**Final Report – Team 5**

**How Much Does Bias in the Media Affect Election Outcomes?**

Lucy Lu, 28273098, [junyil16@uci.edu](mailto:junyil16@uci.edu)

Tram La, 14166227, [latb@uci.edu](mailto:latb@uci.edu)

Nancy Carlson, 15744122, [nacarlso@uci.edu](mailto:nacarlso@uci.edu)

Deborah Rosa Franza, 72533099, [franzad@uci.edu](mailto:franzad@uci.edu)

Vaishnavi Ravinutala, 15538807, [vravinut@uci.edu](mailto:vravinut@uci.edu)

GitHub: <https://github.com/Political-Bias-Analysis>

**1. Introduction and Problem Statement**

How much does bias in the media affect election outcomes? As Americans we expect the election process to be transparent and unbiased to ensure we can make informed decisions about who we choose as our government representatives. To answer the question, this research looks to quantify sentiment on four topics, abortion, race, immigration, and socioeconomics to understand if sway on these topics affects election outcomes. News media outlets hold an undeniable influence in shaping public opinion. However, the rise of social media as a public discourse forum and its ability to amplify correct and incorrect information also has the power to shape a narrative, and frame issues that can also sway public opinion. As these two forums compete for the public’s attention, this research also wants to understand do the sentiments expressed by media outlets CBS, CNN, FOX, and NPR on the four topics, influence public opinion or does public opinion, as expressed on Twitter, influences the sentiments media presents to the public.

The results of this research indicate the sentiments expressed about abortion, race, immigration, and socioeconomics subjects show a range of sentiment variability or opinion with socioeconomics giving the highest variability and race giving the lowest. Additionally, sentiments expressed on all topics by news media are overwhelmingly positive compared to sentiments expressed on Twitter which were negative. Furthermore, the sentiment scores in the media article group had more effect on the compound sentiment score than the Twitter group. The quantification of bias in the media articles obtained for this research was inconclusive due to this model’s data limitation on 3 years of election results. However, the Twitter model found with one unit increase in the collective sentiment score of all topics, we see a tenth of a unit increase in the normalized vote count for Republican candidates running for House office seats. In contrast, when the same model was run on House elections for Democratic candidates no increase in voter count was found. In conclusion, exploring different model options, and confirming sentiment score accuracy is recommended to corroborate these findings.

**2. Related Work:**

For the statistical analysis portion of this project, we used both Linear Regression and ANOVA 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 counts, we discovered aPolishh paper by [Krochmal](https://arxiv.org/ftp/arxiv/papers/2010/2010.03493.pdf) that used a number of OLS regression models to determine if sentiments in polish tweets were a significant predictor in vote share in poviat (which is known as a regional government in Poland) while taking into account other necessary socio-economic covariates for each specific poviat (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 ANOVA**

In a paper by [Park](https://sci-hub.ru/10.1108/jhtt-08-2016-0042) from the University of Central Missouri, an ANOVA model 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, our ANOVA model 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**

Two databases were created for this research project: “electiondb” and “sentimentdb”. The electiondb is 11 MB in size and consists of two tables named “results”, (table 1) and “voters”, (table 2). The sentimentsdb is 86 MB in size and consists of two tables named “articles” (table 3) and “twitter” (table 4).

The dependent variable which addresses our research question is the normalized mean vote count of registered voters to national United States election results. Data on Presidential, Senate, and House election results, by state, and party were obtained from the “Federal Election Commission United States of America.” The independent variables used from this data to build our models are year and state. Election results for the year 2022 are not yet available.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **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/>

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>.

A picture containing text, screenshot, font, number

Description automatically generated

To get a better sense of where party affiliation falls the data is grouped by region. This gives us a sense of how a voter's geolocation influences party affiliation.

**Figure 1: Mean Vote Percentage by**

**Region and Party Affiliation: 2010 to 2020**

News media article data were collected by scraping the websites of CBS, CNN, FOX, and NPR news. The code used to scrape each website targeted the subject of elections and included keywording that related to one of the four topics of interest: abortion, race, immigration, and socioeconomics. A total of 15,914 articles in total were collected. The CSV files are organized by year and topic. After each file was cleaned and tokenized, it was then fed through the VADER algorithm to assign positive, negative, and compound sentiment scores to each article text. The independent variables collected from the news articles table to build our model are year, news source, the tag of the topic of interest to each article, and its sentiment score. Although three sentiment scores were collected for each article, our models used the compound score inclusively.

