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

Since the 2016 US Presidential Election, 'Fake News' is a prevailing term that has led to a lack of trust in mainstream media (van der Linden, Panagopoulos, and Roozenbeek 2020). In 2024, a Gallup poll showed that 69 percent of Americans have little to no trust in the media (Brenan 2024). Political bias in the mainstream media appears to shoulder much of the blame, leading readers and viewers to assume the media is spinning stories to benefit viewership and political parties (van der Linden, Panagopoulos, and Roozenbeek, 2020).

Democracy requires a consistent and accurate flow of information throughout citizens, and misinformation is often seen as the first step in toppling a democratic governmental system (Lewandowsky et. al. 2023). With media trust at an all-time low, citizens need a way to measure whether a given news source shows bias to their consumers.

The purpose of this study is to determine whether political bias can be 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 scraped and 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 separate 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 political bias in the other.

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

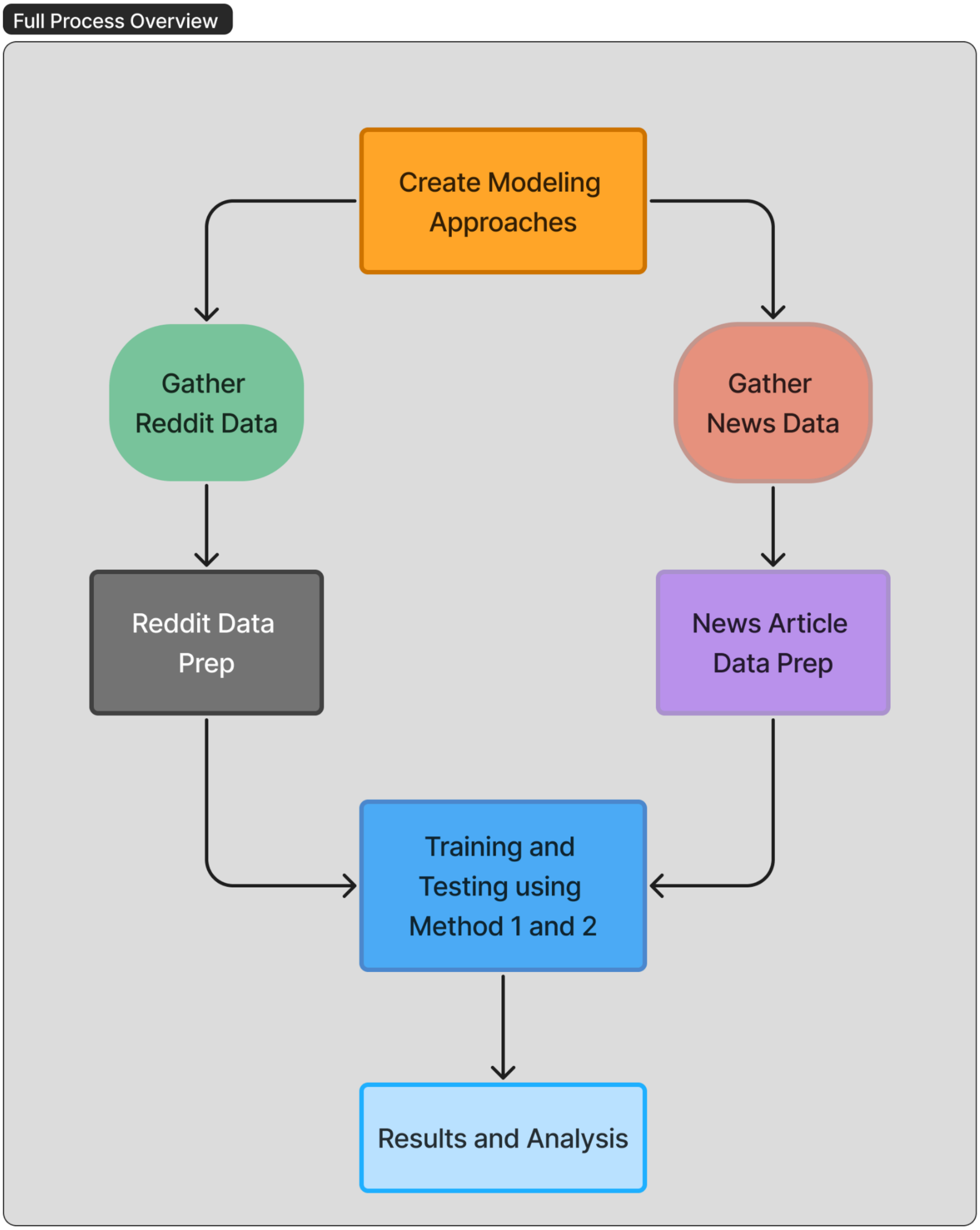


Figure 1: Graphical Abstract. Process flow diagram for analyzing political news bias using AllSides political bias ratings and political subreddits. Colors of the elements in the process flow diagram are consistent and each step is broken down in other diagrams throughout the paper.

Table of Contents

I. Introduction & Problem Statement

Over the last few decades the transition to a digital configuration has amplified online news channels and social media's role in disseminating misinformation and political bias. The polarization of media outlets across the political spectrum has caused concern regarding the radical opinions of democratic identified and republican identified 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 TF-IDF vectorization, tokenization, and sentiment analysis and machine learning (ML) techniques logistic regression, multinomial Naive Bayes (Multinomial Bayes), and support vector machines (SVM) political bias can be identified of across media forms.

