LEVERAGING REDDIT AND NEWS ARTICLES FOR POLITICAL LEANING CLASSIFICATION IN MACHINE LEARNING

Salma Aly[1,2], Christopher Matthews[1,2], Michael Mistarz[1,2], Jamia Russell[1,2,†], Nikita Sharma[1,2]

[1] Northwestern University School of Professional Studies
Master of Science in Data Science Program
633 Clark St. Evanston, IL 60208

[2] Github - NLP Model Political Classification

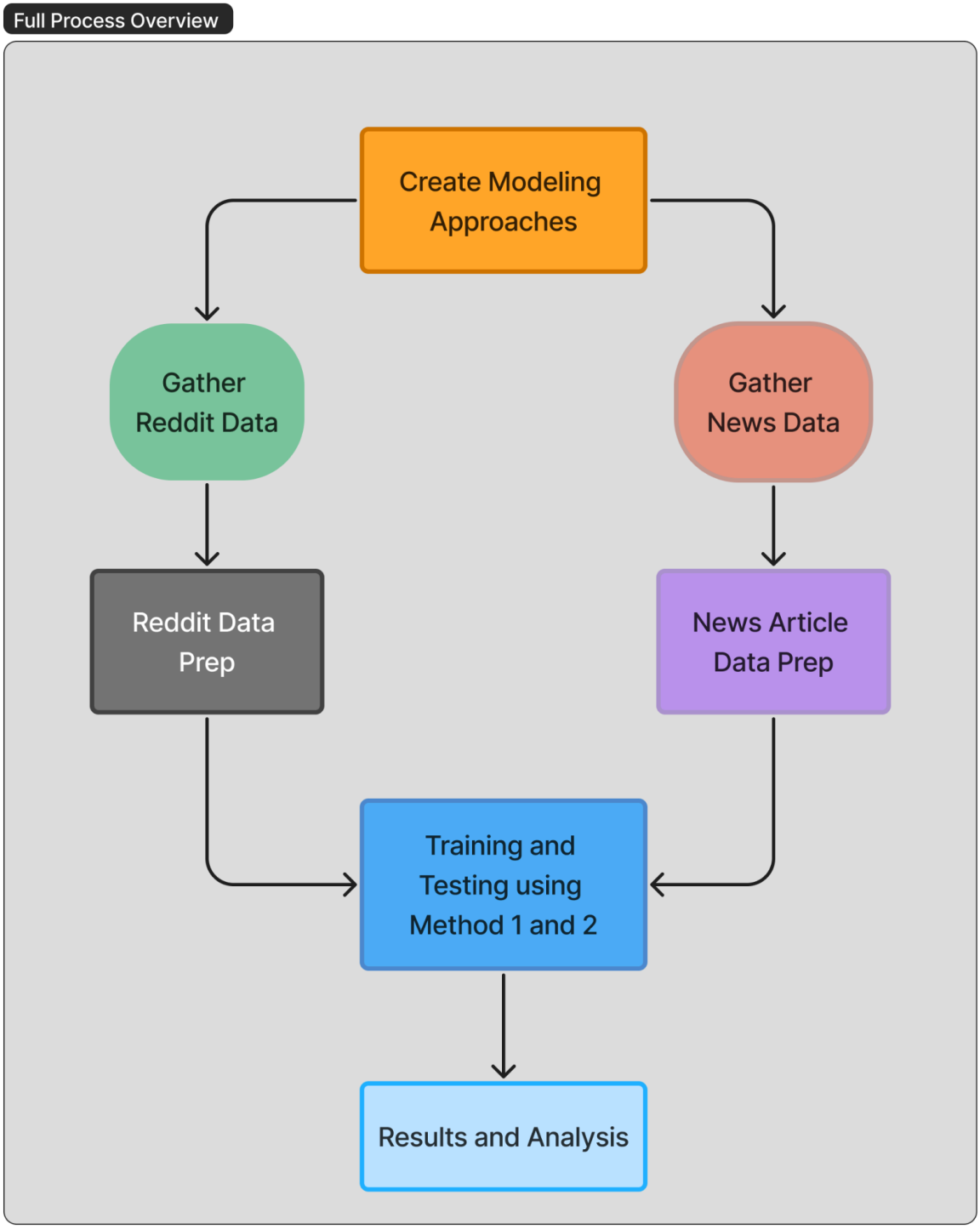
†Address to which correspondence should be addressed:
jamialashe@gmail.com

Abstract

Inferred media bias within US news and media sources has led to the polarization of the political climate. This has been amplified by an increase in the number of people whose main source of news information comes from social media sources, leading to an era of misinformation and mistrust. With the deterioration of trust in media sources, citizens have sought a way to measure how a given news source shows bias to their consumers. Determining bias within information sources has become necessary to correcting the information landscape, and the purpose of this study is to set the groundwork for bias determination in news sources.

This study was designed to determine whether political bias can be classified and measured by comparing the texts of news articles with text comments made in both conservative and liberal subreddits on the Reddit social media site. The comments from each of the top posts of the last year from nine conservative and nine liberal subreddits were aggregated into a text corpus to build the Reddit data. Articles from news sources from across the political spectrum as rated by AllSides, an industry leader in detecting media bias, were aggregated into a news data corpus. These two corpora were vectorized using Term Frequency Inverse Document Frequency (TF-IDF) and analyzed using different statistical techniques to determine whether one corpus could be used to detect and predict political bias in the documents within the other corpus.

Keywords: TF-IDF Vectorization, Bias Identification, Linguistic Variability, News Media Bias, Subreddit Language, Support Vector Machines, Multinomial Naive Bayes



Graphical Abstract. Process flow diagram for analyzing political news bias using AllSides political bias ratings and political subreddits. Colors of the elements in the process flow diagram are consistent and each step is broken down in other diagrams throughout the paper.

Table of Contents

I. Introduction & Problem Statement

Over the last few decades, the transition to a digital configuration has empowered online news channels and social media's role in disseminating misinformation and political bias leading to only 69 percent of Americans having little to no trust in traditional media today (Brenan 2024). Political bias in mainstream media appears to shoulder much of the blame, leading readers and viewers to assume that media sources are spinning stories to benefit viewership and political parties (van der Linden, Panagopoulos, and Roozenbeek, 2020). The polarization of media outlets across the political spectrum has caused concern regarding the radical opinions of liberal leaning and conservative leaning consumers. These populations are susceptible to the media's influence on public opinion, shaped narratives, and controlled democratic discourse. Recognizing and quantifying this bias using Natural Language Processing (NLP) techniques including TF-IDF vectorization, tokenization, and sentiment analysis, then analyzing the models using Machine Learning(ML) techniques like logistic regression, Multinomial Naive Bayes (Multinomial NB), and Support Vector Machines(SVM), can help identify political bias across media forms.

The aim of this paper is to examine machine learning model classification performance on linguistic quantifiers that exhibit systematic patterns in language and language identification using a model trained on existing data relating to political identity. Language and expression are critical in comprehension and understanding of readers as they form and voice their own opinions.

Existing studies examining Reddit, a widely used community-based platform, and similar platforms have highlighted the impact of word choice and ideological bias through language on policy (De Arruda, Roman, and Monteiro 2020). Focusing on language patterns and content, our

model identifies language patterns consistent with conservative or liberal positions across news articles and Reddit comments to provide insight on the effect of language on ideological classification.

The model was prepared to handle large scale textual data and distinguish language consistencies through wording and tone enabling it the ability to ascertain ideological slants.

II. Literature Review

Researchers have no shortage of text availability or tools to limit the amount of insights to build from textual data – especially when it comes to political bias. Multiple resources have been used before to assess political bias and leanings in written platforms. For instance, Arruda, Roman, and Monteiro (2020) and Mayopu, Wang, and Chen (2023) used news articles to develop their research on selection, coverage, statement bias and classification of real and fake news respectively. Reddit was used by Ferrer et al. (2020) and Zahrah, Nurse, and Goldsmith (2022) to investigate sentiment and for topic classification. Iyyer et al. (2014) used Recursive Neural Networks (RNN) to detect political leaning in text and proved that RNN could lead to higher accuracy as compared to the previously used techniques at that time such as bag-of-words-models.

2.1 Detecting Bias and Political Leaning

Detecting political leaning and bias has been one of the interesting areas to investigate using NLP techniques. In 2014, Iyyer et al. investigated detecting political leanings using text data. Their data collection consisted of US Congressional Bills and Tweets from the social platform X. They used word embeddings and employed (RNN) to detect political ideology. Their research advanced the field after models such as bag-of-words were used for similar research problems. They proved that political leaning is better detected in complex structures such as

sentences rather than isolated terms.

Ferrer et al. (2020) used Reddit data to examine the relationship between language, bias, and political discourse in digital spaces. They used word embeddings and applied K-means clustering to discover biases in different Reddit communities with a focus on topics such as gender and race. Not only was Ferrer et al. (2020) work able to detect bias, it was also able to detect the polarization tendencies in the studied Reddit communities.

2.2 Sentiment Analysis in Political Context

According to Antypas et al. (2023) negativity in social media platforms spreads faster. In their work, Antypas et al. (2023) analyzed political tweets in different languages from the social media platform X (formally known as Twitter). They also examined a number of sentiment classifiers such as the SVM, Neural Networks and a lexicon approach using VADER. Neural Network based model Bertweet-Sent unsurprisingly outperformed the two other modeling approaches thanks to its deep learning capabilities that allowed it to detect the complex patterns in the textual data.

In 2023, Alfonso and Rarasati analyzed X, a social media platform, posts to assess how they aligned with 2024 Indonesian presidential surveys. They used TF-IDF for feature extraction and a 10-fold cross validation on an SVM model. Using F1-score and Pearson's correlation coefficient metric, Alfonso and Rarasati (2023) found that X posts can provide a relatively accurate reflection of the public sentiment on the presidential election in real-time. The paper also addresses the ethical implications of detecting language biases, emphasizing the importance of understanding these dynamics to promote media literacy and foster informed citizen engagement in online discussions.

