How Much Does Bias in the Media Affect Election Outcomes

Team 5 Final Report

Lucy Lu, 28273098, junyil16@uci.edu
Tram La, 14166227, latb@uci.edu
Vaishnavi Ravinutala, 15538807, vravinut@uci.edu
Nancy Carlson, 15744122, nacarlso@uci.edu
Deborah Rosa Franza, 72533099, franzad@uci.edu
GitHub: https://github.com/Political-Bias-Analysis

1. Introduction and Problem Statement

In America we often turn to major media outlets for politics and election news. Therefore, we expect these new sources to deliver accurate and unbiased information, however, we all know that this is not the case. Hence, in this research, we aim to investigate the impact of bias in media outlets on election outcomes. We focus on four bias topics: abortion, race, immigration, and socioeconomics, and examine how sentiments expressed by news outlets (CBS, CNN, FOX, and NPR) compare to public sentiment on Twitter. Our findings reveal that sentiments expressed by news outlets and Twitter users differ significantly. Media sentiments tend to be extreme, either very positive or very negative, while Twitter sentiments are more evenly distributed. However, due to the limited three-year election data available, we were unable to find a significant relationship between media article sentiments and election results. On the other hand, our analysis using the Twitter model showed that a one-unit increase in the collective sentiment score across all topics resulted in a tenth of a unit increase in the normalized vote count for Republican candidates in House elections. Lastly, no increase was observed for Democratic candidates.

These findings shed light on the complexity of media bias and its potential influence on election outcomes. Despite the limitations and the evolving landscape of media platforms, our research contributes to understanding the role of bias topic sentiments in shaping public opinion and political dynamics.

2. Related Work:

For the statistical analysis portion of this project, we used both Linear Regression and Statistical Tests to gather evidence for our various hypotheses.

2.2.1 Related Work on Linear Regression

As a reference for our two Linear Regression models between media sentiment scores and Twitter sentiment scores on normalized election vote counts, we discovered a Polishh paper by Krochmal that used a number of OLS regression models to determine if sentiments in polish tweets were a significant predictor in vote share in Powiat (which is known as a regional government in Poland) while taking into account other necessary socioeconomic covariates for each specific Powiat (region). They discovered that sentiment scores in tweets were indeed a statistically significant predictor, which mirrors our regression results in which we found that Twitter sentiments for each bias were a significant predictor.

2.2.2 Related Work on Statistical Tests

In a paper by Park from the University of Central Missouri, an ANOVA test was utilized to examine potential differences in sentiment scores among Twitter posts pertaining to various types of Asian restaurants. The study revealed statistically significant variations in Twitter sentiments across Japanese, Korean, Thai, and Chinese restaurants. Building upon this foundation, for our report the ANOVA by ranks also known as the Kruskal-Wallis Rank Sum Test aims to investigate if there are significant differences in scores based on the source (News or Twitter) and the year from which the scores were obtained.

3. Data Sets

Four databases were created for this research project: “electiondb”, “sentimentdb”, “articlesdb”, and “Twitterdb”. For articlesdb, there are three main schemas (articles, biases, and media_source) to store necessary articles data. As for Twitterdb, there is one general schema that stores clean Twitter data. The electiondb is 11 MB in size and consists of two schemas named “results” and “voters”. The sentimentsdb is 86 MB in size and consists of two schemas, “articles” and “twitter” .

The election data was received as an Excel workbook file. These files were fairly standardized, but still had slight differences that needed to be cleaned separately. Data on Presidential, Senate, and House election results, by state, and party were obtained from the “Federal Election Commission United States of America.”

Year	Office	State	Last Name	Party	Vote %	Vote Count
2020	Pres	AL	Trump	REP	0.620316	1441170
2020	Pres	AL	Biden	DEM	0.365700	849624

Table 1: Election ‘results’ table 2020 to 2010. (5237 rows by 7 columns)
https://www.fec.gov/introduction-campaign-finance/election-and-voting-information/

The response variable which addresses our research question is the normalized mean vote count of registered voters. To normalize the vote count of each state, data on voter registration counts were obtained from the “United States Census Bureau”.

