Navigating the Currents: Demographics, Sentiment, and Media Influence in American Political Waters

Naresh Kumar Kaushal, Rutuja Abhijit Kale

March 20, 2024

Abstract

This study investigates the interplay between demographic factors and public sentiment towards key national and international events impacting U.S. politics, as well as the role of media bias in shaping public opinion. Utilizing sentiment analysis on a vast array of Reddit comments from all 50 states on pivotal issues such as the economy, abortion, healthcare etc. This research integrates demographic data from the U.S. Census Bureau—including income levels, education, and adult population sizes—to conduct a comprehensive regression analysis. Additionally, the study delves into media framing by scrutinizing language patterns within news articles from allsides.com, focusing on divisive topics like the Israel-Hamas conflict and the Ukraine-Russia war. By examining publication trends and public alignment with media perspectives through advanced visualizations, this research aims to illuminate the complex dynamics between demographic variables, significant political events, and media influence, offering profound insights into the fabric of American public discourse.

Introduction

The rise of digital platforms has transformed the landscape of public discourse, making the analysis of public sentiment and media influence more pertinent than ever. This study probes into the synergy between public sentiment on key national issues and media bias, and its subsequent effect on public perception. It examines two main areas: the sentiment reflected in state-specific Reddit discussions on major topics, and the portrayal of these topics by media outlets with different political biases. The study investigates the extent to which international events may influence media coverage, potentially shaping the narrative presented to the public. For this research, we have formulated two key research questions to guide our investigation and analysis.

Research Questions

1. What is the relationship between state level public sentiment on national issues and various demographic factors and how this might affect political landscape?

2. How does media bias manifest in the coverage of national issues and major international events, and what patterns can be identified from this coverage? and how might they influence public opinion?

The timing of this study is critical, examining data from 2016 to 2020, a period that culminated in a significant U.S. election. It aims to correlate demographic and sentiment data with electoral outcomes, providing a predictive framework for future elections. The inclusion of media bias analysis offers insights into the dynamics that may shape the 2024 elections.

Data Acquisition

Our project undertook a three-pronged approach to data collection, encompassing demography data, public opinion, and news articles to provide a comprehensive landscape for our analysis. Initially, we gathered demographic data to understand the underlying social and economic fabrics of different regions. Subsequently, we gauged public opinion through sentiment analysis of social media and forum comments, capturing the pulse of the populace on various issues. Lastly, we augmented our dataset with news articles, offering a perspective on media framing and its potential impact on public sentiment. Each dataset was meticulously sourced and processed to ensure robustness in our ensuing analysis.

The U.S. Census Bureau's website, census.gov [10], was our primary source for demography data, offering detailed insights into various demographic factors. Our focus was on key demographic indicators such as the young and adult populations, gender distribution, income levels, and educational attainment across the states. The Census Bureau provides a robust set of APIs that allow for streamlined access to their vast repositories of data. To identify the specific datasets we needed, we initially delved into the comprehensive list of available tables, accessible at Census API variables [11]. This preliminary research was crucial in pinpointing the exact tables that contained the demographic data relevant to our study.

Table 1: Census Data Variables Description

Table Name	Description
B01001_006E-B01001_011E	15 years to 29 years males
B01001_030E-B01001_035E	15 years to 29 years females
B19013_001E	Estimate median household income
B15003_022E	People with bachelor's degree
B15003_023E	People with master's degree
B15003_024E	People with other professional degree
B15003_025E	People with PhD degree
B01001_011E-B01001_019E	30 years to 65 years males
B01001_036E-B01001_043E	30 years to 65 years females
B01001_002E	Approximate male population
B01001_026E	Approximate female population

Upon identifying the necessary tables, we utilized the Census API to extract the data. The API calls were structured as follows:

```
https://api.census.gov/data/2020/acs/acs5?get=[table_name]&for=state:[state_fips_code]
```

By replacing [table_name] with the name of the required table and [state_fips_code] with the FIPS code [9] for each state, we could efficiently pull the demographic data for each state from 2015 to 2020. Then to consolidate multiple demographic data tables into a single, comprehensive dataframe we merged all demography related tables using their FIPS codes column as a common identifier.

For capturing public opinion on key national issues across different states, we turned to Reddit, a platform with dedicated subreddits for each state such as r/California, r/Oklahoma etc. Our goal was to analyze the public sentiment on topics like the economy, COVID-19, BLM, healthcare, immigration, LGBTQ+ rights, and abortion ban. Given the challenges of data acquisition from reddit through scraping, such as the high volume of comments and the complexity of filtering relevant discussions, we sought an efficient alternative.

