 5.6%, with a relatively small difference in mean sentiment scores as well. However, in **Ohio**, there was a notable discrepancy between the sentiment scores and the outcome—Trump won by 8.1%, despite our model predicting otherwise based on sentiment. Despite these exceptions, our model's 8 out of 12 successes in battleground states suggest that mean sentiment scores offer a promising variable for predicting election results in the future.

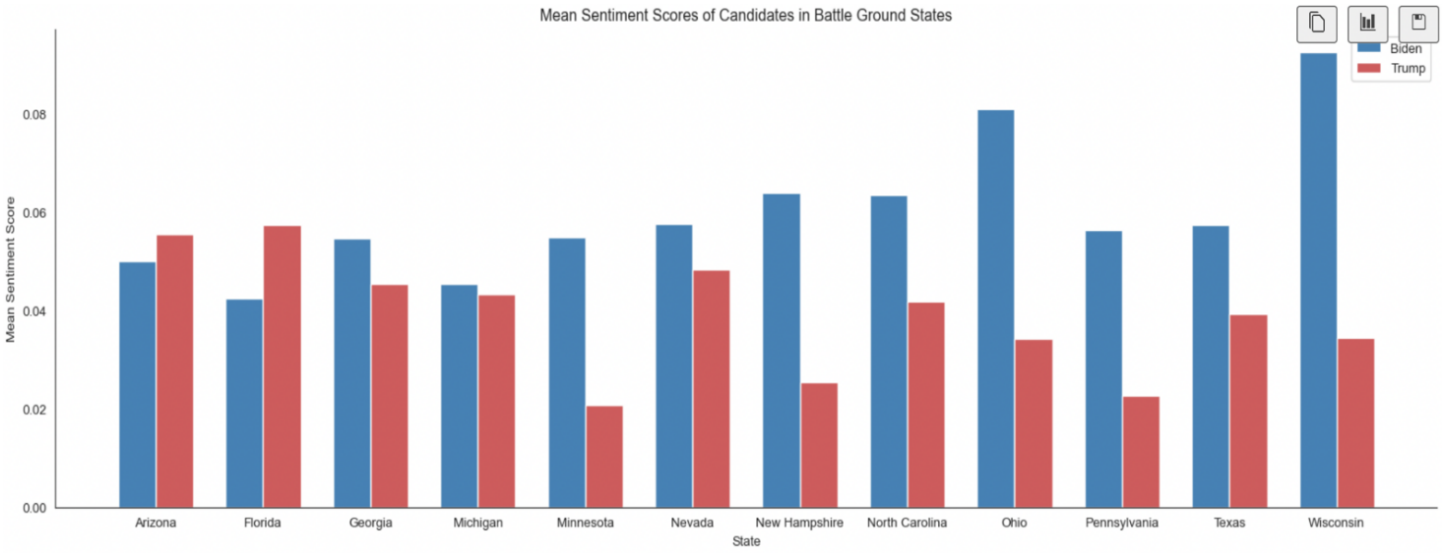

Figure 5: Mean Sentiment Score of Tweets about Candidates in Battleground States

*Note: This bar chart compares the mean sentiment scores of tweets for Biden and Trump in key battleground states during the 2020 U.S. Presidential Election. The analysis shows that in 8 out of 12 states, the candidate with the higher mean sentiment score also had a higher vote share and ultimately won the state. This correlation suggests that sentiment analysis of tweets can provide valuable insights into predicting election outcomes, particularly in closely contested states.

4. Endogeneity Concerns and Engagement Bias

In this section, we address engagement bias observed in our dataset and examine additional variables that could influence both Twitter sentiment and vote share. Omitting these variables from our regression model may lead to endogeneity issues. Specifically, voter demographics,

state-level political climate, and media exposure may affect both Twitter sentiment and election outcomes. Controlling for these factors helps mitigate biases and improve the accuracy of our analysis.

5.1. Engagement Bias

In the overall dataset, we observed that there were more tweets about Joe Biden than Donald Trump. This prompted us to further explore the number of unique user IDs tweeting about each candidate. Our analysis revealed that, like the tweet count, there were also more unique users tweeting about Biden than Trump, as shown in the bar plot below.