The aim of this paper is to examine machine learning model classification performance on linguistic quantifiers and language identification 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 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 conservative or liberal positions across news articles and Reddit comments providing insight on the effect of language on ideological classification.

The model, including two methods of approaching online news and subreddit posts classification, were prepared to handle large scale textual data and distinguish language consistencies through wording and tone. With this information the model ascertains 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 politics 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 and statement bias and classification of real and fake news respectively. Reddit was used by Ferrer et al. (2020) and Zahrah, Nurse, and Goldsmith (2022). 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 use 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. 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 (formally known as Twitter) posts to assess how they aligned with 2024 Indonesian presidential surveys. They used TF-IDF for feature extraction and a 10-folds 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 are a time where a significant amount of polarization, misinformation and fake news can take place. Not only do this kind of news spread fast, they also have negative impacts on the 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) to build 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 and 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

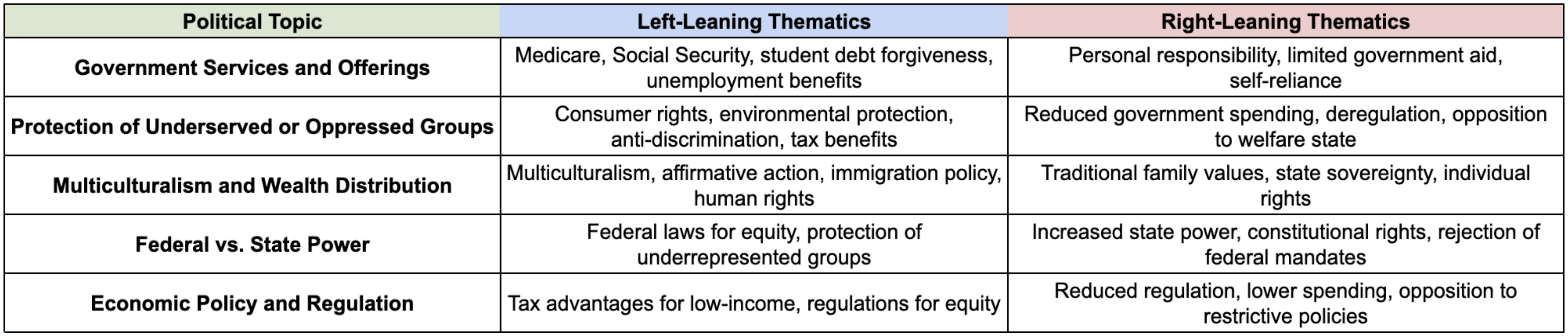
For this analysis, data from both Reddit and news sources were collected over a 12-month time horizon. This time range was chosen to capture recent and relevant discussions that reflect current political climates and trends. By focusing on the latest year, we ensure that our analysis remains timely and 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, a widely used community-based platform, 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 Processing (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, allowing for extraction and pattern identification in political discussions.