Year	State	Population	Registered	Voted	Normalized Registered Vote	Normalized Population Vote
2020	AL	3769	2527	2247	0.889197	0.596179
2020	AK	528	383	330	0.861619	0.625000

Table 2: Census ‘voters’ table 2020 to 2010 in thousands (260 rows by 7 columns)
https://www.census.gov/topics/public-sector/voting/data/tables.2010.List_1863097513.html#list-tab-List_1863097513.

Data is grouped by region to determine party affiliations, thus illustrating how a voter's geolocation influences their party affiliation. For the article data, we collected approximately 15,914 articles from 2016 to 2020, from CBS (https://www.cbsnews.com/), CNN (https://www.cnn.com/), FOX (https://www.foxnews.com/), and NPR (https://www.npr.org/). News media article data were collected by scraping the target word elections with keywords that related to one of the four topics of interest: abortion, race, immigration, and socioeconomics. For each article, we collected six particular attributes, which are *headline*, *author*, *source*, *published date*, *url*, and *article content*. All six attributes are TEXT data types except for *date* which is the DATE type. Three primary keys of our table are headline, author, and published date to avoid duplications

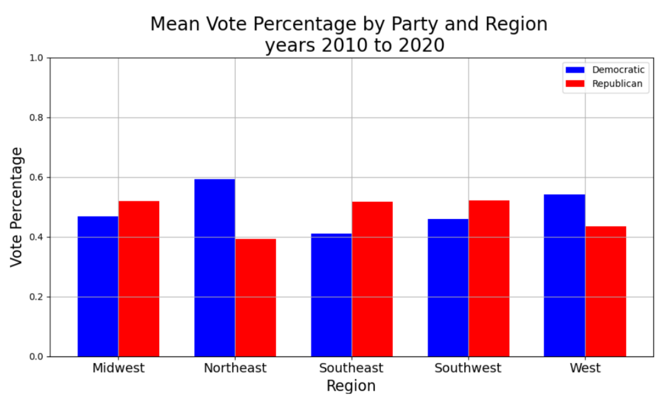

Fig 1: Mean Vote % by Region and Party: 2010 - 2020

Headline	Author	Source	Bias	Publish Date	URL	Article Content
Trump focuses on immigration issues. . .	Grace Segers	CBS	immigration	2018-10-30	https://www .cbs…	Earlier this month, President Trump offered a. . .
Clinton: Trump campaign built on ‘prejudice..’	Dan Merica	CNN	racial	2016-08-25	https://www .cnn…	Clinton is painting Trump as outside the norm of American..

Table 3: Article table 2016 to 2020. (15914 rows by 9 columns)

Figure 2 shows the number of media articles collected on each topic of interest. The topic of immigration stands out as a primary topic covered by news organizations.

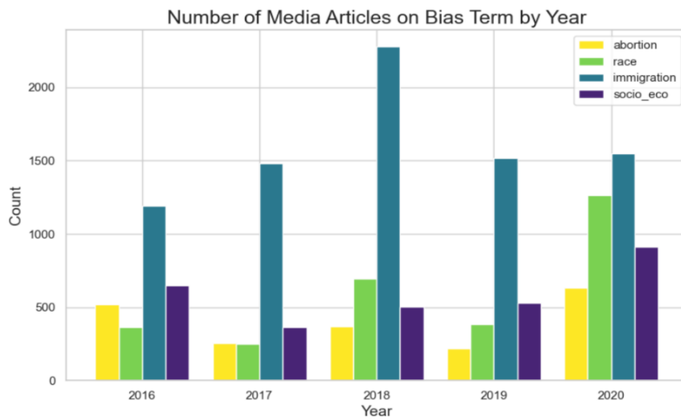

Fig 2: Annual Total Number of Media Articles

The Twitter data is provided by Professor Chan Li under the ICS department. This data was received as unlabeled CSV files with 33 columns and approximately 59,861 tweets from 2016 to 2022. Extracting only the necessary variables reduced the table to six columns. Those columns are *user_id, date, tweet, city, state, and user_name*, all attributes are TEXT data types except for *date* which is the DATE type.