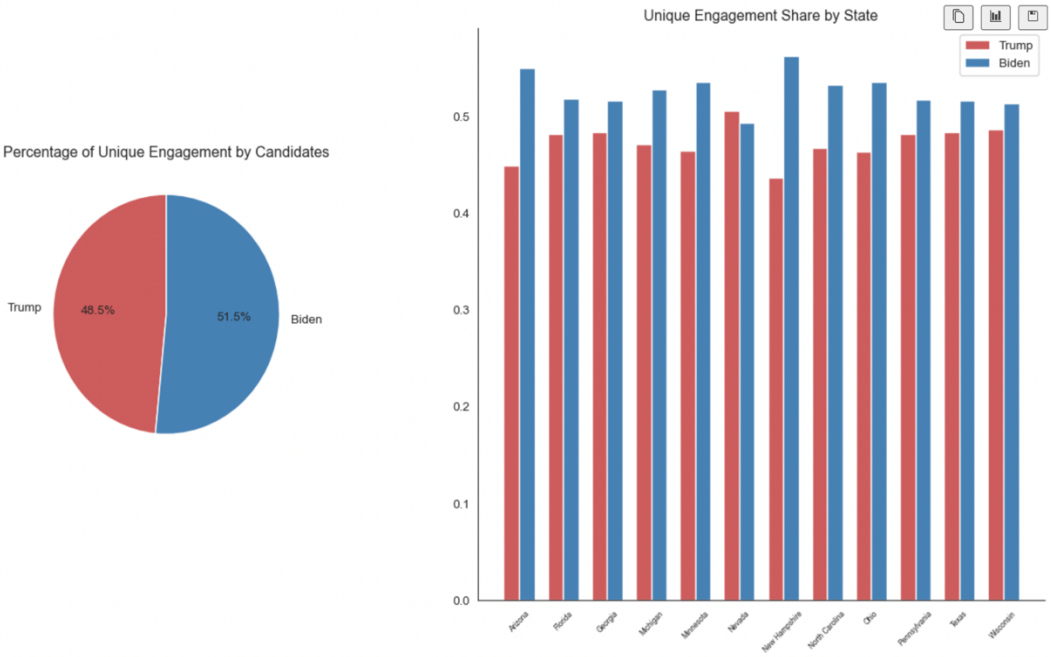

Figure 6: Percentage of Twitter User who Tweet About the Candidates in US and in Battleground States

*Note: This plot compares the percentage of Twitter users tweeting about Donald Trump and Joe Biden in battleground states. The chart illustrates that overall, and in most states, engagement regarding Biden is higher than Trump, indicating a greater volume of discussion and sentiment around Biden during the 2020 U.S. Presidential Election. This trend may reflect the broader public focus on Biden in these critical states.

It makes sense because existing study shows that Twitter users in the US are more likely to be Democrat than Republican. [6]Given this, we recognized that using the mean sentiment score in our model might introduce bias toward Biden, as the higher number of tweets and unique users mentioning him could skew the results. To address this potential bias and better reflect the correlation with vote share, we decided to use a proportional metric. Specifically, we calculated the proportion of positive tweets (i.e., the number of positive tweets divided by the total tweets for each candidate) and the proportion of negative tweets. These proportions offer a more balanced approach to understanding public sentiment and its correlation with the actual vote share.

When we plotted the scatter plots of the positive and negative tweet shares against vote share, we observed a clear positive association between the share of positive tweets and Trump's vote share. Conversely, there was a negative association between the share of negative tweets and Trump's vote share. These findings support our hypothesis that a higher proportion of positive tweets for a candidate corresponds to a higher vote share, while more negative tweets are linked to a lower vote share for that candidate.

Building on this, we introduced a final prediction score by subtracting the negative tweet share from the positive tweet share. The reasoning behind this approach is that if a candidate has a higher positive tweet share and a lower negative tweet share, they are more likely to win in that state. The scatter plot shows a strong association between this prediction score and vote counts,

[6] Perrin, A. (2019). Americans are changing their relationship with Facebook. Pew Research Center

with the strongest relationship observed so far. This result suggests that the prediction score may offer the most accurate method for predicting election outcomes.