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 these 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 are more likely to retain some of the following key thematics:

1. Thematics that reflect positive reviews on generous government services or offerings. This can include positive outlooks on medicare, social security, student debt forgiveness and unemployment benefits.
2. Thematics that reflect positive reviews on federal laws or economic policies that protect groups or individuals that are considered underserved, unprotected or oppressed. This can include positive outlooks on federal laws that protect consumers, the environment and abortion rights, federal laws that fight against discrimination and inequitable outcomes, or tax advantages to those who are on the lower end of the socioeconomic ladder.
3. Thematics that reflect positive reviews that embrace the importance of multiculturalism and the belief that wealth should not be concentrated amongst the few. This can include, positive outlooks on affirmative action, the government's role in immigration policy, and the human rights to healthcare, housing, clean water and a living wage.

Sources with a right or conservative media bias are more likely to retain some of the following key thematics:

1. Thematics that reflect positive reviews on 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, a decrease in federal regulation and an increase in state legislation overriding federal regulation, preserving the fundamental rights and ideals in the constitution, and rejecting equality or equity as an organizing principle.

2. Thematics that reflect positive reviews on 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, rejecting laws that put economic burdens on US and foreign businesses or enterprises, and the rejection of politics related to the “welfare” state such as gender identity and inequality, and affirmative action.

Using political thematics as a guidestone to label media bias, a 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 our machine learning model, two corpora were created to focus on the extreme leftist thematics and the extreme rightist thematics. This leftist article corpus uses sample documents from US Today, CNN and MSNBC while the rightist article corpus uses sample documents from Fox News and the American Conservative.

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. With a list of selected news articles from US Today, CNN, MSNBC, Fox News, and the American Conservative we perform targeted searches using election-related keywords to gather articles relevant to our study’s focus. 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, they undergo data preparation for analysis.

Following data-preprocessing 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, news articles with less than 50 words were dropped. Overall we had 368 news articles aggregated in this DataFrame.

The final step in data preparation was vectorizing 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.

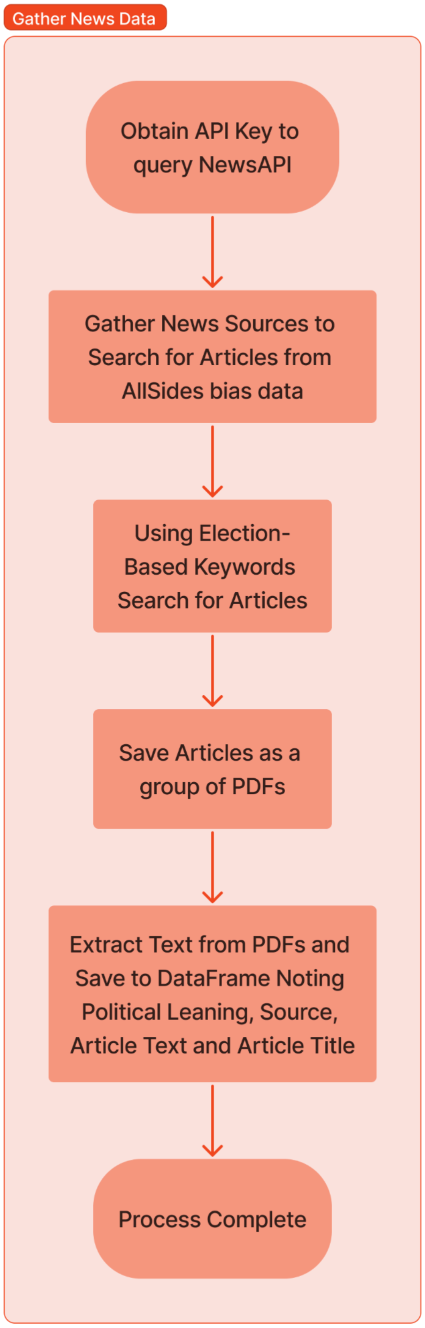

Gather News Data
Obtain API Key to query NewsAPI
Gather News Sources to Search for Articles from AllSides bias data
Using Election-Based Keywords Search for Articles
Save Articles as a group of PDFs
Extract Text from PDFs and Save to DataFrame Noting Political Leaning, Source, Article Text and Article Title
Process Complete

Figure 3: Gather News Data. Process flow diagram displaying news article retrieval using NewsAPI and Python libraries.