User id	Date	Tweet	City	State	User name
245053853	2016-01-01 02:56:36	Richard Shelby holds real FTD meetings with his money men every election cycle. #AnybodyButShelby	Gulf Shores	Alabama	Wally Jimenez

Table 4: Twitter table from 2016 to 2020. (21871 rows by 10 columns)

Figure 3 below shows the total tweets collected on each topic. The tweet on the topic of immigration outnumbers the other subjects until the year 2020 where it is overtaken by the subject of race.

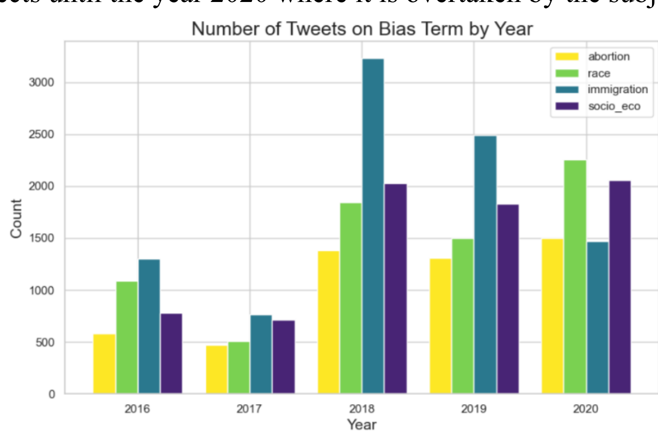

Fig 3: Annual Total Number of tweets

4. Overall Technical Approach

4.1 Scraping Websites and Data Collection

To represent a variety of viewpoints across the United States news articles are scrapped from four main news sources: CBS, CNN, FOX, and NPR. BeautifulSoup and Selenium Python libraries were used to facilitate the website scraping process. Articles were collected that mentioned elections and included wording of the four research topics: abortion, race, immigration, and socioeconomics.

4.2 Preprocessing

The scrapped media article text was identified by its bias group during the scraping process. However, tweets from Twitter came to us unlabeled in separate files by year. Text from both sources went through several Natural Language Processing (NLP) cleaning steps: 1. Removed repeating and meta characters from text, 2. Identify "parts of speech" such as nouns and pronouns, etc. 3. "Lemmatization" breaks words down from plural forms to singular forms. Tweet text is then run through a separate function to classify into one of the four bias topics. The election data came to us in yearly Excel workbooks. The data is organized in a slightly different manner each year. The party labels are standardized as three-letter labels before they are uploaded to the database. In addition, all records were checked for NaN values and removed if found. Lastly, the "state" column in the census data was standardized into two-letter abbreviations to correspond to the election data.

4.3 VADER Sentiment Scoring

Once the text from Twitter and news articles finishes the NLP steps, it is ready to be fed into the VADER machine learning model for sentiment analysis. To do that, we aggregate all the articles for a specific year, source, and bias topic into one data frame. The Twitter data on the other hand, is grouped by year, state, and bias topic. And then, we feed the input into the model to get a compound sentiment score. The VADER (Valence Aware Dictionary and sEntiment Reasoner) compound score is a metric used to represent the overall sentiment polarity of a given text. It uses a pre-trained model that has been trained on a vast amount of social media text to determine the sentiment intensity of words. The Sentiment Intensity Analyzer works by assigning sentiment scores to individual words based on their presence in a lexicon, which is a collection of words and their associated sentiment scores. The compound score is calculated as a normalized, weighted sum of the individual word sentiment scores. This single value combines the individual sentiment scores of each word in the text to provide an overall sentiment intensity ranging from -1 to +1. A compound score of greater than or equal to 0.05 represents positive sentiment, less than or equal to -0.05 represents negative sentiment. Anything in between means neutral.