2.3 Detecting Political Misinformation and Fake News

Presidential elections cycles are a time where a significant amount of polarization, misinformation and fake news can take place. Not only does this kind of news spread fast, it also has negative impacts on communities (Mayopu et al., 2023). Mayopu et al. (2023) analyzed political fake news during the 2016 U.S presidential elections. They used a combination of Natural Language Processing (NLP) and Singular Value Decomposition (SVD) techniques to conduct their research. Through their research, they were able to distinguish between real and fake news. Das et al. (2023) review on the other side, focused on comparing NLP performance for news fact checking to human fact checking. They emphasized the importance of using a hybrid system for fact checking as opposed to a solely automated system. According to Des et al. (2023), taking this "humans-in-the-loop" approach for fact checking gives results more credibility while allowing for a scalable performance.

III. Data

3.1 Overview

For this analysis, data from both Reddit and news sources outlets were collected over a 12-month time horizon. This time range was chosen to capture recent and relevant discussions that reflect current political climates and trends. By focusing on the latest year, we ensure that our analysis remains timely and relevant, capturing recent shifts in public opinion and media coverage related to key events such as elections, policy changes, and social movements.

In political analysis, combining diverse information sources can provide a comprehensive view of public opinion and media narratives. News headlines, recognized for their formal and timely reporting on current events, offer a broad and credible overview, often upholding editorial standards and factual accuracy (Das et al., 2023). In contrast, Reddit serves as an unfiltered

reflection of public opinion. Through longer posts that encourage in-depth discussions, Reddit fosters nuanced discourse on political topics, enriched by community-driven insights and background analysis that may not surface in traditional media (Zahrah, Nurse, and Goldsmith, 2022). Reddit has become a popular resource for Natural Language Process (NLP) studies, such as those focused on classifying discussions around mental health and domestic abuse (Ferrer et al., 2020). These studies illustrate that platforms like Reddit are not merely reflections of offline sentiment but increasingly serve as active spaces for shaping contemporary ideologies and social processes.

Here, Python was used to pre-process the data, preparing it for analysis in this study. According to Mayopu, Wang, and Chen (2023), the pre-processing stage entails several steps, including tokenization, data cleaning, lemmatization, vectorization and word frequency counting. Through this structured NLP approach, we were able to refine and prepare the text for meaningful exploration.

3.2 Bias Segmentation in Pre-processing for Media Outlets

We first explore data pre-processing in which we aim to segment and control our data analysis of identifying political biases in media outlets. When sourcing media outlets, it was essential to understand the current political landscape of media outlets and how they can be segmented. Researchers have been able to uncover and segment political bias in major US based news outlets by interpreting the average view of the American and dissecting recurrent thematics. Thematics refers to recurring themes that indicate ideological leanings in a body of text, such as media content.

Political Topic	Left-Leaning Thematics	Right-Leaning Thematics
Government Services and Offerings	• Medicare, Social Security, student debt forgiveness, unemployment benefits	• Personal responsibility, limited government aid, self-reliance
Protection of Underserved or Oppressed Groups	• Consumer rights, environmental protection, anti-discrimination, tax benefits	• Reduced government spending, deregulation, opposition to welfare state
Multiculturalism and Wealth Distribution	• Multiculturalism, affirmative action, immigration policy, human rights	• Traditional family values, state sovereignty, individual rights
Federal vs. State Power	• Federal laws for equity, protection of underrepresented groups	• Increased state power, constitutional rights, rejection of federal mandates
Economic Policy and Regulation	• Tax advantages for low-income, regulations for equity	• Reduced regulation, lower spending, opposition to restrictive policies

Figure 2: AllSides Media Bias Thematics. Chart that compares and contrasts thematic elements in left-leaning and right-leaning media outlets.

According to AllSides, a media bias and detection platform, sources with a left or liberal media bias reflect positive reviews on:

1. Generous government services such as Medicare, Social Security, student debt forgiveness, and unemployment benefits.
2. Federal laws and economic policies protecting underserved groups, including consumer protections, environmental and abortion rights, anti-discrimination laws, and tax benefits for lower-income individuals.
3. Multiculturalism and wealth equity, highlighting support for inclusive immigration policies, and access to healthcare, house, and clean water.

Sources with a right or conservative media bias reflect positive reviews on:

1. Traditional family values and the sovereignty of the individual over the collective. This can include positive outlooks on the reliance of personal responsibility rather than government intervention.
2. Limiting the government's scope and power to manage domestic, economic and social affairs. This can include positive outlooks on decreasing government spending and their involvement in economic issues, or social issues related to the “welfare” of the state such as gender identity and inequality.

Using political thematics as a guidestone to label media bias, the AllSides Media Bias Chart was developed after years of analysis and review to illustrate and structure the landscape of US media outlets tendencies. When pulling sample news article documents that would serve as the underlying data in the construction of the ML model, two corpora were created to focus on the extreme liberal left-leaning thematics and the conservative right-leaning thematics.

AllSides Media Bias Chart

Figure 3: AllSides Media Bias Chart

3.3 Data Collection and Preparation for News Media Outlet

The data collection process for news articles begins with obtaining an API key to query the NewsAPI, a service that provides access to recent and historical articles from a wide range of

news sources. We performed targeted searches using the following news sources to gather relevant articles:

- US Today
- CNN
- MSNBC
- Fox News
- American Conservative

The extracted content is organized into a structured DataFrame, where each entry includes essential metadata such as source, article title, and article text.

Once the articles are compiled and structured, the dataset was then preprocessed by:

1. Breaking down text into tokens
2. Stripping any stop words and punctuation
3. Identifying any keywords and phrases for further analysis
4. Vectorizing text using Term Frequency-Inverse Document Frequency (TF-IDF).

This process established a consistent feature space and weighting scheme.

political_leaning	source	text	title	type
Conservative	FOX News	3 reasons why Kamala Harris still can't define...	FOX News_3_reasons_why_Kamala_Harris_still_can...	news_articles
Conservative	FOX News	5 key takeaways from Kamala Harris' '60 Minute...	FOX News_5_key_takeaways_from_Kamala_Harris___...	news_articles
Conservative	FOX News	64 days: Kamala Harris has yet to do formal pr...	FOX News_64_days_Kamala_Harris_has_yet_to_do_f...	news_articles
Conservative	FOX News	79 days: Kamala Harris has yet to do formal pr...	FOX News_79_days_Kamala_Harris_has_yet_to_do_f...	news_articles
Conservative	FOX News	81 days: Kamala Harris has yet to do formal pr...	FOX News_81_days_Kamala_Harris_has_yet_to_do_f...	news_articles

Figure 4: Example screenshot of the news corpus DataFrame used in the evaluation.

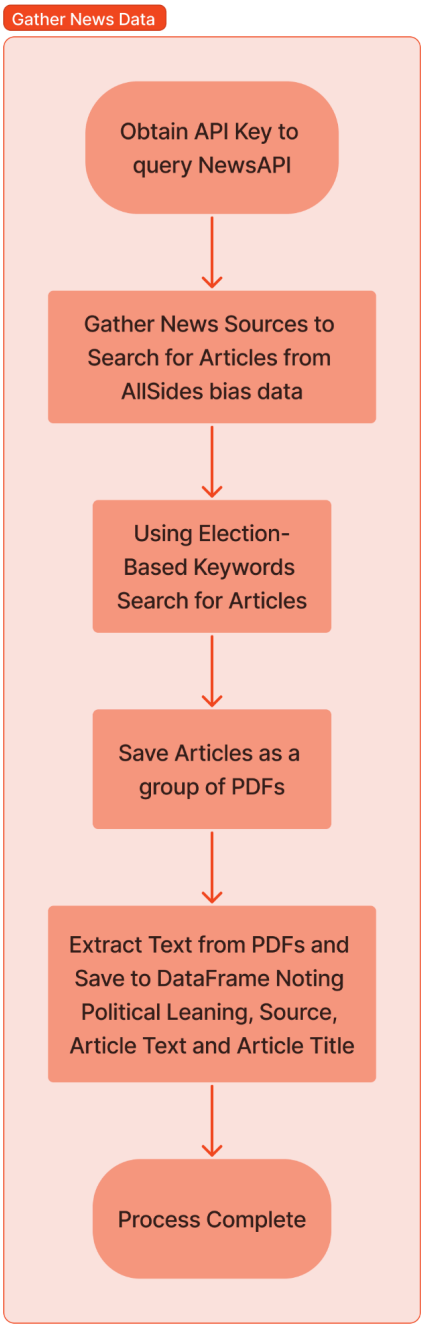


Figure 5: Gather News Data. Process flow diagram displaying news article retrieval using NewsAPI and Python libraries.

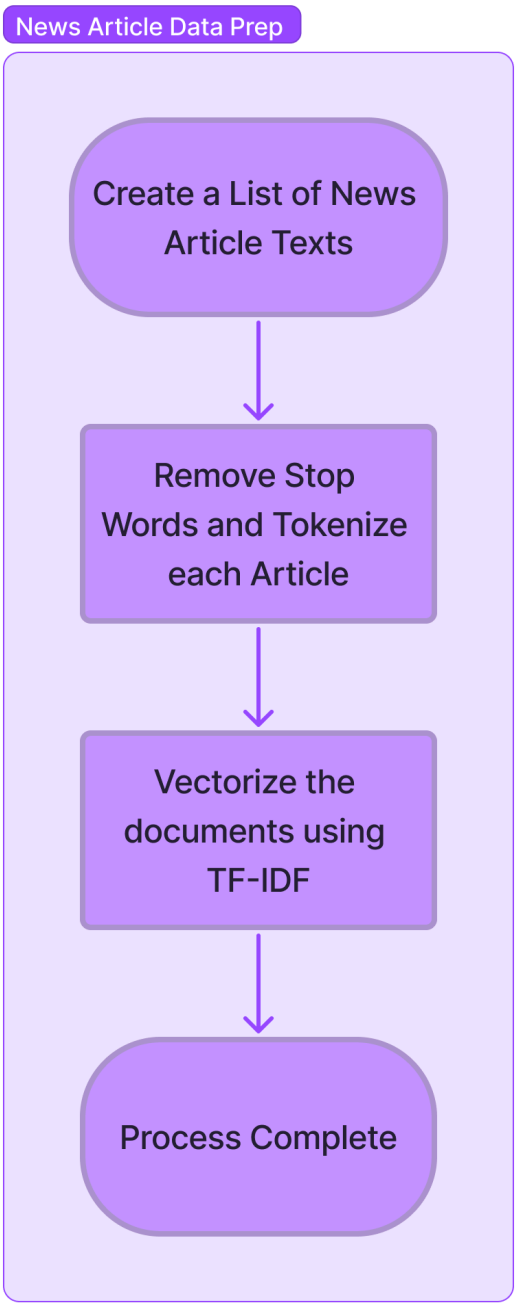


Figure 6: News Article Data Prep. Process flow diagram displaying the process of creating the news article dataframe implementing natural language processing (NLP) techniques.