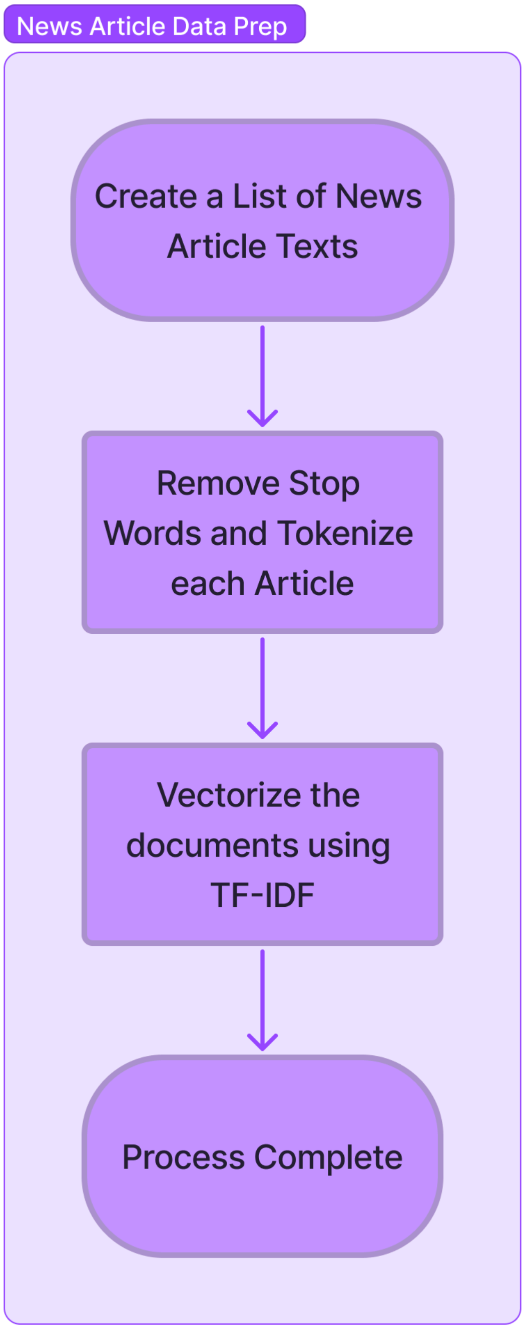

News Article Data Prep
Create a List of News Article Texts
Remove Stop Words and Tokenize each Article
Vectorize the documents using TF-IDF
Process Complete

Figure 4: 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. Reddit is a web platform for social news aggregation, web content rating, and discussion. 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 natural language processing 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 or liberal. 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 data-preprocessing 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 in data preparation was vectorizing 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.