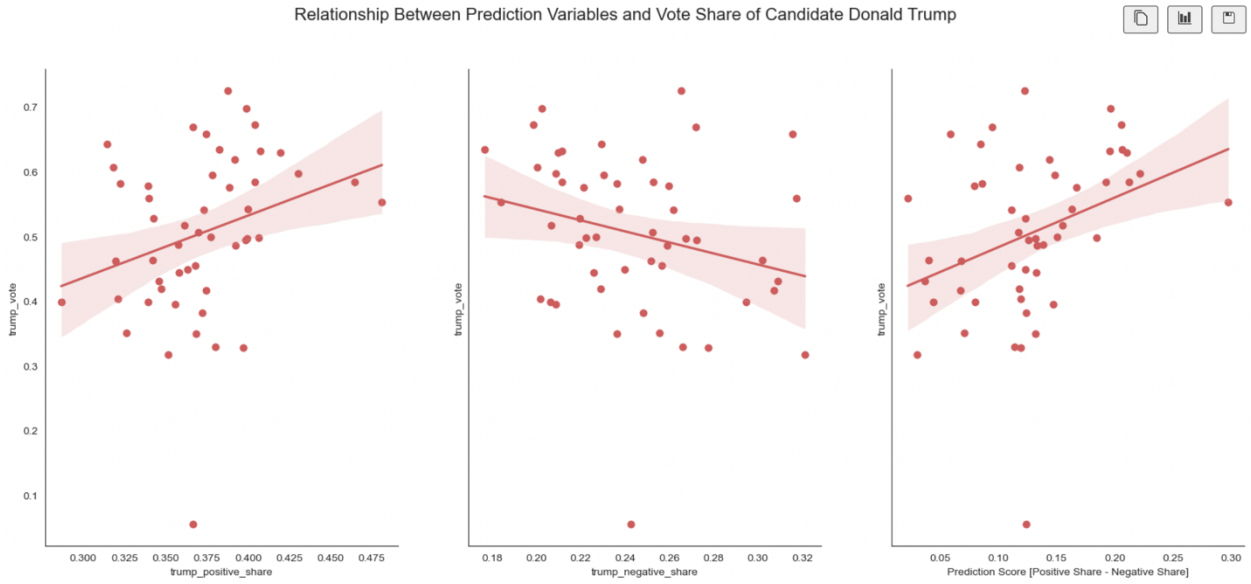

Figure 7: Scatterplot of Different Candidate Variables and Vote Share of Trump.

*Note: The first scatter plot illustrates the relationship between Donald Trump's vote share and the positive share of tweets about him, showing a clear positive correlation. The second scatter plot demonstrates the relationship between Trump's vote share and the negative share of tweets, exhibiting a negative correlation. The third scatter plot compares Trump's vote share with the net sentiment (positive share minus negative share of tweets), showing a positive relationship. Together, these plots highlight the impact of sentiment in tweets on Trump's vote share, with more positive sentiment corresponding to higher vote shares and more negative sentiment correlating with lower vote shares.

The plot below shows the association between the positive tweet share, negative tweet share, and prediction score with Biden's vote share. While the trend is less distinct than with Trump, there is still somewhat positive association between Biden's positive tweet share and his vote share, and a negative association with his negative tweet share. The prediction score also shows a relatively stronger association with Biden's vote share, indicating that in states where Biden had a higher proportion of positive tweets and fewer negative tweets, he tended to perform better in the election.

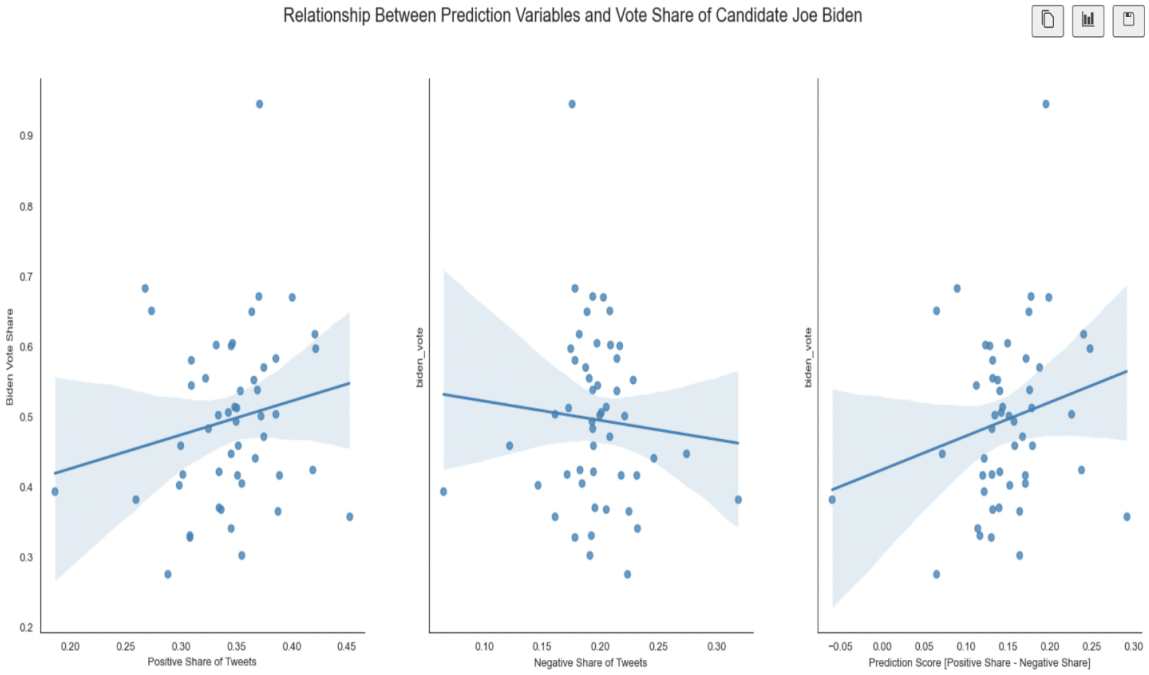

Figure 7: Scatterplot of Different Candidate Variables and Vote Share of Biden.

5.2. Controlling for Demographic Characteristics

In our dataset, we identified the fraction of positive and negative tweets for each candidate across counties and linked this with census data to determine the majority demographic groups—primarily White, Hispanic, and Black—in those counties. We then analyzed the distribution of negative and positive tweet shares for each demographic group. The goal was to explore how Twitter sentiment aligns with well-known demographic leanings, such as Democratic support among Black and Hispanic populations and Republican support among White populations.