<https://www.cnn.com/> <https://www.foxnews.com/>

<https://www.npr.org/> <https://www.cbsnews.com/>

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **year** | **source** | **published** | **content** | **abortion** | **race** | **immigration** | **socio\_eco** | **compound** |
| 2016 | CBS | 2016-11-16 | in the wake of the election the antidefamation... | planned parenthood | None | None | None | -0.9989 |
| 2016 | CBS | 2016-12-15 | u intelligence source be not just point the fi.. | planned parenthood | None | None | None | 0.8768 |

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

A picture containing text, screenshot, diagram, plot

Description automatically generated

**Figure 2: Annual Total Number of Media Articles**

Figure 2 shows the number of media articles collected on each topic of interest. The topic of immigration stands out as a primary story covered by news organizations. Although our research found news articles had more influence over election outcomes than Twitter posts, there was no direct link to any specific news media topic that affected election outcomes.

The Twitter data was provided by Professor Chen Li from the University of California at Irvine’s Information and Computer Science department. We requested Twitter data that covered our topic of interest during the years 2016 to 2022. The Twitter data came to us unlabeled in the form of CSV files with 33 columns and approximately 59, 861 tweets. The independent variables used from this data to build our models are year, state, and topic tag with its compound sentiment score.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **year** | **created** | **state** | **owner** | **tweet** | **abortion** | **race** | **immigration** | **socio\_eco** | **compound** |
| 2020 | 2020-08-01 00:19:13.0 | OR | 1033162882471477248 | not to worry he already demonstrate he not sma… | None | Black | None | None | -0.4082 |
| 2020 | 2020-08-01 14:19:56.0 | OK | 4805323446 | please explain to me why trump close the china… | None | None | Wall | None | -0.9377 |

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

A graph of a number of tweets

Description automatically generated with low confidence

**Figure 3: Annual Total Number of tweets**

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

A picture containing text, line, plot, diagram

Description automatically generated**Figure 4: Comparison of Twitter Weighted Sentiments to Media Articles**

**Grouped by U.S. Region & Source**

An analysis of weighted sentiment scores (figure 4) helps us see which U.S. region and which news source is speaking the most about a topic. Here we see the Midwest and CBS give the most mentions on immigration. By summing the weighted compound scores the magnitude of the positive versus negative sentiments over time are revealed.

**4. Overall Technical Approach**

**4.1 Scraping Websites and Collection Data**

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

**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 identify the subject of the tweet as one of the four research topics.

The election data came to us in Excel workbooks by year. The data is organized in a slightly different manner each year. The most recent years had all three federal office elections - President, Senate, and House election results in one Excel Workbook, with the results of each office on a separate sheet within the workbook. While the older years had separate workbooks for each office. The worksheets were defined and put into a dictionary of data frames that were then extracted one at a time to clean and analyze individually.

**4.3 Sentiment Scoring and Analysis**

Once the text from Twitter and news articles finishes the NLP steps, it is ready for Sentiment Analysis. Valence Aware Dictionary and sEntiment Reasoner (VADER) is a popular tool to assess the sentiment of a text. VADER is especially skilled at web-based and social media text since it is a lexicon, rule-based feeling analysis tool that can handle informal language, slang, and emotions. One of VADER’s limitations is its struggle to capture nuanced sentiments such as sarcasm, irony, or sentiment expressed through complex sentence structures. The VADER algorithm assigns a compound sentiment score to each row of the Twitter and Article tables.

The compound sentiment score ranges from -1 to 1, where -1 represents the most negative sentiment and +1 represents the most positive sentiment, and 0 is a neutral sentiment.

**4.4 Exploratory Data Analysis (EDA)**

Several aggregated tables were built to look at the election, media articles, and Twitter data in various ways. We created visualizations such as histograms, bar charts, box plots, scatter plots, choropleth maps, and density plots to see distributions, relationships, and patterns that are within the data. We derived new features such as normalizing the vote count value with voter registration. We extracted model variables to build plots on model summary results. These visuals helped us make observations and interpret the data to build the storyline of our findings.