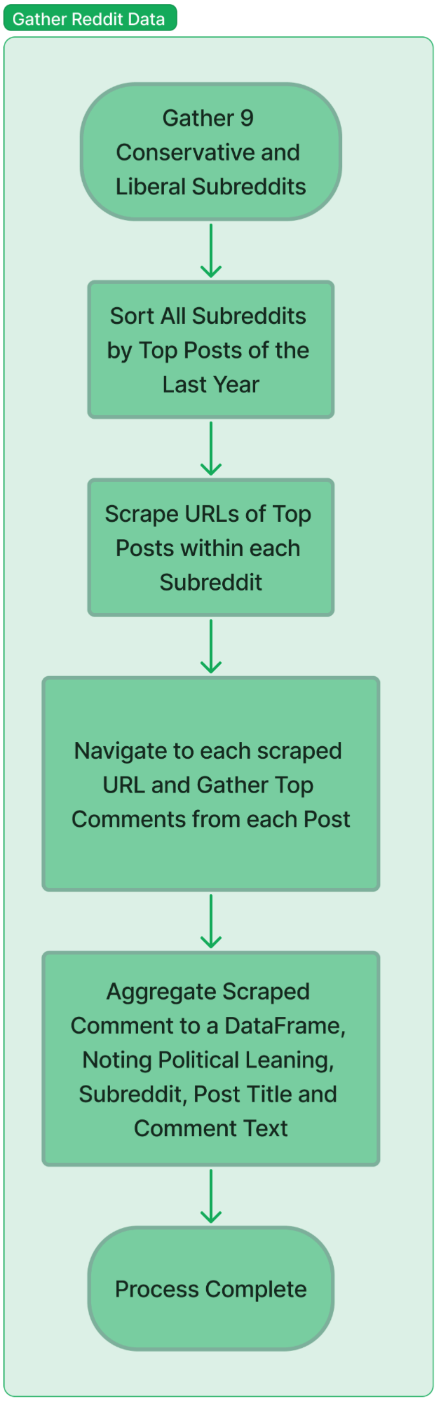

Gather Reddit Data
Gather 9 Conservative and Liberal Subreddits
Sort All Subreddits by Top Posts of the Last Year
Scrape URLs of Top Posts within each Subreddit
Navigate to each scraped URL and Gather Top Comments from each Post
Aggregate Scraped Comment to a DataFrame, Noting Political Leaning, Subreddit, Post Title and Comment Text
Process Complete

Figure 4: Gather Reddit Data. Process flow chart displaying the retrieval of Reddit comments in liberal and conservative subreddit groups using Selenium.

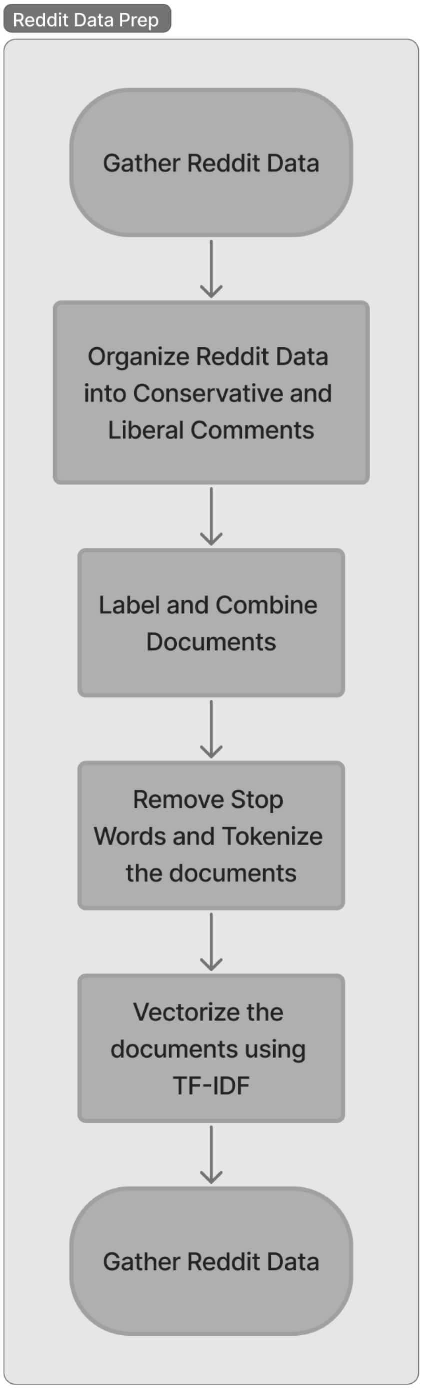

Reddit Data Prep
Gather Reddit Data
Organize Reddit Data into Conservative and Liberal Comments
Label and Combine Documents
Remove Stop Words and Tokenize the documents
Vectorize the documents using TF-IDF
Gather Reddit Data

Figure 5: 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. Second,we take a deep dive to investigate the difference in tone between the different political groups.