4.5 Kruskal-Wallis Rank Sum Test

To analyze the assumptions underlying the sentiment scores from Twitter and news media outlets, we employed a combination of visualizations (see Fig 4 and 5) and statistical tests. These assessments aimed to evaluate the normality and homogeneity of variance among scores from different sources. Upon identifying evidence of non-normality and unequal variances among score source groups, we selected an appropriate test given these assumptions to examine the relationship between sentiment scores from news media and Twitter. We opted for the Kruskal-Wallis Rank Sum test (also known as a one-way ANOVA on ranks). To facilitate the analysis, we consolidated sentiment tables based on bias topics, encompassing sentiments from both news articles and tweets. Subsequently, we randomly sampled without replacement 500 observations of sentiment scores from each source for each bias table (evident in Tables 5 and 6).

year	bias source	compound
2020	Abortion Tweet	-0.4043
…	…	…
2016	Abortion Article	0.9998

Table 5: Abortion Sentiments (1000 rows)

year	bias source	compound
2020	Immigration Tweet	0.2732
…	…	…
2016	Immigration Article	0.9406

Table 6: Immigration Sentiments (1000 rows)

4.7 Linear Regression

To measure how the sentiment in news media and tweets affects election outcomes, two separate Linear Regression (LR) models will be run. The model's response variable, "vote count", represents the votes given to candidates running for seats in the Federal House of Representatives (House) from both the Republican (REP) and Democratic (DEM) parties.

The model for media articles' sentiment scores with the election vote count on house elections for Republicans was built by an aggregated table. The table was created by matching the media article table and the voting results table by the year column. However, the matching limited the model to having only three data points. As a result, we concluded the articles table cannot be used to build a valid LR model.

The model for the Twitter sentiment scores with the same election vote count was built using the following aggregated table. This table matches the year and state columns from the Twitter sentiment table and the election results table. As a result, this model ended up with 140 data points, which gave more strength to this model.

year	state	bias	compound	norm_vote_count
2016	AK	Abortion	0.180325	0.433207
2016	AK	Race	0.464467	0.433207
2016	AK	Immigration	-0.155167	0.433207
2016	AK	Socioeconomic	0.433350	0.433207

Table 7: Twitter Data on years 2016, 2018, and 2020. (560 rows by 5 columns)

5. Software

Software(s)	Descriptions
PostgreSQL	**Usage**: Store collected data such as articles, election results, and Twitter data **Issues/Frustrations**: Since everyone in our group needs to have the same data in their local machine, it is difficult for us to keep the data up to date. **Solutions**: We have separate SQL files that are connected to Github. Once a file is changed, our group will be notified of the changes and act accordingly.
psycopg2	**Usage**: Implement scripts to create Postgresql tables and insert data from json files
Sqlalchemy	**Usage**: Implement scripts to create Postgresql tables and insert data from json files
Pandas	**Usage**: Read data from csv and json files, store necessary data into organized tables that can be used later on for data visualization and analysis
Selenium	**Usage**: Create a program that can interact with dynamic websites (to scrape article links) **Issues/Frustrations**: Since each website is built differently, we need to write a script for each individual news site. Additionally, this process takes a while to run because we need to put a sleep timer between each action (to avoid getting blocked by the sites). **Solutions**: The only solution that we were able to come up with was to write different scrape files for each individual site. While scraping, we focused on completing other tasks and left the program running in the background.
BeautifulSoup4	**Usage**: Receives links and scrape article headlines and contents.
Matplotlib/Seaborn	**Usage**: Python libraries that we use to create visualization for data explorations and analysis.
Ggridges, patchwork, ggplot2, tidyverse, car, irr	**Usage**: Create data visualization, clean data, and run statistical tests in R (Intrclass Correlation, Levene's Test, Kruskal-Wallis Test)

ReactJs	**Usage**: We are planning to use ReactJs as our front-end framework. Currently, we have the skeleton setup. We would like to display our graphs and results in a website format.
Node Js/Express	**Usage**: The backend framework that acts as a middleman between our database and the front-end.
Recharts	**Usage**: A React Js library that allows us to create interactive graphs **Issues/Frustrations**: Limited graphs provided.
Material UI	**Usage**: A React Js library that allows us to apply pre-built components that enhances the UI aspect of our website **Issues/Frustrations**: Since Material UI provided pre-built components, it would take a lot of effort to re-style them. **Solutions**: Since our dashboard is not the main priority, we focused on functionality instead of styling.