Researchers have shown that Black and Hispanic Americans are more likely to identify with the Democratic Party than with the Republican Party.[7]

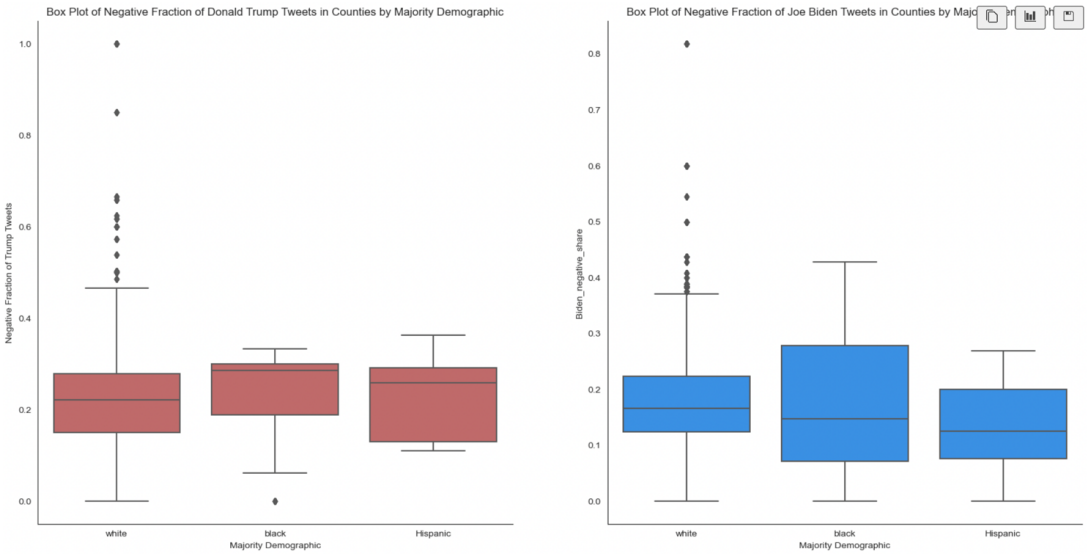

Figure 2: Distribution of Proportion of Negative Tweets In Black, White, Hispanic Majority Counties.

*Note: The boxplot illustrates the mean sentiment scores about Trump and Biden in counties with a Black, Hispanic, or White majority. In White-majority counties, there tends to be less negative sentiment about Trump, while in Hispanic and Black-majority counties, there is higher negative sentiment toward him. A similar pattern is observed for Biden, where White-majority counties show more negative sentiment about him compared to Black and Hispanic-majority counties. This indicates that racial demographics play a significant role in shaping public sentiment toward the candidates.

The first subplot shows the distribution of negative tweet shares about Trump in counties with Black, Hispanic, and White majorities. Counties with a Black majority exhibit the highest median negative sentiment toward Trump, followed by Hispanic-majority counties. White-majority counties, by contrast, had the lowest median negative tweet share, reflecting a relatively more favorable sentiment toward Trump.

[7] Race and the American Electorate: An Examination of the Relationship between Racial Identity and Political Behavior" Michael C. Dawson, Annual Review of Political Science,1994 .