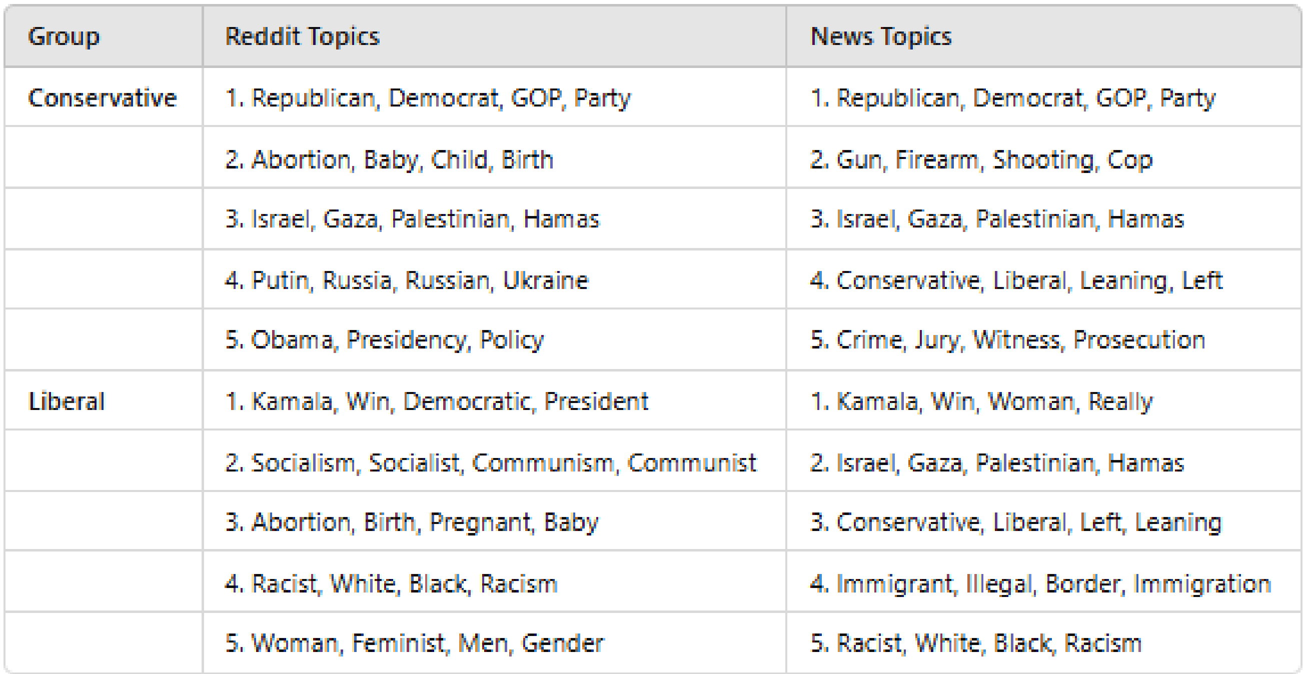
VADER (Valence Aware Dictionary and sEntiment Reasoner) from the NLTK (Natural Language Toolkit) Python package was used to perform this sentiment analysis. VADER is known to be sensitive for web-based media thanks to its ability to analyze contextual information and language nuances (Hutto and Gilbert 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. Below is a summary of the mean sentiment scores for each group:

- Reddit Conservative: 0.003
- Reddit Liberal: 0.058
- News Conservative: 0.299
- News Liberal: 0.259

3.6.2: Discovering Topic

In this section we analyze different topics discussed by the different political leaning groups across Reddit and news content. We used BERTopic for this analysis. BERT leverages embeddings to capture similarities between documents based on their topics (Wu et al. 2024).



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

It is not surprising that the Reddit groups appeared more polarized than the news articles, as political party-related topics dominated the top of their topic lists. The topics were also different between conservative and liberals in general. 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. On the other hand, we do not see other topics related to women’s rights or in the conservative content. 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.

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 Natural Language Process-based Machine Learning models in detecting and predicting political bias in news articles and social media discussions. We used three different ML models; logistic regression, SVM and a multinomial Naive Bayes. Prior work has shown that these models, while simple to interpret, can effectively predict political context (Iyyer et al. 2014; Das et al. 2023; Antypas et al. 2023).