We opted for the PullPush API [3], which provided a streamlined solution for accessing Reddit comments. This API allowed us to collect data on the specified topics from all 50 states during the period of November 08, 2016 to November 03, 2020 encompassing the Trump administration. The PullPush API furnished us with essential information, including the text of comments, the number of upvotes, the posting date, and the subreddit name. We stored the acquired data in a JSON format, subsequently converting it into a pandas DataFrame for ease of analysis. By aggregating the comments from various states, we compiled a comprehensive DataFrame that encapsulated public opinion on each topic across the nation, providing a rich dataset for our sentiment analysis.

For media analysis part we used allsides [6] to get news articles and their political inclinations. The initial objective was to collect media bias ratings for various news sources. Subsequently, the project's scope was expanded to include data on international topics, such as the Israel-Hamas and Ukraine-Russia conflicts, capturing publication dates, article titles, and media bias ratings. Additionally, the research focused on crucial U.S. national topics like the economy, healthcare, abortion, and immigration, with an emphasis on collecting not only article titles and their sources but also the full text.

In the absence of a direct API for data access on allsides.com, web scraping techniques were employed. The initial challenge involved scraping dynamically loading HTML content displaying media bias ratings,

which necessitated innovative methods to scrape the complete dataset. Although Selenium was considered, it proved to be time-consuming. This issue was resolved by leveraging the URL's page parameter, which allowed for efficient navigation through different sections of the table. The second phase of data collection presented additional complexity as it required fetching full-text content from various news sources' websites, each with its unique HTML structure. Python's 'newspaper' library [2] was employed to download and parse the content, proving to be an effective solution.

The task of fetching the full-text content of the articles proved to be time intensive and to address this, multithreading and caching techniques were employed, significantly optimizing the data retrieval process. The careful management of the number of worker threads was imperative to prevent system crashes, and data was saved in batches to conserve memory and ensure the stability of the process.

The resulting data was meticulously structured into a DataFrame, integrating columns for media bias ratings, their agree vs disagree counts, article URLs, and additional fields like article text and title. This structured approach facilitated in-depth analysis of media bias across different topics and regions, providing detailed insights into the news articles for key national topics such as abortion, healthcare, the economy, and immigration.

Preprocessing and Analysis of Data

Sentiment analysis on reddit data

Before sentiment analysis, we employed thorough data cleaning using regex to remove embedded URLs, images, unnecessary special characters and numbers, superfluous blank spaces, and hashtags from the comments. Additionally, comments that surpassed the input size limit of the pre-trained transformer model were succinctly summarized using the '**facebook/bart-large-cnn**' summarizer, ensuring that the essence and sentiment of longer comments were preserved.

To analyze the sentiment of comments from state-specific subreddits, we first tokenize the comments using **AutoTokenizer** then we utilized the Hugging Face pre-trained transformer model **ardiffnlp/twitter-roberta-base-sentiment**, optimized for social media texts. This model's contextual understanding surpasses simpler methods like TextBlob, offering a nuanced analysis of sentiment. We calculated the overall sentiment score for each state by aggregating the weighted sentiments of individual comments, with weights proportional to their upvote counts. This approach prioritizes comments with higher engagement, providing a more representative sentiment analysis.

	state	topic	comment	sentiment_score	score
0	Alabama	abortion	No one can legally force you to get an abortio...	-0.848300	6.0
1	Alabama	abortion	Abortions are not illegal in Alabama	-0.587977	2.0
2	Alabama	abortion	He cannot legally make you have an abortion	-0.736194	7.0
3	Alabama	abortion	gtCan he legally remove me without formal noti...	-0.934370	7.0
4	Alabama	abortion	Must give credit where credit is due Trump has...	-0.608584	6.0

Figure 1: Comments on abortion from r/Alabama subreddit

In our subsequent analysis, we explored the relationship between demographic factors and topic sentiments by fitting a regression line. While acknowledging that correlation does not equate to causation, this method helped illuminate potential connections between demographic characteristics and public opinion trends.

LDA and topic modelling

Similarly, after acquiring the data from allsides.com, the next phase involved cleaning the text that includes converting the text to lowercase, removing punctuation, digits, and specific commonly used words that could

skew the analysis (like 'said' and 'would'). Regular expressions were employed to eliminate unwanted characters and numerals, preparing the text for Natural Language Processing (NLP) tasks.

The cleaned text underwent tokenization, breaking it down into individual words. Each token was then subjected to part-of-speech tagging, a crucial step for the subsequent lemmatization process. The lemmatization, performed by the Natural Language Toolkit (NLTK)'s '**WordNetLemmatizer**', aimed to reduce words to their root form, considering their part-of-speech tags. This step ensured a more meaningful and uniform representation of the text by eliminating variations of the same word. Furthermore, stop words, which are common words that contribute little to the overall meaning of the text, were removed to refine the content further.