**4.5 ANOVA**

To explore the sentiment score relationship between news media and Twitter, the Analysis of Variance (ANOVA) statistical model is run on the sentiment scores of each topic. This model tests the null hypothesis, “There is no difference between media article sentiments and Twitter sentiments.” If a difference is detected, the alternative hypothesis, “There is a difference between article sentiments and Twitter sentiments” is accepted. Four sentiment tables were aggregated by bias topic to contain both the sentiments of news articles and tweets. (Example: Tables 5 and 6)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **year** | **bias** | **compound** |  | **year** | **bias** | **compound** |
| 2020 | Abortion Tweet | -0.4043 |  | 2020 | Immigration Tweet | 0.2732 |
| **…** | **…** | **…** |  | **…** | **…** | **…** |
| 2016 | Abortion Article | 0.9998 |  | 2016 | Immigration Article | 0.9406 |

Table 5: Abortion Sentiments 2016-2020. 7328 rows Table 6: Immigration Sentiments 2016-2020 17271 rows

**4.6 Tukey’s Honestly Significant Difference (HSD)**

To determine which group has a greater compound sentiment score, media articles, or tweets, Tukey’s Honestly Significant Difference (HSD) test is performed. The summary results give the mean difference between the two groups showing which group has the higher mean sentiment score.

**4.7 Linear Regression**

To quantify each topic’s sentiment scores from both news media articles and public tweets and how the sentiment score affects election outcomes, a separate linear regression model (LR) is run on both groups. The model’s dependent 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 the aggregated table below. This table matches the year columns from the media article table and the year mean vote count from the voting results table. Matching the two tables by the year column limited the model to having only three dependent variables that repeated across the independent variables. 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 used 140 dependent variables to repeat across the independent variables. This 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)

**4.8 Model Validations**

Checking the ANOVA model assumptions involved a Q-Q plot to check for normality, residual plots to assess linearity and homogeneity of the variances, and scatter plots on the residuals of both group’s news article sentiments and Twitter sentiments to assess the independence of the observed values.

The LR model assumptions that were checked in our model are: Linearity, Multivariant Normality, Multicollinearity, and Homoscedasticity. Multiple graphs were created to check for these assumptions. Details are model validations are mentioned in section 6.

**5. Software**

Postgressql

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, ggplot

Usage: Create data visualization plots in R

ReactJs

Usage: We are planning to use ReactJs as our front-end framework. Currently, we have the skeleton setup (such as headers and footers). Our next plan is to use publicly available libraries to create interactive visualizations of our data.

Node Js/Express

Usage: The backend framework that acts as a middleman between our database and the front-end.

Sequelize

Usage: A library that allows us to write Postgres queries from our backend

Recharts

Usage: A React Js library that allows us to create interactive graphs

Issues/Frustrations: Limited graphs provided.

Leaflet

Usage: A React Js library that allows us to create the choropleth graph displaying election results by state

Issues/Frustrations: It took us a while to understand how we should organize the data to input into the given graph component.

Solutions: We solved this issue by reading tutorials and looking at how they would structure their data.

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.

Git/Github

Usage: We use Github to maintain our code version control. We have four independent repositories to store our data, scraping scripts, ML models, and the dashboard.

Issues/Frustrations: It is difficult to work with Git for such a big project when all of us have limited knowledge of how this software works. For instance, we had to deal with merge conflicts, pull/push requests, etc. However, we are getting more comfortable with using Git as time goes on.

Solutions: To avoid code writing over each other, we worked on separate branches and merged accordingly.

**6. Experiments and Evaluation**

**6.1 ANOVA Models**

The ANOVA model provides insights into the impact of the ‘bias’ variable on the ‘compound’ sentiment scores when compared across the four different groups: abortion, race, immigration, and socioeconomics. Table 8 shows a summary of the four ANOVA models. We can see from these results that across all groups there are differences in sentiment scores between the media articles and Twitter tweets. Figure 5 shows the variability of each model or group’s sentiment scores. A large mean square value indicates a wide variety of sentiments or opinions on this topic, whereas a smaller mean square value indicates the sentiment values or opinions are similar.

|  |  |  |  |
| --- | --- | --- | --- |
| **ANOVA Model** | **Mean Square** | **F-statistic** | **p-value** |
| Socioeconomic | 847.8397 | 2107.1243 | 0.0 |
| Immigration | 804.7663 | 1518.8621 | 1.654339e-318 |
| Abortion | 469.6311 | 1147.7965 | 1.150284e-233 |
| Race | 345.1253 | 714.9507 | 2.942149e-152 |

A picture containing text, screenshot, diagram, plot

Description automatically generatedA picture containing text, screenshot, diagram, line

Description automatically generatedTable 8: ANOVA Models Summary

Figure 6: ANOVA Compound Mean Figure 7: ANOVA Sentiment Variability

**6.1.1 ANOVA Model Assumptions**

To check for the validity of the ANOVA model the following assumptions must be checked. Independence of the observed data. Normality, which is the distribution of the residuals. Homogeneity of variances in that the variability of the residuals should be similar across both groups, article sentiments to Twitter sentiments, and linearity of the data which expresses the relationship between the independent variable (bias) and the dependent variable compound sentiment score. Once analyzed, we conclude this model shows two distinct groups where data is clustered.