6. Experiments and Evaluation

6.1 Kruskal-Wallis Rank Sum Assumptions

To perform the Kruskal-Wallis Rank Sum Test, several assumptions need to be considered. First, it is assumed that the observations in the dataset are independent of each other. In our analysis, this assumption somewhat holds for media scores as we collected articles from different sources that are unlikely to influence one another. For Twitter scores, independence is assumed since we did not include reply posts in our dataset. Second, the assumption of data not necessarily being normal and variances not necessarily being homogenous across different groups is made. Figures 4 and 5 provide strong evidence for the variation in sentiment scores based on the data source (Twitter or media). The patterns observed in the immigration ridgeline plots can be generalized to other biases as well, as the distributions exhibit similar trends across different biases and sources. Therefore, the assumption of non-normality and non-homogeneity of variances is reasonable for our analysis.

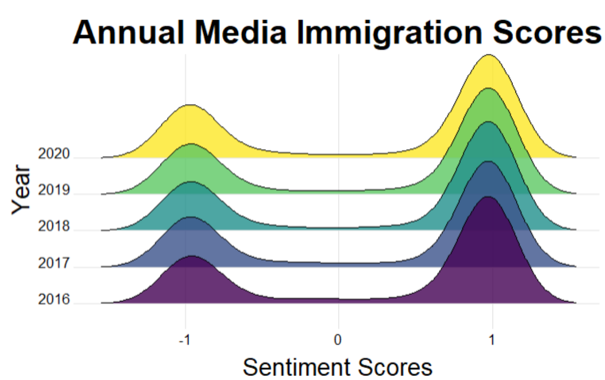

Fig 4: Media Sentiments for Immigration

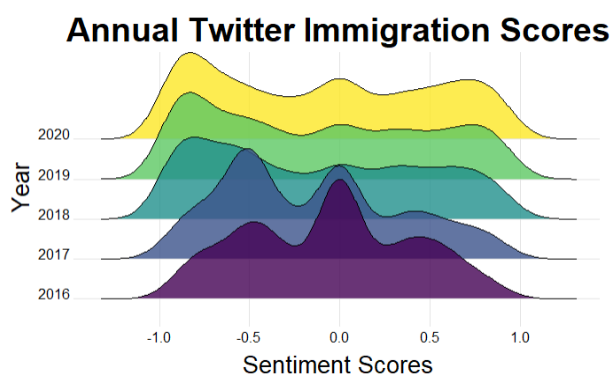

Fig 5: Twitter Sentiments for Immigration

Lastly, to satisfy the assumption of randomly sampled observations, we randomly sampled 500 compound sentiment scores from Twitter posts and 500 from media articles for each bias table. This resulted in a total of 1000 rows of sentiment scores for each bias (Tables 5 & 6), which were used for the Kruskal-Wallis test. By considering these assumptions and ensuring that they are met, we can confidently apply the Kruskal-Wallis Rank Sum Test to assess the differences in sentiment scores between Twitter and media sources for each bias category.

6.1.1 Kruskal-Wallis Rank Sum Test

The Kruskal-Wallis test was used to compare the mean ranks of sentiment scores between Twitter and news media, indicating whether the two sources have the same medians and distributions across different bias categories of abortion, race, immigration, and socioeconomics. Table 8 provides a summary of the four tests conducted. All tests resulted in statistically significant p-values, leading to the rejection of the null hypothesis and supporting the alternative hypothesis. The null hypothesis stated that "Sentiment scores from Twitter and news media have the same distribution and medians," while the alternative hypothesis suggested that "The sentiment scores from Twitter and news media have different medians and distributions." Furthermore, the large chi-squared values obtained from the tests indicate substantial differences between Twitter and media sentiment scores.