3.4 Bias Segmentation in Pre-processing of Social Media Outlets - Reddit

We next explore data pre-processing in which we aim to segment and control our data analysis of identifying political biases in a popular social media and discussion platform called Reddit. Members of the community can submit content such as text posts, pictures, or direct links, which is organized in distinct message boards curated by interest communities. These 'subreddits' are distinct message boards curated around particular topics, such as /r/pics for sharing pictures or /f/funny for posting jokes (Ferrer et al. 2020; Iyyer et al. 2014; Das et al. 2023).

Researchers have identified Reddit as a valuable platform for NLP studies, primarily due to its structured organization of discussion spaces, known as "subreddits." These specialized forums foster distinct ideological communities (Ferrer et al., 2020; Iyyer et al., 2014; Das et al., 2023), where political themes emerge organically, reflecting a range of perspectives from left-leaning to right-leaning. This organization allows researchers to effectively classify and curate data for studies focused on detecting media bias. For example, Reddit subreddits associated with specific political leanings were compiled into respective corpora, providing classified datasets that support accurate bias analysis. Further details on the subreddit selections for left-leaning and right-leaning corpora are provided in Appendix A.

3.5 Data Collection and Preparation for Social Media Outlets - Reddit

To ensure a balanced representation of political perspectives, we selected nine subreddits, categorized as either conservative (right-leaning) or liberal (left-leaning). This selection process was guided by the subreddit's political leanings, user engagement, and topic relevance to political discourse. For each subreddit, we gathered the top posts from the previous year to capture high-engagement content likely to reflect core community perspectives. URLs for these

posts were scraped, and the top comments were collected to provide a focused yet diverse sample of opinions within each subreddit. The comments were then compiled into a structured DataFrame, capturing essential metadata such as subreddit name, political leaning, post title, and comment text.

Following the data-preprocessing methodology for NLP analysis, the dataset was then stripped of any stop words and punctuation and then tokenized to filter and find keywords or phrases for further analysis. As a data cleaning step, comments with 1 word or less were dropped. Overall, we had 7,427 comments aggregated in this DataFrame.The final step included vectorizing text using the text using Term Frequency-Inverse Document Frequency (TF-IDF). This transformation converted the textual content into numerical features to establish a consistent feature space and weighting scheme.

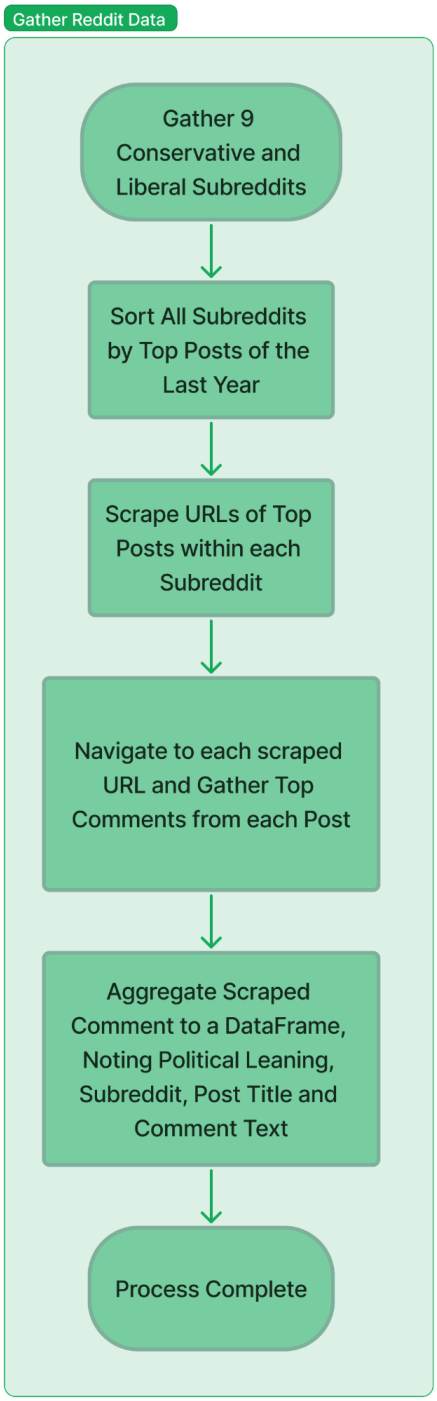


Figure 7: Gather Reddit Data. Process flow chart displaying the retrieval of Reddit comments in liberal and conservative subreddit groups using Selenium.

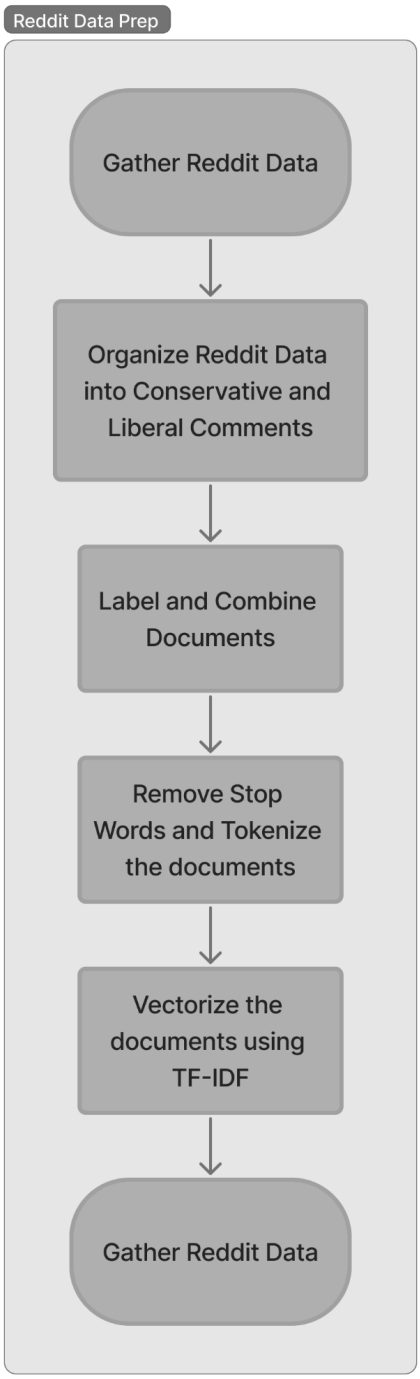


Figure 8: Reddit Data Prep. Process flow diagram displaying creation of subreddit comment database using tokenization and TF-IDF vectorization.

3.6 Data Exploration

3.6.1: Sentiment Analysis

In this section, we investigate the sentiment difference in the language used across the four groups: Reddit conservative, Reddit liberal, news conservative, and news liberal. The goal is to first compare and contrast between professional media tones and community written comments. We then thoroughly investigate the difference in tone between the different political groups.

VADER (Valence Aware Dictionary and Sentiment Reasoner) from the NLTK (Natural Language Toolkit) Python package was used to perform this sentiment analysis. VADER is known to be sensitive for web-based media thanks to its ability to analyze contextual information and language nuances (Hutto and Gilbert 2014).

After running sentiment analysis, we found that news articles tend to have a relatively positive sentiment compared to Reddit comments. The difference in sentiment between conservative and liberal Reddit comments was more pronounced with the conservative sentiment leaning towards a neutral tone and the liberal sentiment leaning towards a positive tone. Figure 9 shows a summary of the mean sentiment scores for each group:

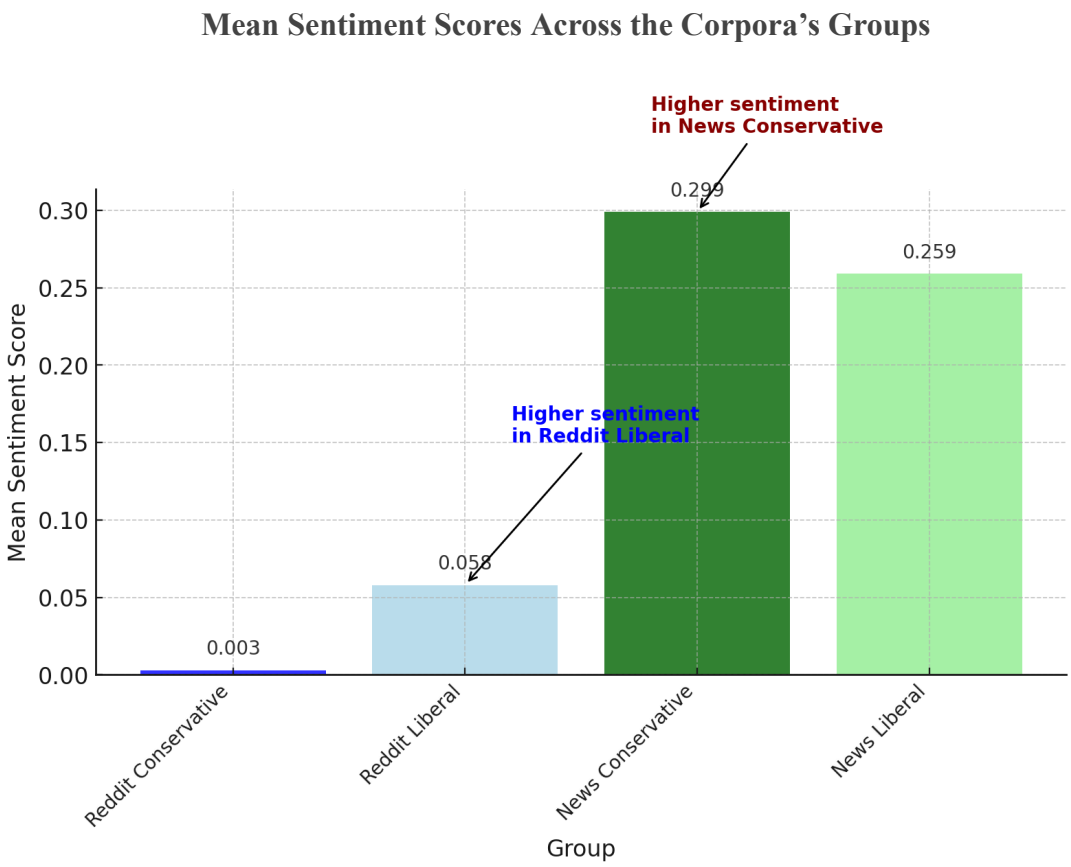


Figure 9: Comparison of Mean Sentiment Scores Across Groups

3.6.2: Discovering Topic

In this section we analyze different topics discussed by the different political leaning groups across Reddit and news outlet content. We used BERTopic for this analysis. BERT (Bidirectional Encoder Representations from Transformers) leverages embeddings to capture similarities between documents based on their topics (Wu et al. 2024). Figure 10 shows a comparative summary of keywords or topics across corpora.

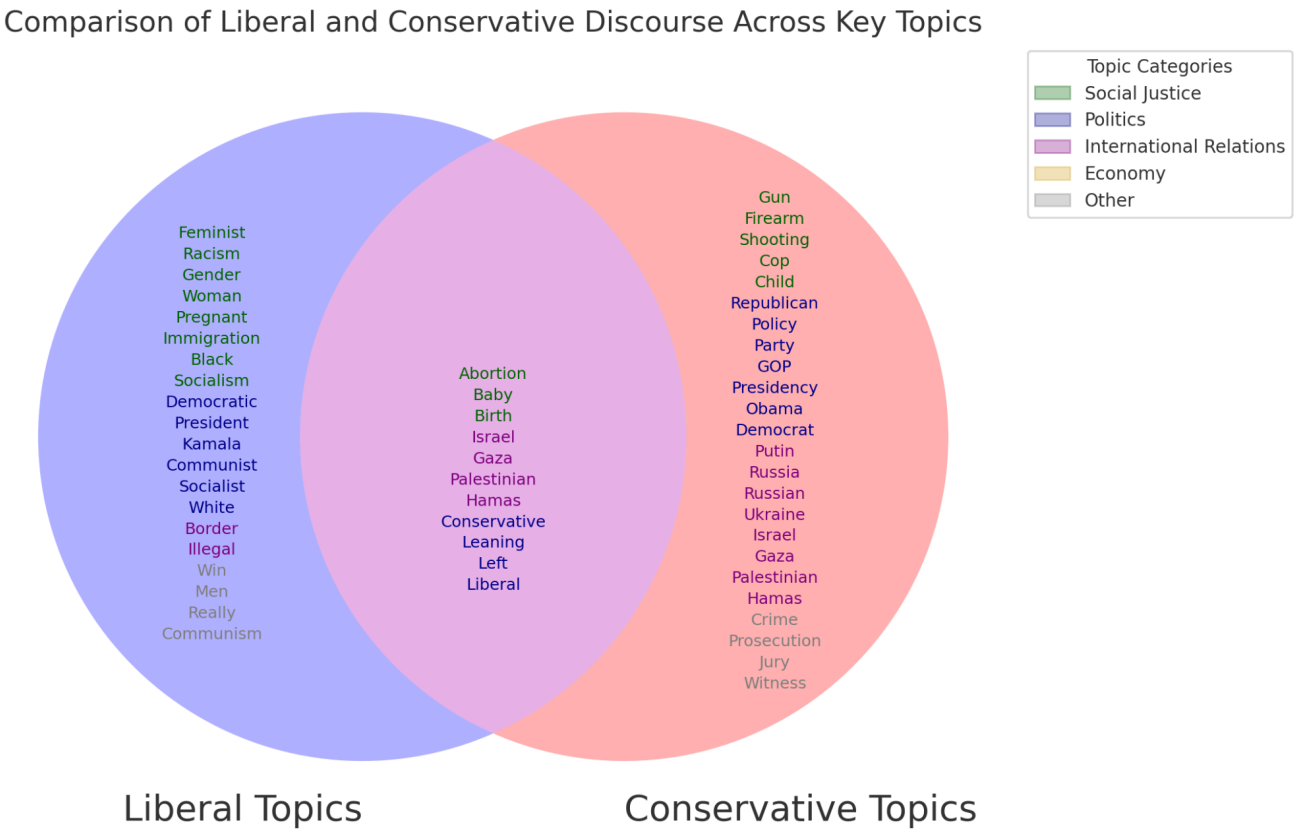


Figure 10: Comparison of Liberal and Conservative Discourse Topics

The high level comparison between the topics reveals some differences. For instance, abortion and reproductive rights were strong topics in both groups. However, it appears that liberal leaning content is more focused on women's rights, feminist and social justice in general. In the conservative content, on the other hand, we do not see other topics related to women's

rights. This is indicative of the ideological differences when the same topic is discussed between the different groups. Moreover, while global issues ranked second on liberal news articles, they were less prominent in the conservative news articles. A deeper look at the community written content shows that the Reddit groups appeared more polarized than the news articles, as political party-related topics dominated the top of their topic lists. Table 5 in the appendix provides more details on the topics discussed by each of the groups in the different platforms.

IV. Methods

4.1 Methodologies to Identify and Predict Political Bias

After establishing the political thematics and corpus foundation, this paper investigates the accuracy of NLP-based ML models in detecting and predicting political bias in news articles and social media discussions. We used three different ML models: logistic regression, SVM and Multinomial Naive Bayes. Prior work has shown that these models can effectively predict political context (Iyyer et al. 2014; Das et al. 2023; Antypas et al. 2023).

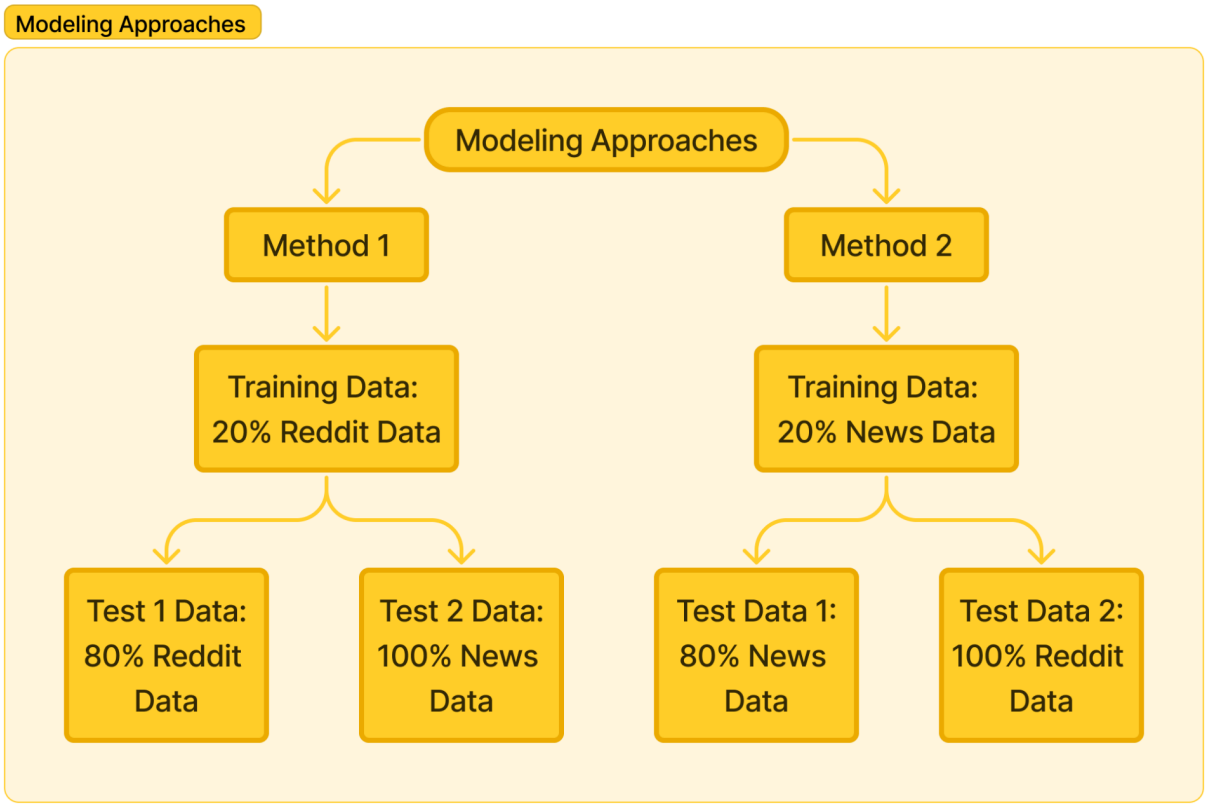


Figure 11: Two modeling approaches used for analysis of political subreddit comments and political articles.

Our methodology involves two approaches that can be seen in Figure 11: first, develop a model to classify political messaging from Reddit to evaluate news articles, and second, develop a model trained on news articles to assess social media discussions in Reddit. The model's effectiveness will be scored using a train-test split of the training data and grading the models with an F Score, a measure of the predictability of the model. Then, each model was graded on how accurately it could predict a testing group, such as predicting news articles based on their political leaning of the publisher from a Reddit-trained model.

Approach 1:

Develop a ML model to classify political messaging into left and right-leaning categories based on content extracted from social media communities in Reddit. This model will then be used to test and evaluate political bias in media articles from major news outlets.