The clean text data obtained was then analyzed using Latent Dirichlet Allocation (LDA) [8], implemented through the 'gensim' library. LDA is a statistical model that discovers the underlying topics in a collection of documents, making it suitable for uncovering prevalent themes across the articles categorized by media bias (left, right, center). For the LDA analysis, a dictionary and corpus were constructed from the cleaned text, which were essential for modeling the topics. The LDA model was configured to identify a set number of topics, iterating through the documents multiple times to enhance the topic quality and coherence.

Visualizations

Our visualizations transform intricate data sets of the reddit and demography data into clear, engaging narratives. A choropleth map highlights regional sentiment trends, while demographic plots expose socioeconomic patterns. Regression analyses reveal correlations in an understandable format.

For media analysis we employed visualizations like bubble chart depicting public opinion on allsides [6] media bias rating. Stacked area charts to track changes in media coverage over time. Word clouds visualize the main themes from LDA-derived topics, providing a qualitative snapshot of media coverage's thematic focus. Collectively, these visualizations synthesize raw data into insights, aiding the exploration of media bias and public sentiment. Along with these visualizations we also created 2 Dash apps one for the state wise correlation between sentiment scores of various national topics and demographic features Dash App - 1, and second to see how media can play a role in shifting public opinion Dash App - 2 (These apps sometimes take 50 sec to load because of unpaid version of hosting site). These applications, enhanced by their creation with Dash Python libraries and their hosting on Render.com, enable deep interaction with the data, promoting an immersive and dynamic data experience.

Results and Discussion

Demography vs public opinion on various national topics

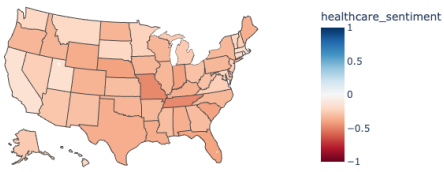

Figure 2: State wise healthcare sentiment

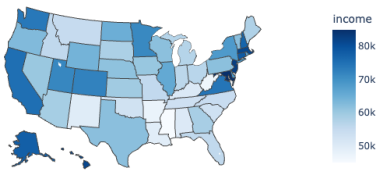

Figure 3: Median Household income across different states

The visualizations for healthcare sentiments and median household income suggest that while there's a positive correlation Figure 4 between income and healthcare sentiment across the US, no state shows a positive sentiment Figure 2, reflecting a nationwide concern over healthcare policies during Trump's tenure.

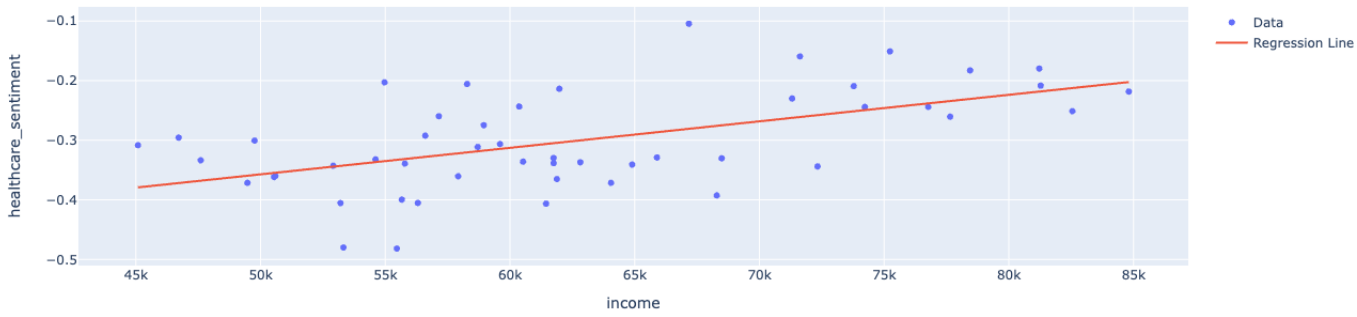

Figure 4: Correlation between state wise healthcare sentiment and median household income

This could be attributed to the debates surrounding the Affordable Care Act and its potential repeal, which was a significant aspect of Trump's policy agenda. High-income states tend to have less negative sentiments, which may indicate that although more affluent populations are somewhat more satisfied with healthcare policies, there's an overarching national discontent that transcends income levels. This discontent could have influenced voters in swing states, where healthcare was likely a critical issue, contributing to the volatile political landscape of the period.