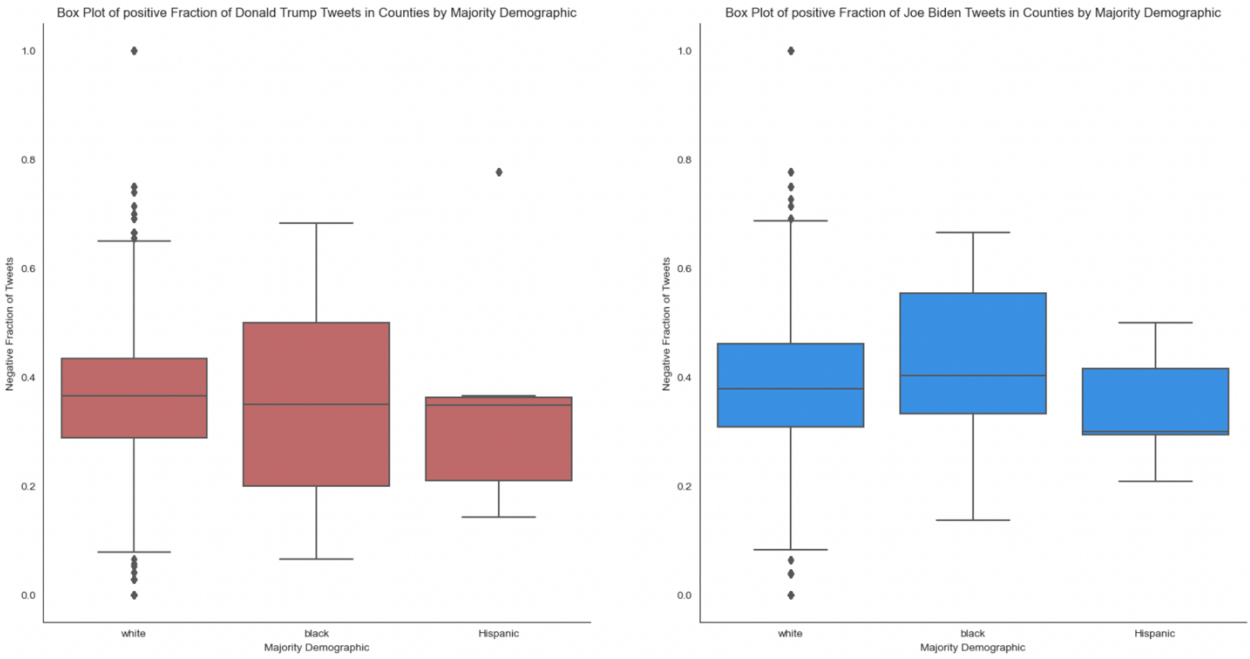

Figure 3: Distribution of Proportion of Positive Tweets In Black, White, Hispanic Majority Counties.

The second subplot illustrates the distribution of negative tweet shares for Biden in the same demographic groups. Here, counties with Black and Hispanic majorities show lower negative sentiment toward Biden than White-majority counties. Notably, the median negative tweet share for Biden is lowest in Hispanic-majority counties, indicating stronger support for him in these areas. These findings are consistent with broader research showing that Black and Hispanic Americans tend to align more with the Democratic Party.

Therefore, when analyzing the correlation between Twitter sentiment and vote share, it is crucial to control for these demographic factors. As our findings suggest, different racial and ethnic groups exhibit distinct sentiment trends, which could bias the results if not accounted for. By including demographic variables in our model, we can better isolate the true relationship between

sentiment expressed on Twitter and the vote share for each candidate, ensuring a more accurate and unbiased analysis.

5.3. Controlling for Median Income

While plotting median income of states and candidate vote share we observed a negative relationship between Donald Trump's vote share and median income, where states with lower median incomes were more likely to vote for Trump, while higher-income states leaned towards Biden. This trend aligns with broader research showing that lower-income individuals, particularly in rural areas, have increasingly supported Republican candidates like Trump due to factors such as economic insecurity, opposition to globalization, and cultural identity concerns . Trump's appeal to working-class voters, especially in regions affected by deindustrialization, has been widely documented as a significant factor in his electoral success.

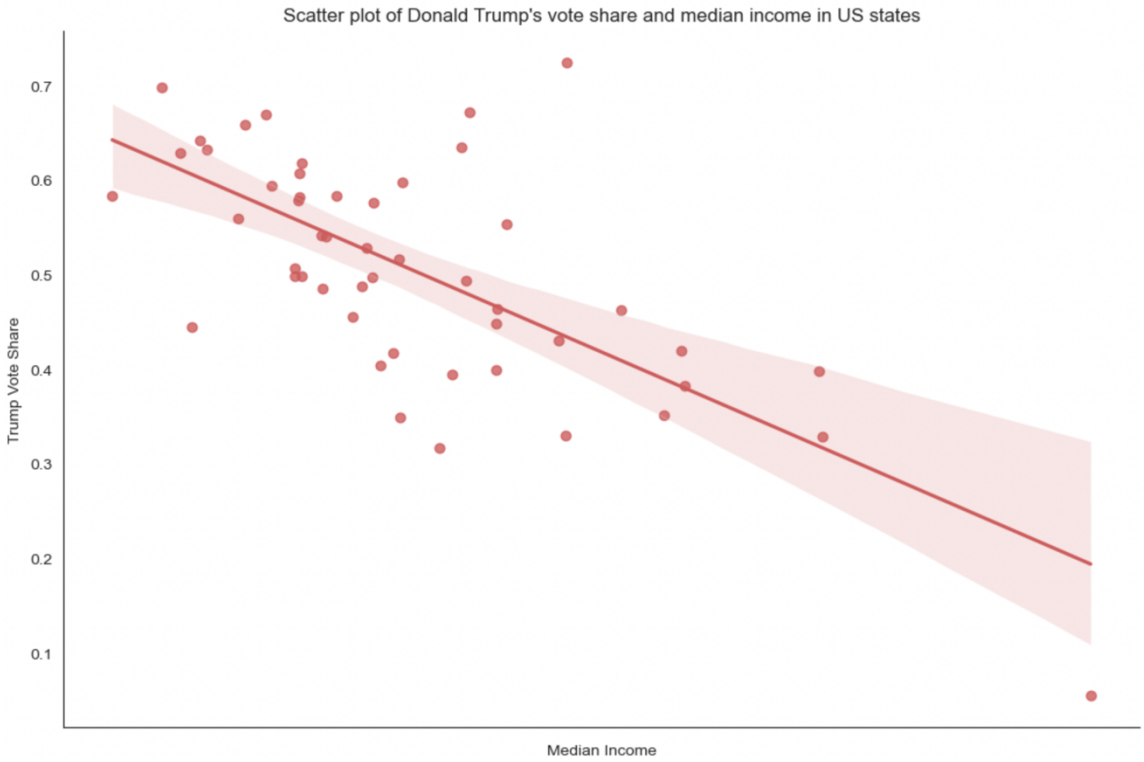

Figure 10: Scatterplot of Trumps Vote Share in US States and the Median Income of those States