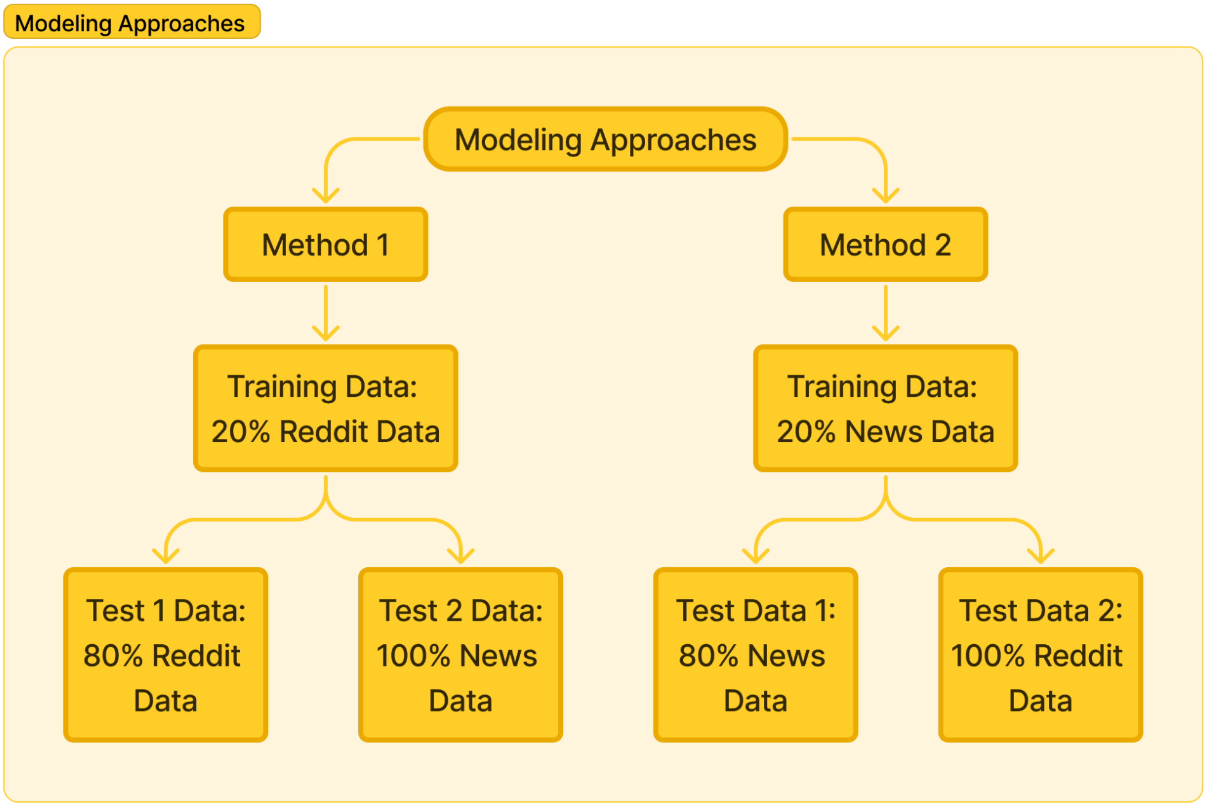

Modeling Approaches
Modeling Approaches
Method 1
Method 2
Training Data: 20% Reddit Data
Training Data: 20% News Data
Test 1 Data: 80% Reddit Data
Test 2 Data: 100% News Data
Test Data 1: 80% News Data
Test Data 2: 100% Reddit Data

Figure 4.1: Two modeling approaches used for analysis of political subreddit comments and political articles.

Our methodology involves two approaches: 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. To quantify the model's effectiveness, we will employ a probability scale to assess the extent of political leaning in various news articles.

Methodology 1:

Develop a machine learning 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.

Methodology 2:

Develop a machine learning model 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.

Method 1: Reddit Training - Reddit and News Testing

The reddit documents were divided into conservative comments and liberal comments, vectorized using TF-IDF. This vectorized set of documents was split into a training and test group using a randomizer, setting 20 percent of the vectorized documents as the training group and leaving 80 percent of the documents to test on. This was analyzed using logistic regression, multinomial Naive Bayes (Multinomial NB), and support vector machines (SVM) models to determine whether the training reddit documents could be used to predict whether a test comment was from a liberal or a conservative subreddit. This test-train split allowed us to see whether this was a viable model to use going forward, and the data can be seen in Table 4.1 below.