Approach 2:

Develop a MLmodel to classify political messaging into left and right-leaning categories based on content extracted from media articles from major news outlets. The model will then be used to test and evaluate social media communities such as Reddit.

Approach1: Reddit Training - Reddit and News Testing

After preprocessing the Reddit documents and splitting them into their appropriate liberal left and conservative right leaning corpora, they were then split into training and test groups, using a randomizer, setting 20 percent of the vectorized documents as the training group and leaving 80 percent of the documents to be tested on. Figure 12 illustrates the flow diagram of this process.

The training dataset was analyzed using three ML models to determine whether a test comment could be accurately classified as originating from a liberal or conservative subreddit. The results of this analysis, based on the train-test split, are presented in Figure 14 in the results section and detailed further in Table 1 in Appendix A.After training and testing with the Reddit data, the models were restructured to allow the Reddit data to be the training data and the news articles to be the testing data for each of the three models. The Reddit training models made a prediction on whether the news article being considered was written by a conservative-leaning news source or a liberal-leaning news source. The full outputs of these analyses can be seen in Table 2 within Appendix A.

This had marginal effects on the performance and accuracy of our models, with no overall benefit seen to the models. The ML model that produced the highest average F Score was used as the primary model moving forward.

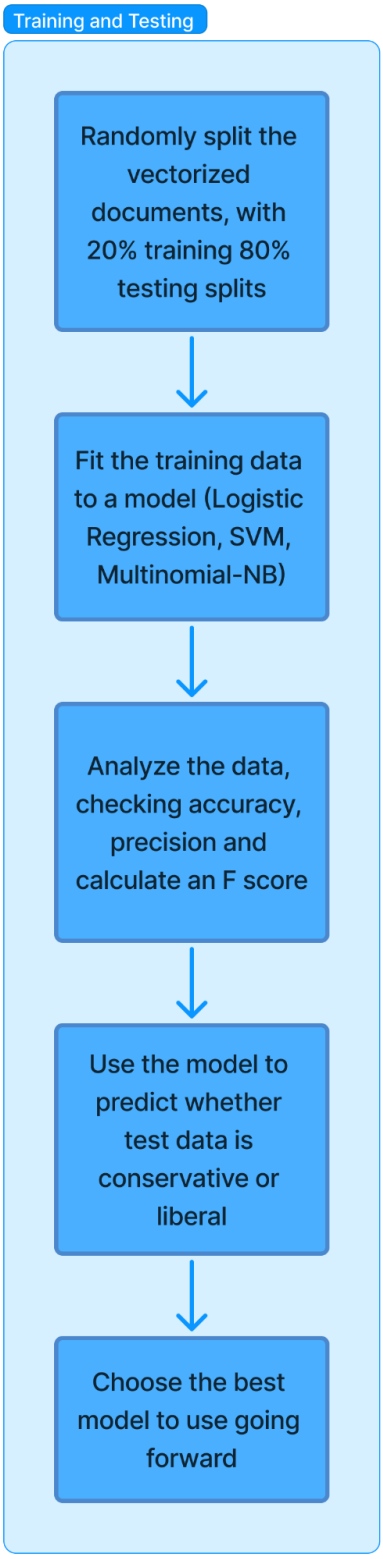


Figure 12: Training and testing process flow diagram.

Approach 2: News Article Training - News Article and Reddit Testing.

We repeated Approach 1 but interchanged the news articles and Reddit data. After preprocessing the news outlet documents and splitting them into their appropriate liberal left and conservative right leaning corpora, they were then split into a training and test group, using a randomizer, setting 20 percent of the vectorized articles as the training group and leaving 80 percent of the articles to test on. This was analyzed using Multinomial NB to determine whether training news documents could be used to predict whether a test news article was from a liberal or conservative-leaning news source based on their AllSides political bias rating. The results can be seen in Table 3 within Appendix A.

After training and testing with the news data, the model was moved to allow the news data to be the training data and the Reddit comments to be the testing data. The news source training models made a prediction on whether a Reddit comment being considered was written by a conservative or liberal subreddit. The outputs of these analyses can be seen in Table 4 within Appendix A.

V. Results

5.1 Sentiment Analysis

The aggregated sentiment analysis revealed that there is a noticeable difference between the four groups we used in this study; conservative and liberal Reddit comments, as well as conservative and liberal news articles. Professional tone and editorial standards were evident in news articles compared to Reddit comments. News articles had an overall more positive sentiment compared to the Reddit comments. The news article sentiment score was almost 9 times higher than the Reddit sentiment score as illustrated in Figure 13

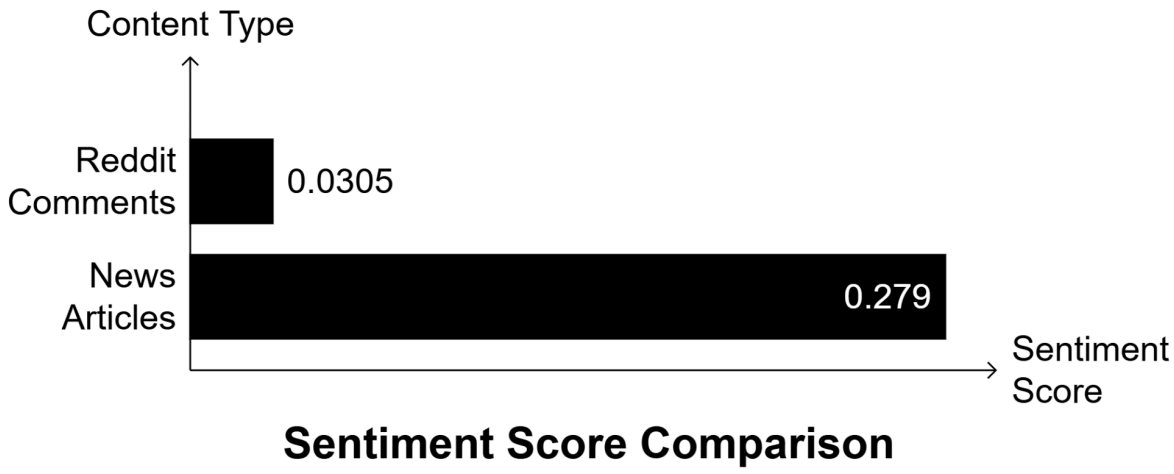


Figure 13: An aggregated sentiment score comparing the average sentiment score in the news articles and the average sentiment score in the Reddit comments.

The sentiment analysis scores (illustrated in Figure 14) segmented between political thematics showed that liberal Reddit communities have a mildly positive sentiment compared to conservative RedditConservative communities What makes this observation interesting is the sentiment is slightly reversed in news outlets. Conservative news article sources showed a slightly higher sentiment score compared to liberal news article sources.

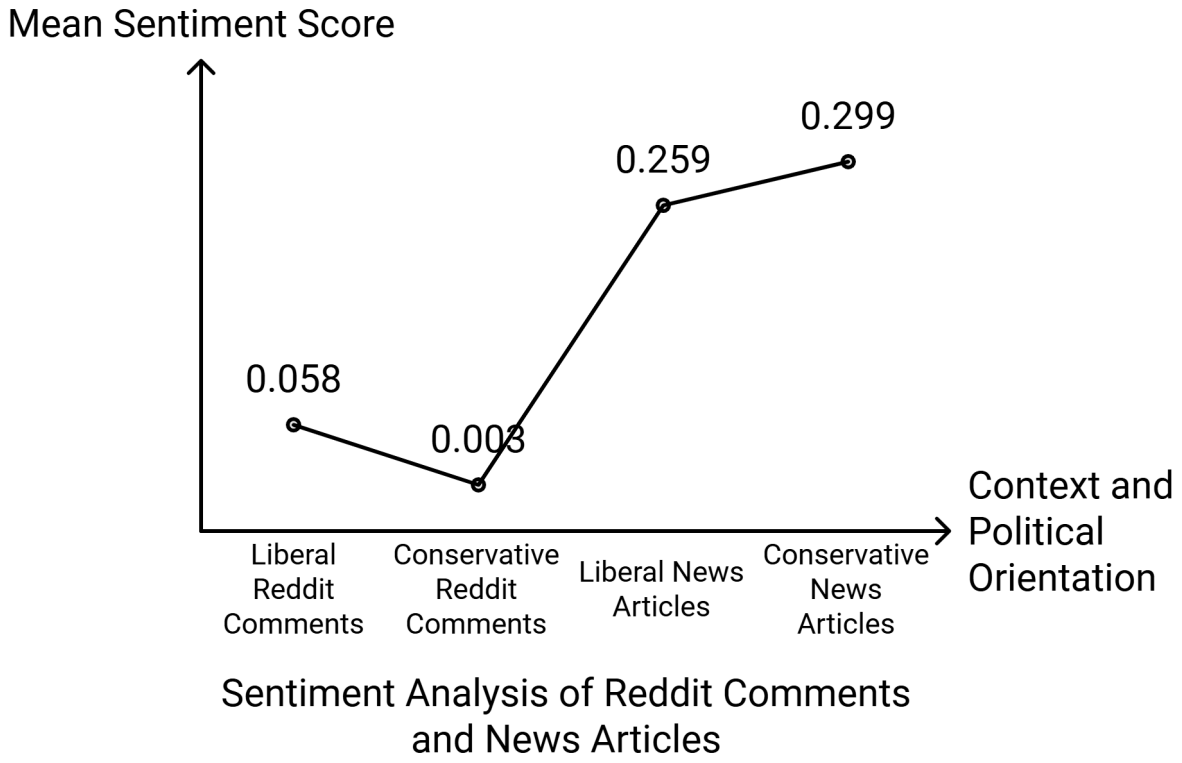


Figure 14: Average sentiment score of Reddit comments and news articles split by their political leaning.

5.2 Model Results

The viability of the model was obtained by using a test-train split on the Reddit data to determine whether we could train the model with a group of documents containing a randomized set constituting 20 percent of the Reddit documents and testing on the remaining 80 percent. The resulting F Score was used to determine whether the model was viable as well as which model to use going forward. The model with the highest average F Score was the Multinomial NB model, with all three models being shown below in Figure 15. This same procedure was used on news articles to prove viability of the news article model to test on Reddit comments. These results can be seen in Figure 16.