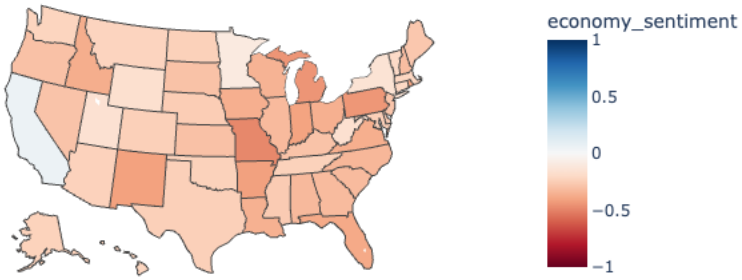

Figure 5: State wise public sentiment on economy

The economy sentiment map Figure 5 during Trump's tenure indicates a predominantly neutral to slightly negative perception of the economy across most states, with no states exhibiting strong positivity. This sentiment likely reflects the economic challenges of the period, such as trade tensions and the pandemic's impact, which could have influenced voter behavior. The lack of deep negativity suggests that while there were concerns, there may have been a degree of confidence in the economic policies or resilience, underscoring the complexity of economic sentiment as a factor in the political landscape.

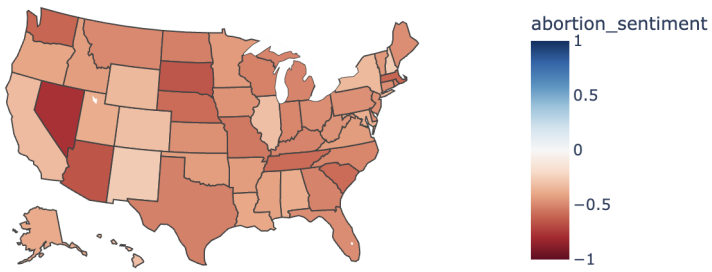

Figure 6: State wise abortion sentiment

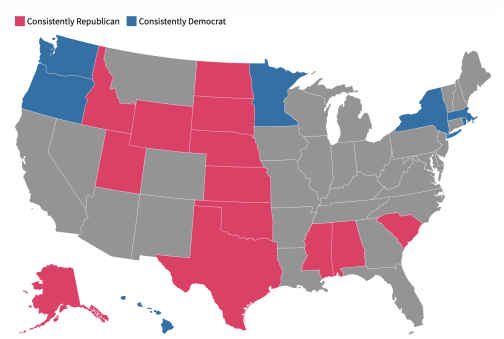

Figure 7: Blue and Red Divide: Mapping Consistent Party Affiliation by State
Source: Federal Election Commission

The abortion sentiment map Figure 6 reveals a predominantly negative sentiment across the majority of states, with some showing particularly strong negative sentiments. Comparing this map to the one indicating political affiliations Figure 7 suggests that states with consistently Republican voting patterns may correlate with stronger negative sentiments toward abortion. This alignment may reflect the influence of political ideologies on public opinion regarding abortion, a highly polarized issue that remains at the forefront of American politics, especially during Trump's tenure when it was a significant point of contention.

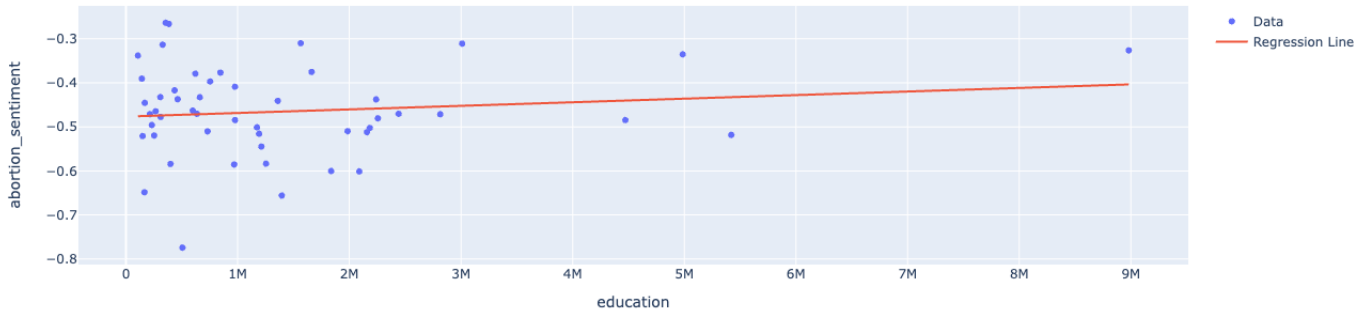

Figure 8: Correlation between state wise abortion sentiment and educated population

The regression plot of educated population vs abortion sentiment Figure 8 shows a discernible positive slope. This could imply that higher education levels in a state are associated with slightly less negative sentiments on abortion, possibly due to more exposure to diverse viewpoints or educational materials covering reproductive rights.