Looking at the two graphs below, we can observe a clear trend where states with lower median incomes tend to have a higher vote share for Donald Trump. In the first map showing Trump's vote share, many states with significant support for Trump, such as **West Virginia, Arkansas, and Mississippi**, also appear in the second map as states with lower median incomes. Conversely, states with higher median incomes, such as **California, New York, and Massachusetts**, show a lower vote share for Trump. This negative association between median income and Trump's vote share suggests that lower-income states were more likely to support him, while wealthier states leaned toward his opponent. This pattern reflects the broader demographic and economic appeal that Trump has among working-class voters, particularly in states with lower average incomes.

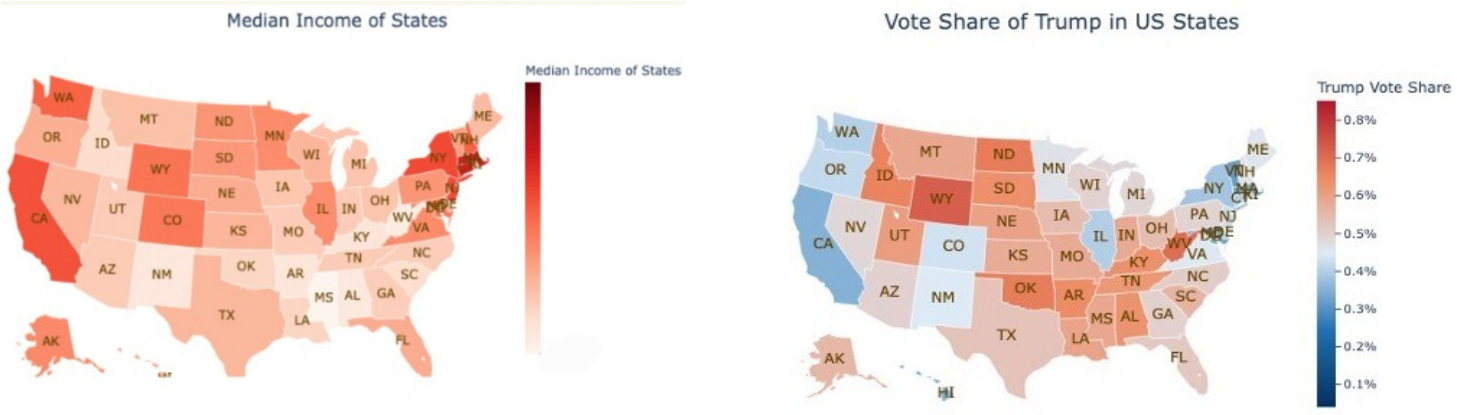

Figure 11: Heat Maps of Median Income and Vote Share of Candidates in US States.

Note: By comparing the two heat maps of the United States, it is evident that in regions where median income is higher, Biden tends to have a higher vote share, while in areas with lower median income, Trump garners more support. This suggests a clear relationship between income levels and voting patterns during the 2020 U.S. Presidential Election.

6. Regression and Results

To confirm whether sentiment of tweets have association with a particular candidate's vote share we ran few OLS regressions to check whether the coefficients are statistically significant. We controlled for median income, percentage of white, black and Hispanic population.

To investigate whether sentiment is associated with a candidate's vote share, we conducted a regression analysis where the dependent variable was the vote share for each candidate. Independent variables included the log of median income, the percentage of Hispanic, White, and Black populations, as well as positive and negative sentiment shares. Initial results indicated that the demographic variables were not statistically significant, and only Trump's negative sentiment share showed a significant relationship with his vote share.

To refine our analysis, we regressed vote shares on the prediction score which we introduced before, calculated by subtracting the negative share of tweets from the positive share for both candidates. When we re-ran the regression using this variable, we found that it was statistically significant for both Trump and Biden, suggesting that the overall balance of positive and negative sentiment is a more reliable predictor of vote share than individual sentiment measures alone. The model is depicted below:

$$\boldsymbol{Trump\ Vote\ Share} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \times (\boldsymbol{Prediction\ Score}) + \boldsymbol{\beta}_2 \times \boldsymbol{log}(\boldsymbol{Income}) + \boldsymbol{\beta}_3 \times (\boldsymbol{Percent\ White}) + \boldsymbol{\beta}_4 \times (\boldsymbol{Percent\ Black}) + \boldsymbol{\beta}_5 \times (\boldsymbol{Percent\ Hispanic}) + \boldsymbol{\varepsilon}$$

The objective is to estimate the values of the coefficients that minimize the sum of the squared errors between the actual and predicted values of the dependent variable which is Trump's Vote Share. This can be expressed as the following function:

$$\min_{\beta_0,\beta_1,\beta_2,\beta_3} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

The corresponding output is:

OLS Regression Results

Dep. Variable:	trump_vote	R-squared:	0.626
Model:	OLS	Adj. R-squared:	0.583
Method:	Least Squares	F-statistic:	14.72
Date:	Sat, 21 Sep 2024	Prob (F-statistic):	1.75e-08
Time:	00:12:06	Log-Likelihood:	66.092
No. Observations:	50	AIC:	-120.2
Df Residuals:	44	BIC:	-108.7
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.7250	0.946	4.996	0.000	2.819	6.631
posneg	0.4200	0.186	2.262	0.029	0.046	0.794
log_income	-0.3928	0.077	-5.128	0.000	-0.547	-0.238
Percent White	0.0021	0.001	1.435	0.158	-0.001	0.005
Percent Black	-0.0007	0.002	-0.398	0.693	-0.004	0.003
Percent Hispanic or Latino	-0.0008	0.001	-0.636	0.528	-0.003	0.002

Omnibus:	2.075	Durbin-Watson:	2.272
Prob(Omnibus):	0.354	Jarque-Bera (JB):	1.261
Skew:	0.099	Prob(JB):	0.532
Kurtosis:	3.752	Cond. No.	7.94e+03

The OLS regression analysis of Trump's vote share yielded an R-squared of 0.626, meaning 62.6% of the variation in vote share is explained by the model. The F-statistic (14.72, $p < 0.001$) indicates strong model significance. Trump's net sentiment ($\beta = 0.420$, $p = 0.029$) positively influenced his vote share, while log of median income had a significant negative effect ($\beta = -0.393$, $p < 0.001$).

Demographic factors, such as the percentages of White, Black, and Hispanic/Latino populations, were not statistically significant. This suggests that demographic information may not be as useful at the state level but could be more informative at the county level. Overall, sentiment and income are key predictors of Trump's vote share in this model.

We had a similar model for Biden:

$$\begin{aligned}\boldsymbol{Biden\ Vote\ Share} &= \boldsymbol{\beta_0} + \boldsymbol{\beta_1} \times (\boldsymbol{Prediction\ Score}) + \boldsymbol{\beta_2} \times \boldsymbol{log}(\boldsymbol{Income}) \\ &+ \boldsymbol{\beta_3} \times (\boldsymbol{Percent\ White}) + \boldsymbol{\beta_4} \times (\boldsymbol{Percent\ Black}) \\ &+ \boldsymbol{\beta_5} \times (\boldsymbol{Percent\ Hispanic}) + \boldsymbol{\varepsilon}\end{aligned}$$

The OLS regression analysis for Biden's vote share shows an R-squared of 0.59, meaning 59% of the variation in Biden's vote share is explained by the model. The F-statistic (13.21 $p < 0.001$) indicates the model is statistically significant overall.

The net sentiment score for Biden (labeled as posnegb) has a positive coefficient ($\beta = 0.380$, $p = 0.03$), indicating that net positive sentiment toward Biden significantly increases his vote share.. We see that in this model as well the above the demographic variables are not significant.

7. Future Work:

In the 2020 US election, Biden's victory was significantly influenced by voter turnout. According to the U.S. Elections Project, a record-breaking high turnout of 66.4% of those eligible to vote was achieved in the 2020 election.[8] This high turnout was driven by several factors, including the highly polarized political climate, increased attention on voting methods and access, and the COVID-19 pandemic, which motivated many voters to vote early or by mail.