**6.2 Tukey’s HSD test**

This test provides a summary table that includes the mean difference between the groups, the standard error, the lower and upper confidence intervals, and the p-values. By examining the p-values, we can determine which bias variable (abortion tweets or abortion articles) has a more significant effect on the compound sentiment score. This test showed in all four bias groups, the article sentiments had a greater effect on the compound sentiment scores.

**6.3 Linear Regression Model**

Several linear regression models that examine the relationship between sentiment scores on bias topics and the dependent variable normalized vote count were attempted. Because the articles data table only had three dependent variables, we did not pursue a Linear Regression model on this data.

A picture containing text, screenshot, diagram, plot

Description automatically generated However, the linear regression model that examined the Twitter sentiment scores to vote counts had 140 vote count dependent variables because it matched tweets to election results on the year and state columns. This model gave a small result on the average bias sentiment scores. In this case, both R-squared and Adj. R-squared are 0.983, which means that the model explains 98.3% of the variation in the dependent variable (norm\_vote\_count). This high value suggests that the model fits the data extremely well. The compound coefficient is statistically significant (p < 0.005) with a value of 0.0108. 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 gave a slight positive result in Republican voting results on House elections.

**6.3.1 LR Model Assumptions**

The Twitter linear regression model revealed a cluster of data on the compound and normalized vote count axes, indicating a positive relationship but with limited strength. Variability above a certain level suggests the influence of other factors 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 and year variables, indicating independent information. The model exhibited slight underestimation and heteroscedasticity, suggesting variable prediction variance. Further investigation is needed to assess the impact and explore remedies. Alternative models should be considered for stronger results.

**7. Notebook Description**

Data on election outcomes, news media articles, and tweets posted on Twitter between the years 2010 to 2022 were collected from various sources. Once correctly labeled the election data was fed into the database called “electiondb.” The data from news articles and the tweets from Twitter were sorted and labeled into one of the research topics: abortion, race, immigration, socioeconomics. The text from both tables is run through a series of NLP cleaning steps and then fed into the VADER algorithm to assign sentiment scores. The sorted data along with sentiment scores are uploaded to a database called “sentimentdb.” To explore the sentiment score relationship between news media and Twitter, the ANOVA statistical model is run on the sentiment scores of each topic. This model tests the null hypothesis, “There is no difference between media article sentiments and Twitter sentiments.” If a difference is detected, the alternative hypothesis, “There is a difference between article sentiments and Twitter sentiments” is accepted. To determine which group has a greater compound sentiment score, media articles, or tweets, an HSD test is performed. The summary results give the mean difference between the two groups showing which group has the higher mean sentiment score. To quantify each topic’s sentiment scores from both news media articles and public tweets and how the sentiment score affects election outcomes, a separate linear regression model (LR) is run on both groups. Tables were formed that merged the sentiments of each topic with the normalized vote count of registered voters to the vote counts of candidates running for seats in the Federal House of Representatives (House) from both the Republican (REP) and Democratic (DEM) parties. Both models identified the dependent variable as “vote count”. The media articles LR model identified the independent variables as, “year”, “source”, “topic tag”, and “compound sentiment score”. The Twitter LR model identified the independent variables as, “year”, “state”, “topic tag”, and “compound sentiment score”.

**8. Member Participation**

Tram wrote the script and led the team in scraping URLs for article content. Next, as a team, we worked on cleaning the data and running the articles and tweets through the VADER sentiment algorithm. After creating the databases, we worked on exploratory data analysis by creating various aggregated tables from the “electiondb” and “sentimentdb.” This was followed by exploring various statistical models that would answer our research questions. Lastly, Tram developed the dashboard website using ReactJs as the front-end framework. See the appendix for team participation details.

**9. Discussion and Conclusion**

Team Summary

1.

2.

3.

4.

Appendix:

|  |  |
| --- | --- |
| **Name** | **Task(s)** |
| 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 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 and twitter data. * Created graphs on article and twitter counts. * Created 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 statistically significant ANOVA models and accompanying interpretations * Validated ANOVA model using Boxplots, Ridgeline graphs, and Tukey Test * Assisted in hypothesis formulations * Created various data visualizations for assessing distribution of sentiment scores between media and twitter * Explored media & election data in its ability to meet assumptions for linearity |
| 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 * Led team in deciding which sentiment analysis model to utilize by trying hugging face BERT transformer, and finally settling on VADER model |