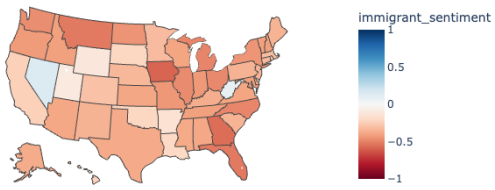

Figure 9: State wise sentiment on immigration policies

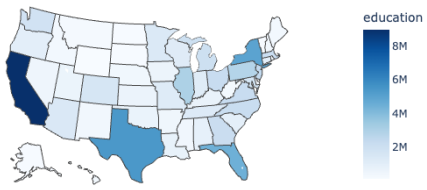

Figure 10: State wise education level

The regression plot Figure 11 between immigrant sentiment Figure 9 and education level Figure 10 indicates a slight negative slope, suggesting that as education levels increase, the sentiment towards immigrants becomes

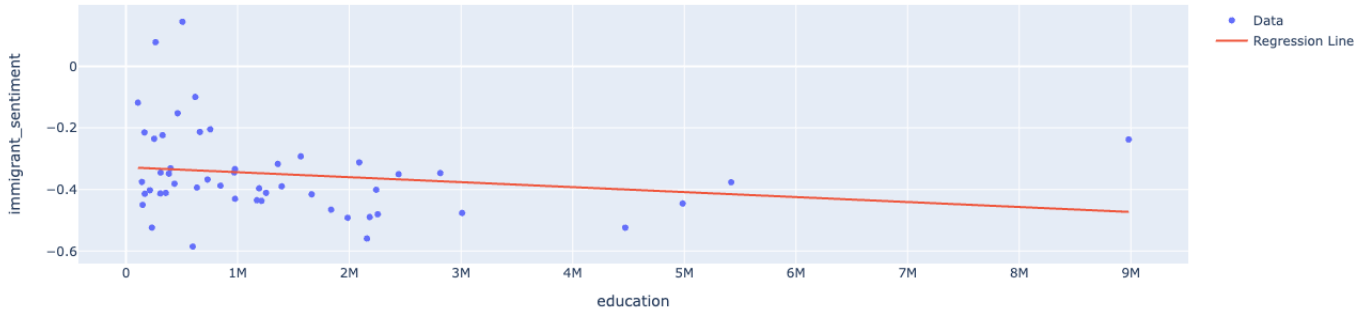

Figure 11: Correlation between state wise immigration sentiment and level of education

slightly more negative. This might seem counterintuitive if we expect higher education levels to correlate with more progressive views. However, this could reflect complex social dynamics where higher education populations have nuanced views on immigration, possibly due to economic competition in skilled job markets or differing perceptions on cultural integration. The insights from these visuals, taken in the context of Trump's tenure, could suggest that while traditionally higher education is associated with more liberal views, when it comes to immigration, there might be a more varied or complex relationship. This complexity could be rooted in the specific immigration policies during this period, such as debates on DACA, border security, and the publicized political rhetoric surrounding immigration.

Media analysis

The bubble chart Figure 12 visualization from allsides.com [6] data reveals public agreement or contention with media bias ratings shared by allsides. Bubble size, indicative of total responses, highlights the news source's potential influence on public perception. Larger bubbles signal significant interaction, suggesting these sources may more strongly shape public sentiment. The chart directly compares the number of agreements and disagreements on the bias ratings, with media outlets positioned on the diagonal line indicating strong conflicting views among people. Outlets located to either side of the diagonal are regarded as having a more pronounced bias.

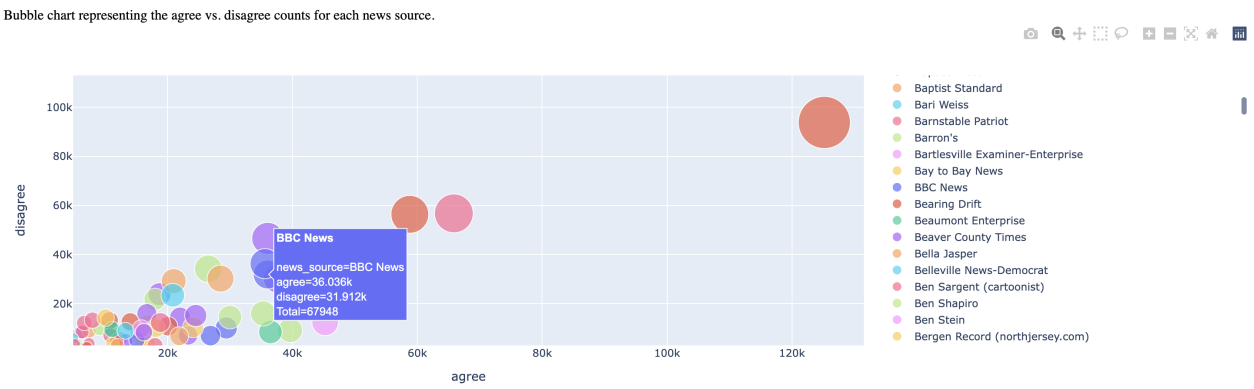

Figure 12: Bubble chart showing public opinion to bias ratings of various media houses.