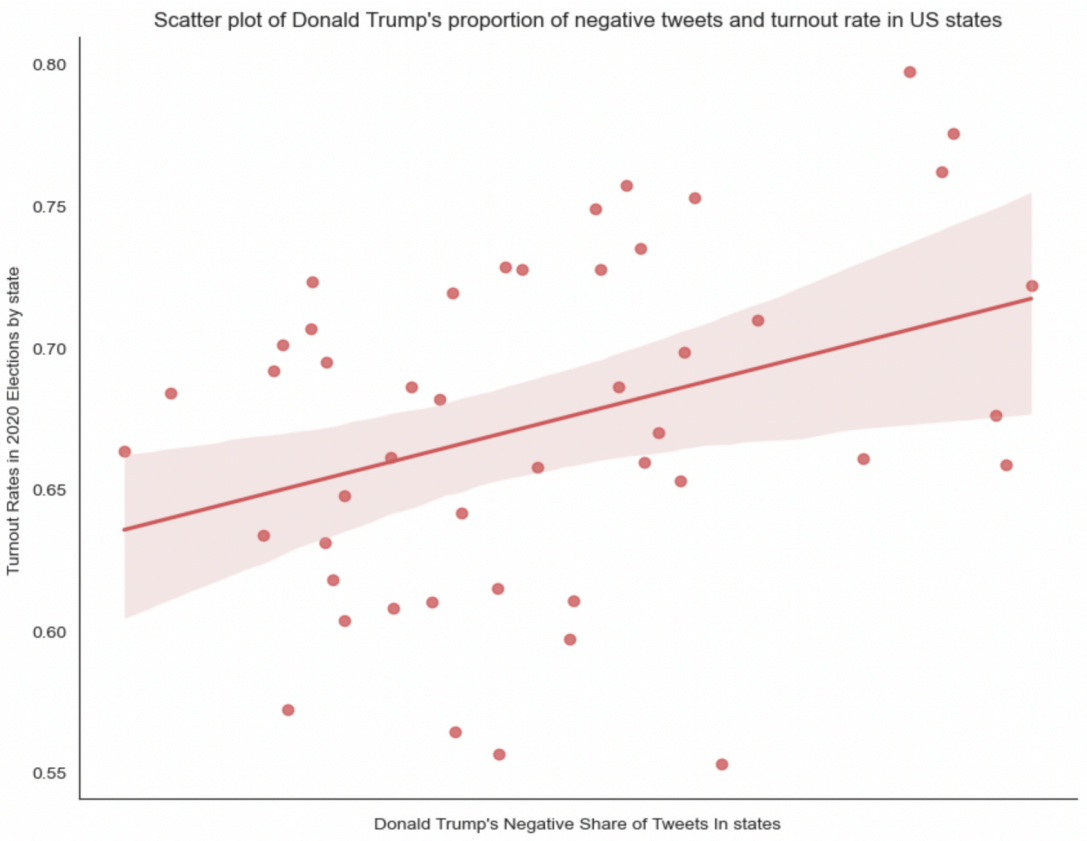

Figure 12: Scatterplot of Negative Share of Tweets and Voter Turnout Rates in US States.

*Note: The 2020 U.S. Presidential Election saw a historically high voter turnout rate. We hypothesized that the increased negative sentiment toward Trump contributed to this high turnout. To test this, we plotted the fraction of negative tweets about Trump in each state against the voter turnout rate. The plot reveals a positive association, suggesting that higher negative sentiment toward Trump may have motivated more people to participate in the election.

[8] The 2020 U.S. Presidential Election: High Turnout, Partisan Polarization, and a Pandemic by Michael McDonald and Robin Best (2021)

Our hypothesis was that in states where individuals expressed more negative sentiment about Trump on social media platforms, there is a higher likelihood of greater voter turnout. The rationale behind this hypothesis was that negative sentiment towards a political candidate might have had served as a motivating factor for individuals to participate in the election and cast their vote for the other candidate to remove Trump from power. Therefore, by examining the scatter plot between negative sentiment towards Trump in Twitter and voter turnout in various states, we noticed a positive association. Future work could be looking deeper into this association of tweeter sentiment and voter turnout.

Conclusion:

Our analysis of Twitter sentiment data during the 2020 US Presidential Election provides significant insights into the relationship between online sentiment and actual voting outcomes. Using data from 1.72 million tweets related to Donald Trump and Joe Biden, we explored whether public sentiment expressed on Twitter could be a reliable predictor of state-level vote shares. We employed several sentiment analysis techniques, primarily focusing on the mean sentiment scores and the proportion of positive and negative tweets for each candidate. Our results indicate a positive correlation between Trump's mean sentiment score and his vote share, with this relationship being stronger for Trump than for Biden. By comparing which candidate had the higher mean sentiment score, we accurately predicted the election outcome for 64% of all states and 67% of battleground states.

We also noticed that engagement regarding Biden was higher; therefore, for our final regression analysis, we used the proportion of positive and negative tweets to better account for this disparity. Additionally, we introduced the concept of a prediction score, which subtracts the negative tweet share from the positive tweet share. This approach proved to be a stronger predictor of vote share for both candidates. While we explored the correlation between population demographics and sentiment, we found it made more sense at the county level, though it was not a significant predictor at the state level when controlled for. Our regression analysis indicated that the prediction score was a significant predictor of vote share, and we recommend using it for future sentiment analysis.

A key limitation of Twitter data is the disproportionate use of the platform, with certain demographics and political groups being overrepresented, making it challenging to generalize the findings to the broader population. Future work could involve developing methods to adjust for these biases, such as weighting the data or incorporating external sources to create a more representative analysis.

Bibliography

Duggan, M., & Smith, A. (2022). Politics on Twitter: One-third of tweets from U.S. adults are political. Pew Research Center.

O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series." *ICWSM*, vol. 11, no. 122-129, pp. 1–2, 2010.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. International AAAI Conference on Weblogs and Social Media (ICWSM).

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives, 31*(2), 211–236. https://doi.org/10.1257/jep.31.2.211

B. Joyce and J. Deng, "Sentiment analysis of tweets for the 2016 US presidential election," *2017 IEEE MIT Undergraduate Research Technology Conference (URTC)*, Cambridge, MA, USA, 2017, pp. 1-4, doi: 10.1109/URTC.2017.8284176.

Perrin, A. (2019). Americans are changing their relationship with Facebook. Pew Research Center

Race and the American Electorate: An Examination of the Relationship between Racial Identity and Political Behavior" Michael C. Dawson, Annual Review of Political Science,1994 .

The 2020 U.S. Presidential Election: High Turnout, Partisan Polarization, and a Pandemic by Michael McDonald and Robin Best (2021)