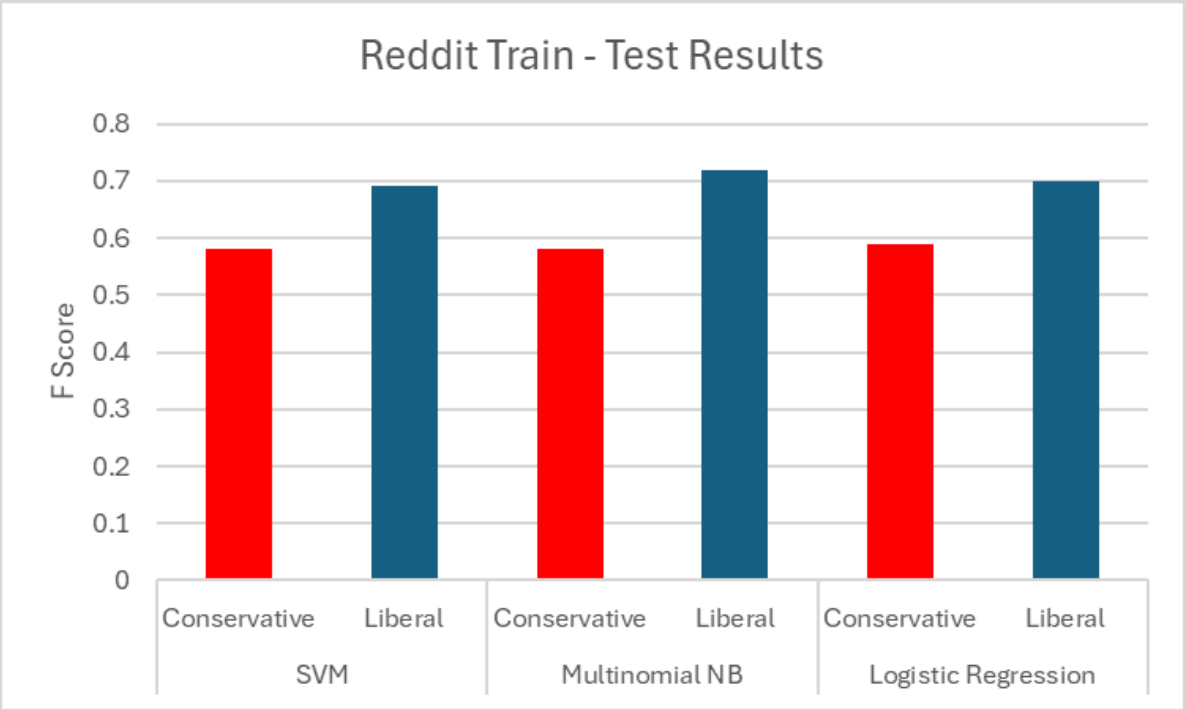


Figure 15: Graphical representation of the Reddit Test-Train split results, proving viability of the model with Multinomial NB analysis resulting in the highest average F Score between the two political leanings.

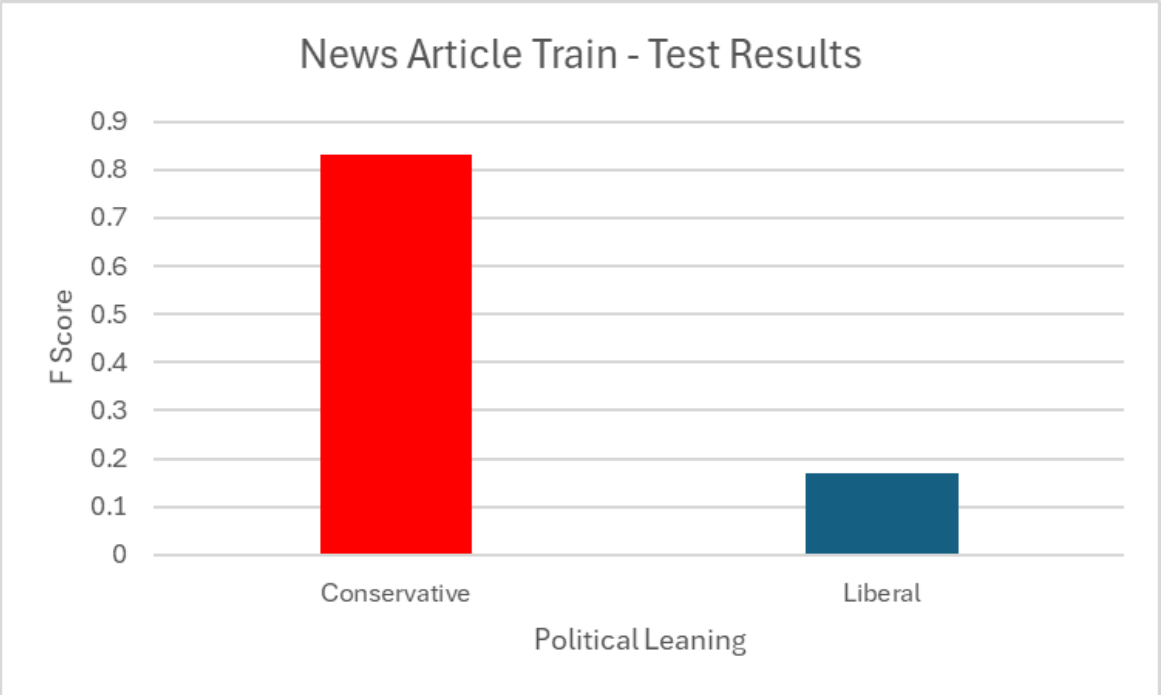


Figure 16: Graphical representation of news article test-train split results based on political leaning.

After showing model viability in both directions, the Reddit training data was used to determine the political leaning of news articles used as test data. The Reddit training data was able to predict liberal news documents at a 95.7 percent rate and conservative documents at a 4.8 percent rate, with overall model accuracy of all documents of 39.1 percent. The reverse was also tested, setting the news articles as the training data and the Reddit documents as the testing data. The news article training data was able to predict liberal Reddit comments at a 0.3 percent rate and conservative Reddit comments at a 99.7 percent rate. The results of both testing models can be seen below in Figure 17.

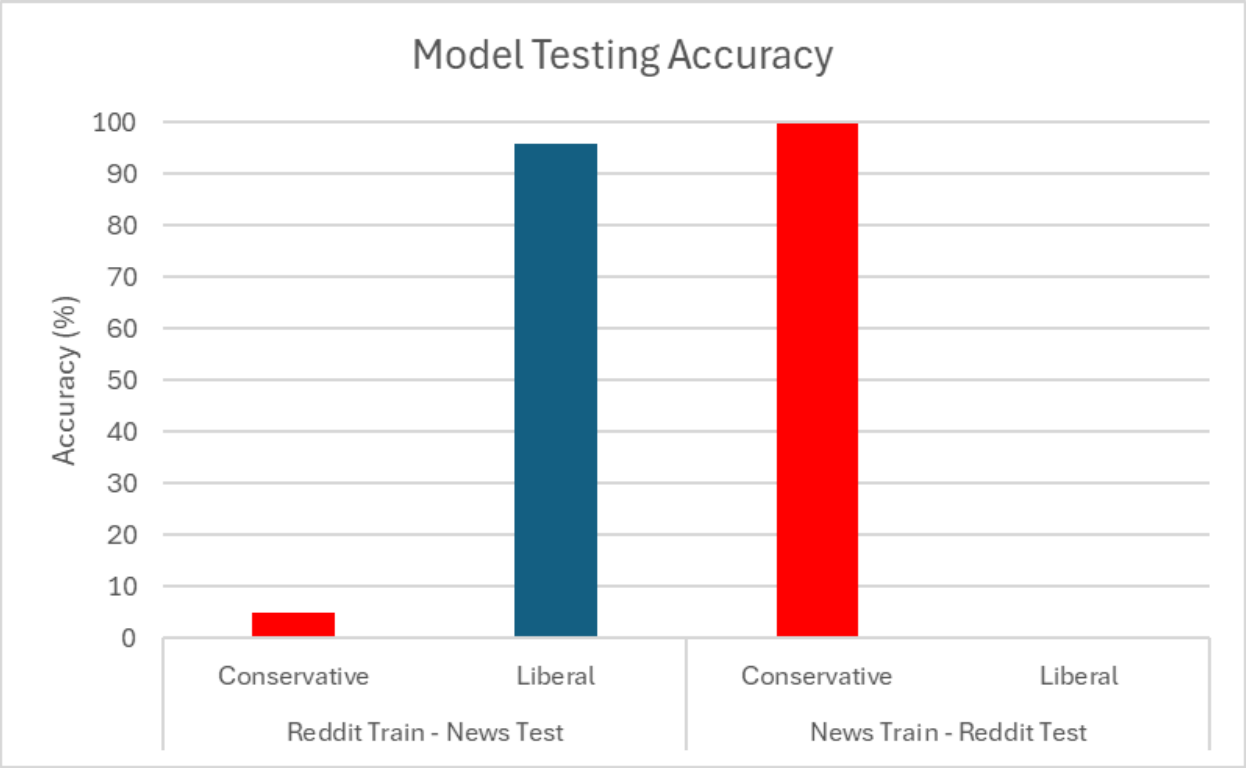


Figure 17: Overall model testing accuracy, with Reddit training leading to a more accurate prediction of liberal news article political leaning and news article training leading to a more accurate prediction of conservative Reddit comments.

VI. Analysis and Interpretation

The sentiment analysis results reveal an interesting dynamic between professionally written news articles and Reddit comments written by ambiguous community members. New articles had nearly ninefold higher sentiment scores compared to Reddit content. These results show how news articles can use positively polished language to engage and inform its audience. Reddit communities, on the other hand, use more neutral and unstructured language to express their opinion.