Kruskal-Wallis Test	H Test Value (Chi-Squared)	df	p-value
Socioeconomic	358.27	1	2.2e-16
Immigration	138.52	1	2.2e-16
Abortion	178.79	1	2.2e-16
Race	89.681	1	2.2e-16

Table 8: Kruskal-Wallis Rank Sum Test Summary

6.2 Linear Regression Model Assumptions

Most of the Residual plot data points are clustered around the red 0 residual line between 0.0 and 0.2 on the fitted values, suggesting the model is performing well in that range, however when the data points start to scatter away from the residual line in a cloud from 0.2 to 0.6 on the fitted values line, it suggests that the model's predictions may have larger errors or discrepancies compared to the observed values within that range (Fig. 6). The best fit line in the Regression plot indicates there is a positive relationship between the two variables but with limited strength (Fig. 7). Variability above a certain level suggests other factors may have influence on vote counts. The residuals' multivariate normality graph showed a deviation from the normal distribution, potentially due to inaccurate sentiment scores or volatility in positive scores. The correlation analysis found no strong correlation or multicollinearity between compound score and year variables, indicating independent information. The model exhibited slight underestimation and heteroscedasticity, suggesting variable prediction non-constant variance.

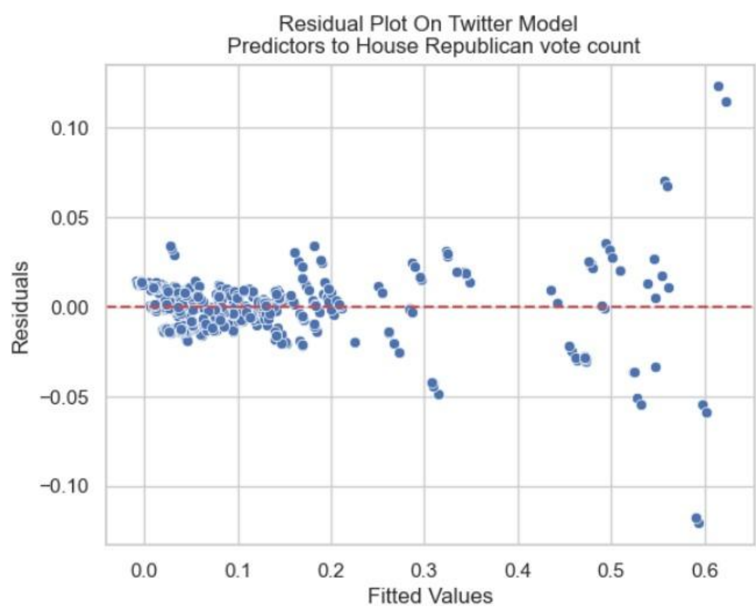

Fig 6: LR Model's Residual Plot

6.2.1 Linear Regression Model

Several LR models were attempted to examine the relationship between sentiment scores on bias topics and the response variable normalized vote count. Because the articles data table only had three data points, we did not pursue a LR model on this data. However, since there are 140 Twitter sentiment data points, we are able to fit this data to a LR model to measure the effect on election outcome. This model gave both R-squared and Adj. R-squared of 0.983, which explains 98.3% of the variation in the response variable (norm_vote_count). The combined group compound score coefficient is 0.0108 and gave a statistically significant p-value, ($p < 0.005$). (Table 9) This suggests that a one-unit increase in the compound sentiment variable is associated with a 0.0108 increase in the expected value of the transformed response variable, when holding other predictors constant. This change, although very small, being less than one unit of change in Twitter sentiment, signals that the collective Twitter sentiments on all groups has an effect on Republican voting results on House elections. Additionally, due to an insignificant p-value on the combined group compound coefficient in the Democratic LR model no increase in vote count was found. (Table 10)