BBC news media has nearly equal number of agree and disagree counts in its bias rating indicating that public sentiment about this bias rating is divided. This proximity of agree and disagree counts suggests that the audience may have a nuanced view of BBC News' bias or that its audience is split on their perception of the news source's political leaning. Popular media outlets such as The New York Times, recognized as left-leaning by sources like allsides, receive similar acknowledgment from the public. Similarly, Fox News, known for its right-leaning stance, is also affirmed by audience perceptions. These news organizations are

widely consumed and prevalent, and their political inclinations grant them the power to sway public opinion subtly, thereby playing a significant role in shaping societal views.

The stacked area charts provide a temporal view of media coverage across biases during critical international events, like the Israel-Hamas conflict and the Ukraine-Russia war. Notable spikes in article counts during these events suggest a surge in media activity, possibly in response to increased public interest or as an attempt to influence public opinion.

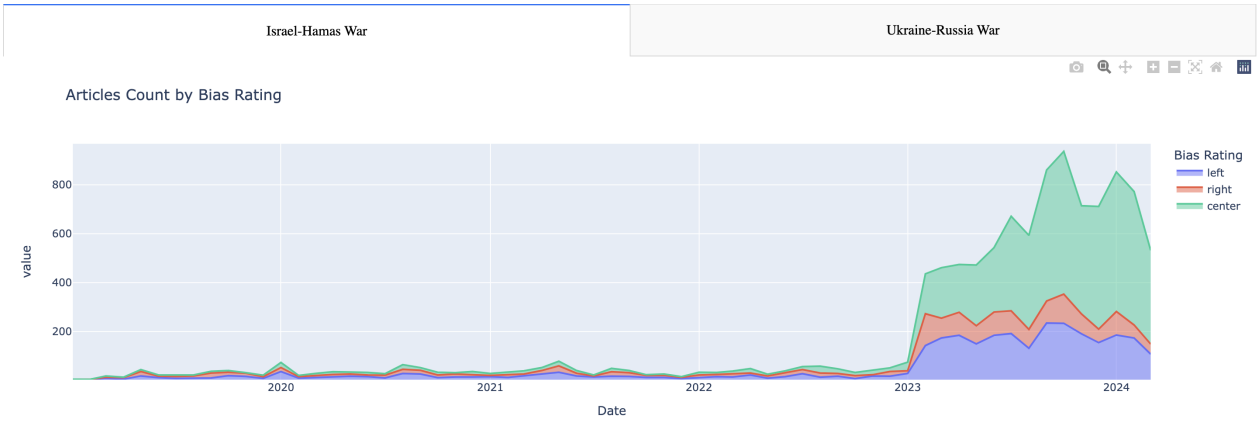

Figure 13: Stacked area chart showing spike in article count for the Israel-Hamas conflict around Oct 7, 2023

For example the stacked area charts for the Ukraine-Russia conflict Figure 14, illustrate a discernible spike in article production from right-wing media outlets, surpassing that of left-leaning sources. This could reflect a variety of factors, including the right's focus on geopolitical strategy or national security concerns that are often intensified during wartime. It may also indicate a response to their audience's interests, which could skew towards international policy and defense topics during such events. Conversely, left-wing media might concentrate on the human aspect of war or diplomatic efforts, leading to a less pronounced spike in article counts. This tendency mirrors the ideological perspectives that shape each side's narrative and editorial choices, influencing how the war is reported and potentially affecting public opinion and sentiment. The variation in article volume among different biases illustrates the diverse narratives presented by the media. This highlights the need for critical evaluation of media coverage, especially during significant global incidents.