To put this in perspective, our corpora was collected in the year prior to a polarizing US presidential election. The conservative media prioritized a tone of optimism or confidence in their messaging while liberal media was more on the cautious side. A possible interpretation is that the conservative media goal was focused on inspiring confidence and positive messages to its audience. On the other hand, the lower scores for liberal media suggest a more critiquing or cautious tone. This aligns with criticisms that the Democratic Party received after the 2024 presidential elections regarding the party's emphasis on attacking the Republican Party as opposed to presenting affirmative messages.

In the absence of editorial considerations, the sentiment in Reddit was different among the two political groups. Liberal communities showed a slightly more positive tone than conservative communities. Our topic discovery revealed that conservative communities on Reddit often discuss conflict related issues such as international conflict and divisive issues such as abortion laws. Liberal communities, on the other hand, evolve around social justice topics and positive elections outcomes. This shows how media and public communities can react differently– especially in politically charged times.

From the model viability testing, the Reddit train-test model showed that liberal comments could be more accurately predicted than conservative comments. The news train-test model showed the opposite phenomenon, with conservative news articles being easier to predict than liberal news articles. These phenomena were reinforced by the results of testing the models on the other corpus, with the Reddit model showing much higher accuracy in labeling liberal articles and the news model showing much higher accuracy in labeling conservative Reddit comments.

After further examination of the classification reports from the Reddit model, conservative subreddit comment predictions showed higher precision, measuring the true positive predictions among all positive predictions. However, the opposite effect was seen in the recall, where the number of true positives were compared to all positive instances even if an instance was a false negative. The Reddit model showed a higher recall for liberal subreddit comments. The low recall score of the conservative subreddit comment led to an overall lower F Score which was mirrored in the actual testing results, as the Reddit model proved to be better at predicting liberal news articles as well.

Examination of the classification reports from the news model indicated the precision of the liberal news articles had a higher precision than that of the conservative news articles, but the vastly lower recall score of the liberal articles made this model nonviable for using the model to predict the political leaning of liberal news. This was mirrored in the test data for the news model, where the model struggled to predict comments that were from liberal subreddits while it performed well on comments from conservative subreddits.

VII. Conclusions

This paper applied different NLP and ML techniques to explore linguistic patterns driven from sentiment analysis and discussed topics and political leanings. The corpora was selected from a specifically politically charged point of time– about a year before the 2024 US presidential elections. We found that the platform used to express one's opinion can make an impact on their content's tone. Professionally written content, such as news articles, tends to have a more positive sentiment when compared to Reddit comments. The conversation nature between the group members can explain the more neutral sentiment in the Reddit community comments.

Ideological differences were also evident in the topics discussed by both parties. For instance, abortion was discussed by both groups. However, liberal content used words such as "consent", "aware", "compromise" indicating a more focus on social justice and individual rights. "Baby", "child", and "birth" were often used around abortion in conservative discourse which highlights the morals and ethical implications of the topic.

Two predictive models were built using two separate corpora, training on one corpus and testing on the other. The first corpus was created using Reddit comments organized by political leaning, and the second corpus was created using news articles organized by political leaning of the publisher. Both models performed well at predicting one side of the political spectrum and poorly on the other. The Reddit-trained model performed well on predicting liberal news articles and performed poorly when predicting conservative news articles. The news-trained model performed well on predicting conservative Reddit comments and performed poorly when predicting liberal Reddit comments.

VIII. Directions for Future Work

While the results demonstrate the potential for automated classification of political leanings, there are several opportunities for future work to build on and improve the findings of this research.

8.1 Time Series Analysis of Sentiment Scoring and Topic Discovery

The political landscape is dynamic, with narratives, sentiments, and topics evolving over time constantly. Media reporting and public discourse through election years, news cycles and even foreign events such global summits and world conflicts that have a direct relationship to domestic politics could alter sentiment scoring and topics as it relates to researching political classifications methodologies using NLP and ML techniques. By analyzing text data over time, researchers could uncover patterns that reveal how ideological narratives evolve and adapt. Context-aware models, which consider surrounding text and historical data, can provide richer insights into these shifts.

Temporal analysis can also enhance the ability to detect emerging political trends and shifts in public sentiment. For example, spikes in specific topics, such as immigration or healthcare through time can provide insights into how public discourse is shaped and reveal long-term trends of media framing.

8.2 Exploration of Cross-Platform Language Processing and Modeling

Given the findings in this research exploring cross-platform language processing between Reddit - a social media platform and various news sources outlets, the intricacies of adapting to the linguistic variations across multiple platforms would be an interesting exploration. Different platforms have distinct communication styles influenced by their audience, purpose, and

functionality outside of one's own specific intention or focus - in this case political ideology. These differences present challenges for NLP models which must be capable of adapting to diverse language patterns while maintaining consistency and accuracy in sentiment analysis and scoring.

One of the key challenges in cross-platform modeling is addressing variations in syntax, vocabulary, and semantics. Words or phrases may hold different connotations on different platforms. For example, sarcasm and slang may not translate effectively to formal tones on alternative platforms which was in this case particularly identified in this research in regards to the analysis on tone. This seems to have some impact on the prediction capability between Reddit and news sources. The exploration of cross-platform modeling has practical applications in identifying and mitigating political bias and misinformation.

References

Alfonso, Michael, and Dionisia Bhisetya Rarasati. 2023. “Sentiment Analysis of 2024 Presidential Candidates Election Using SVM Algorithm.” *JISA(Jurnal Informatika Dan Sains)* 6 (2): 110–15. https://doi.org/10.31326/jisa.v6i2.1714.

AllSides Media Bias Ratings™. AllSides Technologies, Inc. https://www.allsides.com/media-bias/media-bias-ratings. Retrieved November 2024.

Antypas, Dimosthenis, Alun Preece, and Jose Camacho-Collados. 2023. "Negativity Spreads Faster: A Large-Scale Multilingual Twitter Analysis on the Role of Sentiment in Political Communication." *Online Social Networks and Media* 33: 100242. https://doi.org/10.1016/j.osnem.2023.100242.

Brenan, Megan. 2024. “Americans’ Trust in Media Remains at Trend Low.” *Gallup*, October 14, 2024. https://news.gallup.com/poll/651977/americans-trust-media-remains-trend-low.aspx.

Das, Anubrata, Houjiang Liu, Venelin Kovatchev, and Matthew Lease. 2023. “The State of Human-Centered NLP Technology for Fact-Checking.” *Information Processing & Management* 60 (2): 103219. https://doi.org/10.1016/j.ipm.2022.103219.

De Arruda, Gabriel Domingos, Norton Trevisan Roman, and Ana Maria Monteiro. "Analysing bias in political news." J. Univers. Comput. Sci. 26, no. 2 (2020): 173-199.

Ferrer, Xavier, Tom van Nuenen, Jose M. Such, and Natalia Criado. 2020. "Discovering and Categorising Language Biases in Reddit." In *International AAAI Conference on Web and Social Media (ICWSM)*.

Hutto, C.J. and Eric Gilbert. 2014. “VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text.” *Proceedings of the 8th International Conference on Weblogs and Social Media.* (2014): 215-225.

Iyyer, Mohit, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. “Political Ideology Detection Using Recursive Neural Networks.” Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2014. https://doi.org/10.3115/v1/p14-1105.

Lewandowsky, Stephan, Ullrich K.H. Ecker, John Cook, Sander van der Linden, Jon Roozenbeek, and Naomi Oreskes. 2023. “Misinformation and the epistemic integrity of

democracy.” *Current Opinion in Psychology.* Vol. 54 (2023): 101711.

Mayopu, Richard G., Yi-Yun Wang, and Long-Sheng Chen. 2023. "Analyzing Online Fake News Using Latent Semantic Analysis: Case of USA Election Campaign." *Big Data Cogn. Comput.* 7: 81. https://doi.org/10.3390/bdcc7020081.

van der Linden, Sander, Costas Panagopoulos, and Jon Roozenbeek. 2020. “You are fake news: political bias in perceptions of fake news.” SAGE Publications. *Media, Culture & Society.* Vol. 42, no. 3 (2020): 460-470.

Wu, Yichao, Zhengyu Jin, Chenxi Shi, Penghao Liang, and Tong Zhan. 2024. “Research on the Application of Deep Learning-based BERT Model in Sentiment Analysis.” *Applied and Computational Engineering.* Vol. 71 no. 1: 14-20

Zahrah, Fatima, Jason RC Nurse, and Michael Goldsmith. "A comparison of online hate on reddit and 4chan: a case study of the 2020 US Election." In Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing, pp. 1797-1800. 2022.

Appendix A

Reddit:

```
['amazing many people think intelligent savvy dont realize obvious policies support impact personally',
 'people truly intelligent savvy learn experiences others ignorance excusable topics havent studied ana given great deal thought situation yet took personally impacted wake',
 'theres old saying liberal conservative hasnt mugged yet anas case mugging took form sexual assault wish could say sort naivety confined political left theres awful lot bootlicking side aisle comes police brutality misconduct corresponding sayi
 'oh delicious irony ana k became rich famous thing conservatives',
 'everybody anti gun need one',
 'story old time',
 'conservative leftist shamed friends molested homeless man erection',
 'thus always thus shall ever',
 'conservative completely different view government work',
 'bet brought several guns incident decide get ccw permit',
 'wonder cenk say',
 'homeless guy',
 'thats great nobody else base political standing think right vs going along particular side think theyre good people',
 'love communist utopia theyre directly affected thats realize liberal policies work real world',
 'say right cult',
 'fact checking',
 'didnt throw fact check would add bs remark castigating trump way switching kamala',
 'either moderators fact checked trump injected statement opinion trump',
 'unprofessional obviously biased',
 'hope people see bs',
 'oh know wont purpose fact rm cnn kinds negative things say trump glorified harris left seething joy right',
 'ya joke',
 'doesnt use factual statements lives realm subjective feelings',
 'find crazy theres second debate fox balance',
 'definitely needs second debate fox moderators line well known republican haters proved time time',
...
 'ding ding ding',
 'suppose talked bobby jindal dont think helpful political career',
 'forced melodrama something id expect hillary',
 'recalibrate katie need go full mtg watch margaret thatcher learn',
 ...]
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

Stemming Output:

```
['amaz mani peopl think intellig savvi dont realiz obviou polici support impact person',
 'peopl truli intellig savvi learn experi other ignor excus topic havent studi ana given great deal thought situat yet took person impact wake',
 'there old say liber conserv hasnt mug yet ana case mug took form sexual assault wish could say sort naiveti confin polit left there aw lot bootlick side aisl come polic brutal misconduct correspond say back blue til happen',
 'oh delici ironi ana k becam rich famou thing conserv',
 'everybodi anti gun need one',
 'stori old time',
 'conserv leftist shame friend molest homeless man erect',
 'thu alway thu shall ever',
 'conserv complet differ view govern work',
 'bet brought sever gun incid decid get ccw permit',
 'wonder cenk say',
 'homeless guy',
 'that great nobodi els base polit stand think right vs go along particular side think theyr good peopl',
 'love communist utopia theyr directli affect that realiz liber polici work real world',
 'say right cult',
 'fact check',
 'didnt throw fact check would add bs remark castig trump way switch kamala',
 'either moder fact check trump inject statement opinion trump',
 'unprofession obvious bias',
 'hope peopl see bs',
 'oh know wont purpos fact rm cnn kind neg thing say trump glorifi harri left seeth joy right',
 'ya joke',
 'doesnt use factual statement live realm subject feel',
 'find crazi there second debat fox balanc',
 'definit need second debat fox moder line well known republican hater prove time time',
...
 'ding ding ding',
 'suppos talk bobbi jindal dont think help polit career',
 'forc melodrama someth id expect hillari',
 'recalibr kati need go full mtg watch margaret thatcher learn',
 ...]
```

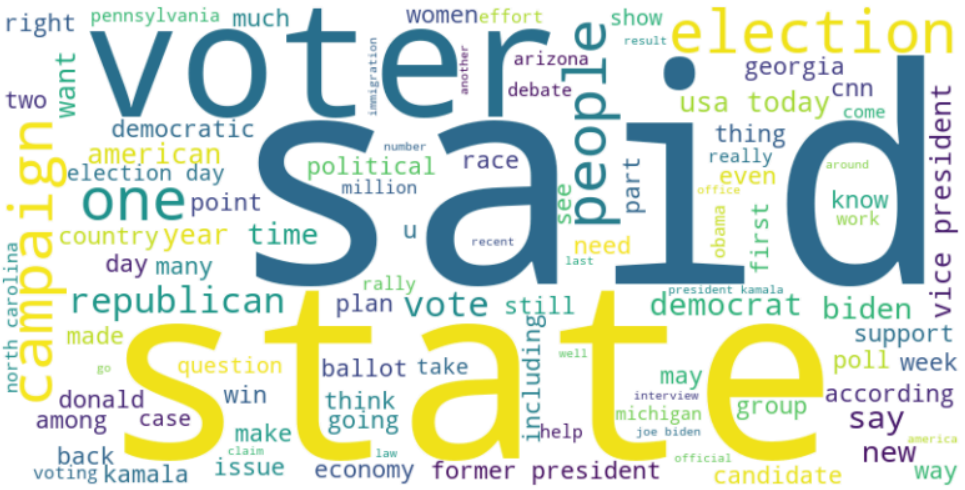
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

Lemmatization:

```
['amazing many people think intelligent savvy dont realize obvious policy support impact personally',
 'people truly intelligent savvy learn experience others ignorance excusable topic havent studied ana given great deal thought situation yet took personally impacted wake',
 'there old saying liberal conservative hasnt mugged yet ana case mugging took form sexual assault wish could say sort naivety confined political left there awful lot bootlicking side aisle come police brutality misconduct corresponding saying
 'oh delicious irony ana k became rich famous thing conservative',
 'everybody anti gun need one',
 'story old time',
 'conservative leftist shamed friend molested homeless man erection',
 'thus always thus shall ever',
 'conservative completely different view government work',
 'bet brought several gun incident decide get ccw permit',
 'wonder cenk say',
 'homeless guy',
 'thats great nobody else base political standing think right v going along particular side think theyre good people',
 'love communist utopia theyre directly affected thats realize liberal policy work real world',
 'say right cult',
 'fact checking',
 'didnt throw fact check would add b remark castigating trump way switching kamala',
 'either moderator fact checked trump injected statement opinion trump',
 'unprofessional obviously biased',
 'hope people see b',
 'oh know wont purpose fact rm cnn kind negative thing say trump glorified harris left seething joy right',
 'ya joke',
 'doesnt use factual statement life realm subjective feeling',
 'find crazy there second debate fox balance',
 'definitely need second debate fox moderator line well known republican hater proved time time',
...
 'ding ding ding',
 'suppose talked bobby jindal dont think helpful political career',
 'forced melodrama something id expect hillary',
 'recalibrate katie need go full mtg watch margaret thatcher learn',
 ...]
```

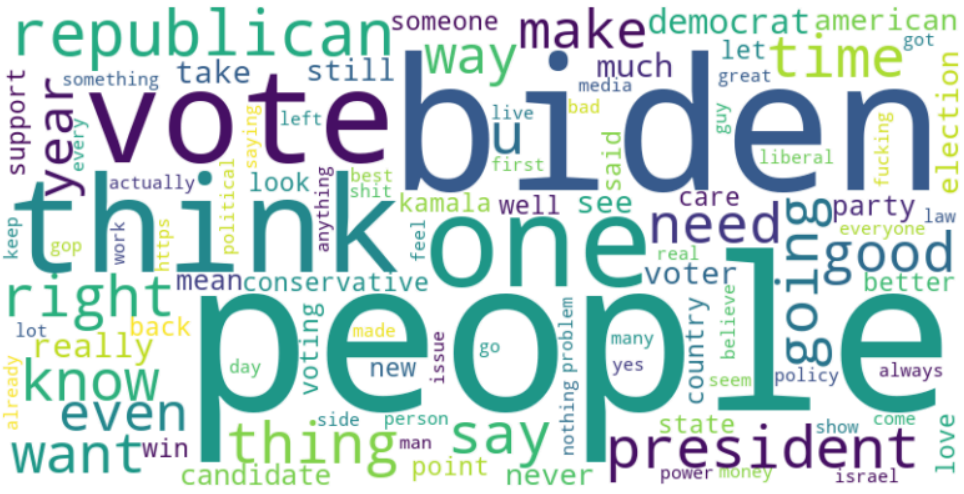
Conservative News Corpus Word Cloud:

Liberal News Corpus Word Cloud:



Conservative Reddit Corpus Word Cloud:

Liberal Reddit Corpus Word Cloud:



A breakdown of the topics by political leaning and outlet:

Reddit Train - Test Data						
Analysis Type	**Logistic Regression**		**Multinomial NB**		**SVM**	
	Conservative	**Liberal**	**Conservative**	**Liberal**	**Conservative**	**Liberal**
Precision	0.68	0.64	0.72	0.63	0.67	0.63
Recall	0.53	0.77	0.48	0.83	0.52	0.76
F Score	0.59	0.70	0.58	0.72	0.58	0.69

Table 1: Reddit data is split into 20% training data and 80% testing data and fitted to three different statistical models to predict whether the test comment was from a conservative or liberal subreddit based on the training data.

Reddit Training - News Article Testing Data						
Analysis Type	**Logistic Regression**		**Multinomial NB**		**SVM**	
	Conservative	**Liberal**	**Conservative**	**Liberal**	**Conservative**	**Liberal**
Accuracy	0.052	0.957	0.048	0.957	0.048	0.964
Overall	0.394		0.391		0.394	

Table 2: Reddit training data using three different statistical models is used to predict whether a given news article was written by a conservative or liberal news source based on AllSides political bias ratings.

News Train - Test Data		
Analysis Type	**Multinomial NB**	
	Conservative	**Liberal**
Precision	0.74	1.00
Recall	1.00	0.10
F Score	0.83	0.17

Table 3: News data is split into 20% training data and 80% testing data and fitted to a Multinomial NB model to predict whether the test article was from a conservative or liberal-leaning news publication based on the training data.

News Article Training - Reddit Testing Data		
Analysis Type	**Multinomial NB**	
	Conservative	**Liberal**
Accuracy	0.996	0.004
Overall	0.481	

Table 4: News outlet training data using a Multinomial NB statistical model is used to predict whether a given reddit comment was scraped from a conservative or liberal subreddit.

Subreddit Left-Leaning Communities:

1. /r/democrats
2. /r/esist
3. /r/wayofthebern
4. /r/liberal
5. /r/askaliberal
6. /r/joebiden

7. /r/progressive
8. /r/dsa
9. /r/murderedbyaoc

Subreddit Right-Leaning Communities:

1. /r/capitalism
2. /r/asktrumpsupporters
3. /r/askconservatives
4. /r/republican
5. /r/progun
6. /r/prolife
7. /r/conservatives
8. /r/walkaway
9. /r/conservative

Group	Reddit Topics	News Topics
Conservative	1. Republican, Democrat, GOP, Party	1. Republican, Democrat, GOP, Party
	2. Abortion, Baby, Child, Birth	2. Gun, Firearm, Shooting, Cop
	3. Israel, Gaza, Palestinian, Hamas	3. Israel, Gaza, Palestinian, Hamas
	4. Putin, Russia, Russian, Ukraine	4. Conservative, Liberal, Leaning, Left
	5. Obama, Presidency, Policy	5. Crime, Jury, Witness, Prosecution
Liberal	1. Kamala, Win, Democratic, President	1. Kamala, Win, Woman, Really
	2. Socialism, Socialist, Communism, Communist	2. Israel, Gaza, Palestinian, Hamas
	3. Abortion, Birth, Pregnant, Baby	3. Conservative, Liberal, Left, Leaning
	4. Racist, White, Black, Racism	4. Immigrant, Illegal, Border, Immigration
	5. Woman, Feminist, Men, Gender	5. Racist, White, Black, Racism

Table 5: This table compares topics of discussion associated with two ideological groups (Conservative and Liberal) across two sources of discourse: Reddit and News Media Outlets.