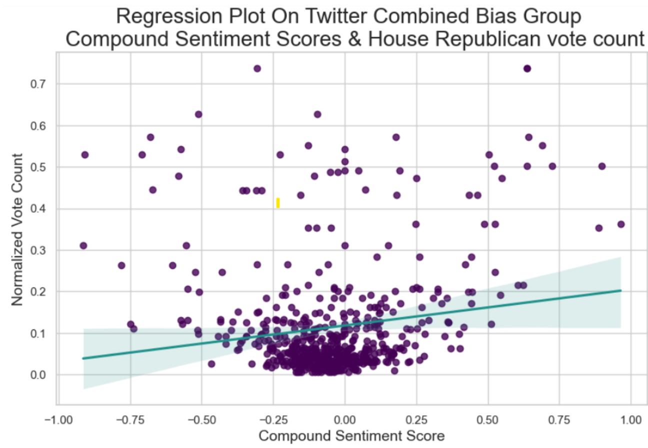

Fig 7: Normalized Votes & Sentiment Scores (REP) Plot

LR-Model	Value	P-value
R-squared	0.982983	NaN
Intercept	0.456195	9.497031e-273
Compound	0.010776	4.627617e-03

Table 9: LR Model (REP) Summary

LR-Model	Value	P-value
R-squared	0.985	NaN
Intercept	0.3102	0.000
Compound	0.0009	0.771

Table 10: LR Model (DEM) Summary

7. Notebook Description

The notebook contains four different repositories. The Scrape repository contains three subfolders. They are scrape, SQL, and election_result_clean. The scrape folder contains the code we use to scrape and clean the article data from the four news outlets and put them into JSON files. The SQL folder includes SQL code we used to create the database and tables. The election_result_clean contains code we use to wrangle the election data. The Data repository contains all the cleaned data/tables we use for modeling, such as article data, Twitter data, article sentiment table, and Twitter sentiment table. The Machine-Learning-Data-Visualization repository contains the code we use to generate sentiment scores using the VADER model. In addition, it also has the code for fitting the Linear Regression models. Moreover, it has the code we use to create visualizations for the presentation and the report. Lastly, the Client repository has the code for creating a dashboard in ReactJs.

8. Member Participation

Tram wrote the script and led the team in scraping URLs for article content. Deborah explored the compound sentiment scores data and utilized a series of hypothesis tests to help the team determine which assumptions were met. Lucy performed rigorous data cleaning of article data and later fed that data into the VADER model to calculate sentiment scores. Nancy created the election database, preprocessed media articles and Twitter data by the bias groups, and developed linear regression models. Vaishnavi assisted in scraping article data and determining which model to use to quantify sentiments. Lastly, Tram developed the dashboard website using ReactJs as the front-end framework.

Name	Contributions
Tram La	- Scraped data from CNN, NPR, FOX, CBS (article links and article contents) - Use OOP to organize article data, prepare data to be inserted into PostgreSQL - Communicated with Professor Chen Li and Yunyan Ding for Twitter data - Designed and programmed project’s website - Created interactive visualizations for website’s dashboard and report - Setup and Manage Github project organization/repositories - Provided inputs on modeling and analysis

Nancy Carlson	- Located the election data on election results, and the census data on population counts and voter registration counts. - Organized, cleaned, and extracted the relative features from the csv documents. - Created the elections database. - Created graphs from aggregated tables to show election data behavior. - Created the sentiment database from scrapped Media Articles and Twitter json files. Data was cleaned and evaluated for sentiments on our bias groups using the VADER algorithm. - Created graphs from aggregated data that show weighted averages sentiment scores for all bias groups on Media Articles and Twitter data. - Created graphs on Media Articles and Twitter counts. - Conducted extensive exploratory data analysis on Twitter, Media Articles, and Election data (ref. EDA_models.ipynb) - Created and tested several Linear Regression models from aggregated tables on Media Articles and Twitter data.
Deborah Rosa Franza	- Assisted in scrapping Fox news articles - Tokenized and lemmatized articles from all media articles to create data visualizations of most common words - Assisted in visualizing media and Twitter data with election data to assess regression fitting viability - Created/validated ANOVA models and accompanying interpretations using Boxplots, Ridgeline graphs - Assisted in hypothesis formulations - Created various data visualizations for assessing distribution of sentiment scores between media and Twitter - Conducted a series of statistical tests (Bartlett, Levene's, Kruskal-Wallis) to find statistical evidence of necessary assumptions of sentiment scores
Lucy Lu	- Converted scraped article data from json file into database format - Cleaned Twitter data and extracted relevant columns for model fitting - Feed in Twitter data into VADER sentiment model and created dataframe from average sentiment score for each bias term by year - Explored Linear Regression and multiple Linear Regression models for statistical analysis - Wrote SQL to create tables and database in PostgreSql - Created visualization on sentiment score and word count
Vaishnavi Ravinutala	- Assisted team in scraping of news articles and design of SQL schema for holding media articles data - Led team in deciding which sentiment analysis model to utilize for articles and twitter dataset by trying hugging face BERT transformer, and finally settling on VADER model - Assisted in making graph visualizations for models, presentation, and report - Provided input and contributed throughout project and towards discussion of statistical models and analysis