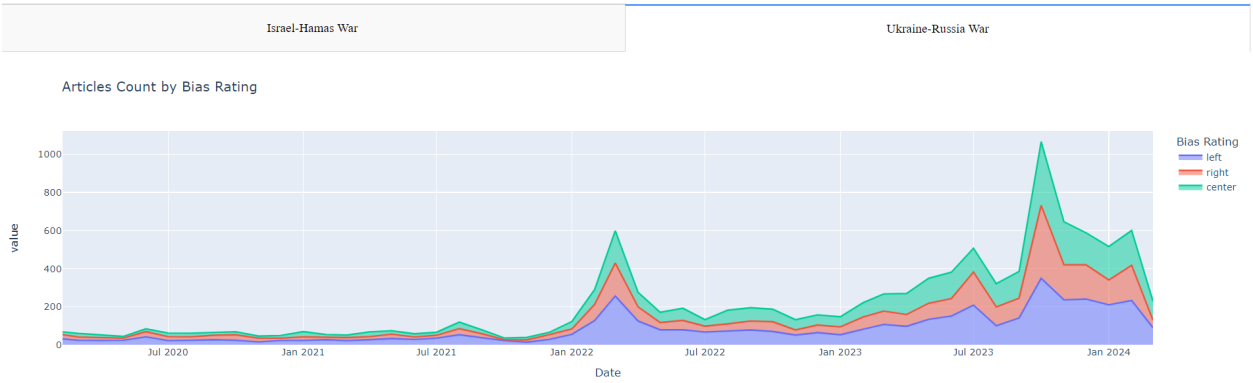

Figure 14: Stacked area chart showing the article count by bias rating during the Ukraine-Russia war.

Word clouds Figure 15 derived from topic modeling provide insight into the thematic emphasis of media coverage. For example, for national topics like healthcare, the left-leaning word cloud emphasizes policy-related terms like "insurance" and "plan," suggesting a focus on reform and accessibility. Moreover, the presence of 'Gaza' and 'Israel' in the left-leaning word cloud may also signify the media's engagement in discourse on how conflicts, like those in Gaza and Israel, impact healthcare services, precipitate humanitarian crises, and shape international healthcare policy and aid. In contrast, the right-leaning cloud features terms

like "Obamacare" and "tax," hinting at fiscal concerns and policy critique. The centrist word cloud, with terms like "hospital" and "patient," leans towards practical aspects of healthcare.

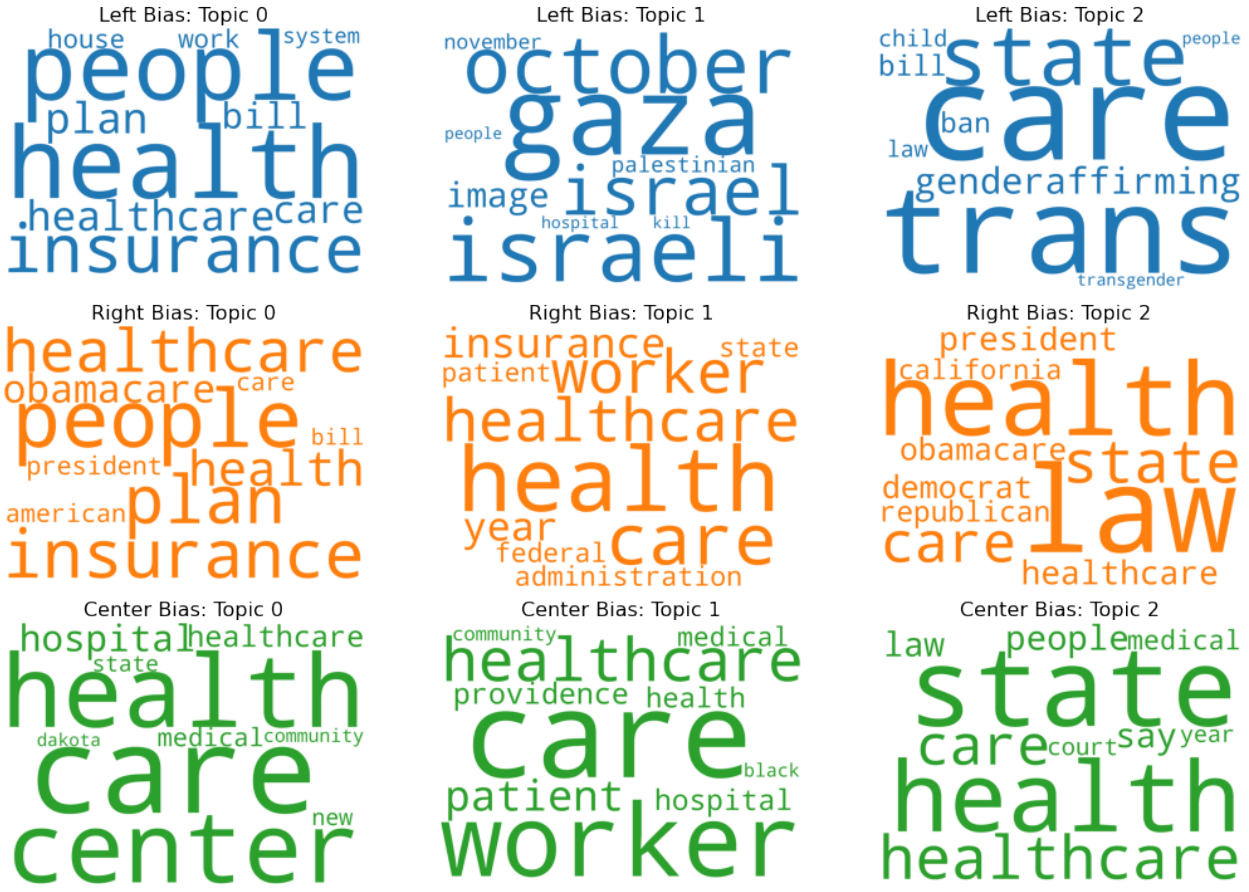

Figure 15: Word clouds for healthcare

These visualizations elucidate the subtle ways media bias permeates public discourse, with healthcare word clouds demonstrating how language used by different biases might inform the healthcare debate and public opinion.

Conclusion

This case study has provided a multi-dimensional view of the American political landscape, spanning the period of Trump's presidency and the transition into the Biden administration. The interplay between demographic factors and public sentiment, addressing our first research question, has revealed a consistently neutral to negative perspective on pivotal issues such as the economy, healthcare, and immigration, with these sentiments being influenced by a variety of socioeconomic indicators including education and median income. Notably, the heightened negative sentiment towards certain policies may have contributed to the political shift observed in the 2020 election, where Trump failed to secure a second term.

The study's integration of media analysis to address the second research question, capturing current events, offers an invaluable lens into how public opinion may be swaying in the early tenure of President Biden. Media's portrayal of ongoing international conflicts and domestic policies, particularly during a time of global instability, provides insight into potential challenges that the current administration will face in seeking public approval and maintaining political footing.

In a broader context, the findings highlight the significant influence of media narratives in either bolstering or eroding political figures, as well as their policies. While our analysis covers a crucial historical epoch marked by Trump's distinctive political style, it also serves as a harbinger for Biden's presidency. Understanding the complex dynamics at play can help anticipate the political headwinds and tailwinds that may impact

future electoral strategies and policy-making decisions. As we continue to navigate through the digital age's vast information landscape, the necessity for critical analysis of both media portrayal and public sentiment becomes increasingly paramount in deciphering the ever-evolving narrative of American politics.

Limitations and Future Work

The scope and depth of this study, while extensive, encounter several limitations that open pathways for future inquiry. One of the primary constraints is the reliance on sentiment analysis which, despite advancements in NLP, may not capture the full spectrum of nuanced human emotions and the complex contexts within which public opinion is formed. Additionally, the demographic data, while robust, does not account for all factors that may influence sentiment, such as cultural nuances, local policies, and non-economic aspects of well-being.

Another limitation is the inherent bias in Allsides and potential bias in the user base of platforms like Reddit. This bias could skew the perceived sentiment and may not be entirely representative of the broader population. Moreover, the dynamic nature of the media landscape and its real-time evolution present challenges in capturing the full extent of media influence on public sentiment.

Future work could expand on the data sources by incorporating a wider array of social media platforms and demographic indicators, as well as by employing longitudinal studies to track changes over time. An interdisciplinary approach combining data science with social science theories could yield richer insights into the drivers of public opinion. There is also a significant opportunity to explore the impact of misinformation and its spread through social media on public sentiment and political outcomes.

The role of international events in shaping domestic politics offers fertile ground for research, particularly in relation to the Biden administration's policy responses and their media framing. In-depth case studies on specific policies or events during the Trump and Biden administrations could provide granular insights into the interplay between policy, public sentiment, and media narratives.

Source code

GitHub repository Link

References

[1] FiveThirtyEight Polls. https://projects.fivethirtyeight.com/polls/.

[2] Newspaper3k: Article scraping & curation. https://newspaper.readthedocs.io/en/latest/.

[3] PullPush: Transforming How We Interact with Data. https://pullpush.io/.

[4] USAFacts: Your Nonpartisan Source for Data. https://usafacts.org/.

[5] YouGov Today. https://today.yougov.com/.

[6] AllSides. Unbiased & balanced news. https://www.allsides.com/unbiased-balanced-news.

[7] Pew Research Center. Pew Research Center. https://www.pewresearch.org/.

[8] Radim Rehurek. Latent dirichlet allocation (lda) in gensim. https://radimrehurek.com/gensim/models/ldamodel.html.

[9] United States Census Bureau. Ansi codes for states. https://www.census.gov/library/reference/code-lists/ansi/ansi-codes-for-states.html, 2024.

[10] U.S. Census Bureau. U.S. Census Bureau Home Page. https://www.census.gov/.

[11] U.S. Census Bureau. American Community Survey 5-Year Data Variables 2020. https://api.census.gov/data/2020/acs/acs5/variables.json, 2020. Accessed: [Insert Date Here].

[4] [7] [1] [5]