9. Discussion and Conclusion

9.1 Lessons Learned

- VADER sentiment scoring allows us to infer the spectrum of positive and negative biases text
- Social media appears to have a smaller impact on election outcomes than initially anticipated
- There is no evidence that social media biases and news article biases align whatsoever

- small p-values aren't a sole indicator that a model fits a given dataset
- Assumption checking the data prior to model fitting is essential
- Media sentiment are surprisingly polarized while twitter sentiments are much more uniform
- Immigration has been the leading topic in news articles out of all of our bias categories over time
- Assuming independence among twitter posts and news articles is potentially a major limitation since a single event may cause multiple articles to be published from multiple news networks and generate a slurry of twitter posts in response to the event and/or the news article publication
- VADER is a pre-trained ML model thus there is no way to assess precision or tune parameters rending us incapable of assessing or improving model accuracy
- No evidence that twitter sentiment are related to Democratic vote counts (Table 10)
- A high R-squared value is not enough evidence of a model's goodness of fit
- 0.15 correlation between compound sentiment scores and normalized vote counts indicates a weakly linear relationship (Fig 7) and evidence that the data may fit other models better
- Evidence of non-constant variance can also be seen in a residuals vs fitted diagnostic plot

9.2 Surprising Difficulties

- Developing our own web scraping code for data acquisition and cleaning
- Finding an appropriate model given the many limitations of our data
- Conveying a clear message while also accounting for data complexity with data visualizations
- Deletion of archived news articles from news networks involved in legal battles limiting our data
- Any inference on relationship direction between election data and sentiment scores
- Graphs may be interpreted differently by different people therefore to make a conclusive statement it's better to gather a variety of evidence for it (statistics and graphs)

9.3 Future Work

- Add more biased terms and collect more data by searching for archived articles
- Explore better sentiment models that can handle article data better, since the VADER model is best toward social media text
- Explore more non-parametric models and statistical tests
- Capture demographic census data so that we can control for relevant variables in predicting election outcomes in different regions
- Scrap online text from other social media platforms such as Reddit and Facebook to get a more generalizable idea of public sentiments on bias topics
- Manually create a test dataset for VADER for accuracy testing, read 100 observations from tweets and articles for each bias group and identify bias language and compare to original scores, or potentially use a series of alternative models that allow us to assess accuracy in score assignments

9.4 Conclusion

Based on the results of our analysis, we find evidence of a discrepancy between news article sentiments and public Twitter sentiments. This indicates that major news networks have expressed opinions that differ from the prevailing public sentiment during the last three election cycles (2016-2020).

Furthermore, our Linear Regression model demonstrates that the collective compound sentiment score derived from Twitter data has a significant impact on the normalized vote count for Republican candidates in House elections. Specifically, for each unit increase in the compound sentiment score, reflecting more positive public sentiments on topics such as abortion, immigration, race, and socio-economics, there is a corresponding increase of 0.0108 in the normalized vote count.

In conclusion, our findings suggest that bias in public forums plays a noteworthy role in influencing election outcomes. The divergence between news article sentiments and public sentiment, along with the demonstrated effect of public sentiment on Republican vote counts, provides little evidence that increasingly polarized bias in the media has an effect on Republican vote counts.