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
F1-Score	0.59	0.70	0.58	0.72	0.58	0.69

Table 4.1: 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.

After training and testing with the reddit data, the model was moved 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 outputs of these analyses can be seen in Table 4.2 below.

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 4.2: 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.

We repeated the modeling process after excluding documents with small word counts as discussed in data prep and applied lemmatization. This had marginal effects on the performance and accuracy of our models, with a slight decrease in accuracy of the logistic regression model, a slight increase in the overall accuracy and performance of the support vector machines model, and equivalent performance of the multinomial Naive Bayes model. Since the Multinomial Naive-Bayes model had the highest F1-Score, we decided to go forward with this model for analysis and for use in Method 2.

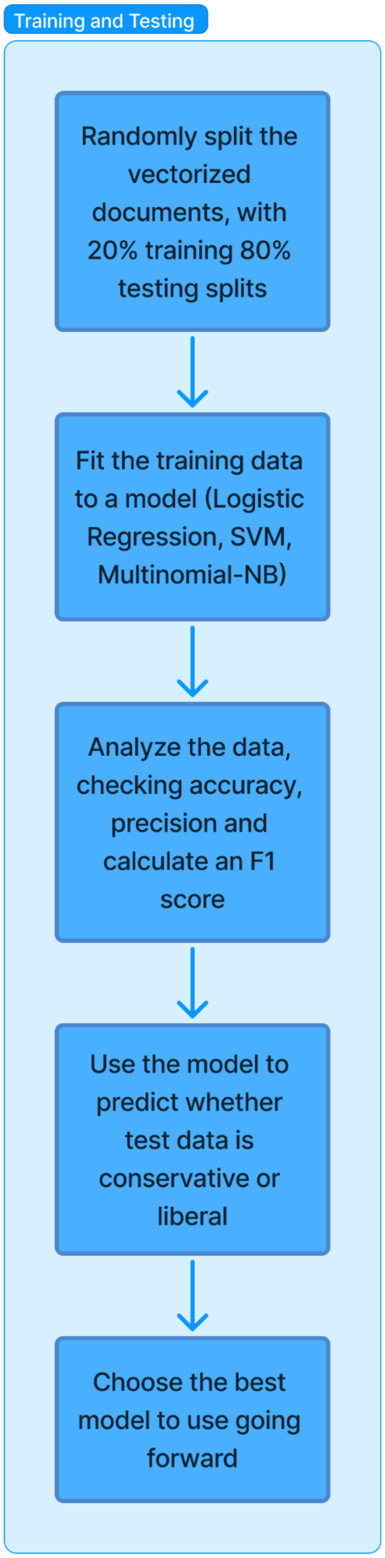

Training and Testing
Randomly split the vectorized documents, with 20% training 80% testing splits
Fit the training data to a model (Logistic Regression, SVM, Multinomial-NB)
Analyze the data, checking accuracy, precision and calculate an F1 score
Use the model to predict whether test data is conservative or liberal
Choose the best model to use going forward

Figure 4.3: Training and testing process flow diagram.

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

We repeated Method 1 but interchanged the news articles and reddit data. The news articles were divided into conservative articles and liberal articles, vectorized using TF-IDF. This vectorized set of articles was 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 the training news documents could be used to predict whether a test news article was from a liberal or a conservative-leaning news source based on their AllSides political bias rating. The results can be seen in Table 4.4 below.

News Train - Test Data		
Analysis Type	**Multinomial Naive-Bayes**	
	Conservative	**Liberal**
Precision	0.71	1.00
Recall	1.00	0.09
F1-Score	0.83	0.16

Table 4.4: News data is split into 20% training data and 80% testing data and fitted to a Multinomial Naive-Bayes model to predict whether the test article was from a conservative or liberal-leaning news publication based on the training data.

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 training models made a prediction on whether the reddit comment being considered was written by a conservative or liberal subreddit. The outputs of these analyses can be seen in Table 4.5 below.

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

Table 4.5: News training data using a Multinomial Naive-Bayes statistical model is used to predict whether a given reddit comment was scraped from a conservative or liberal subreddit.

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.

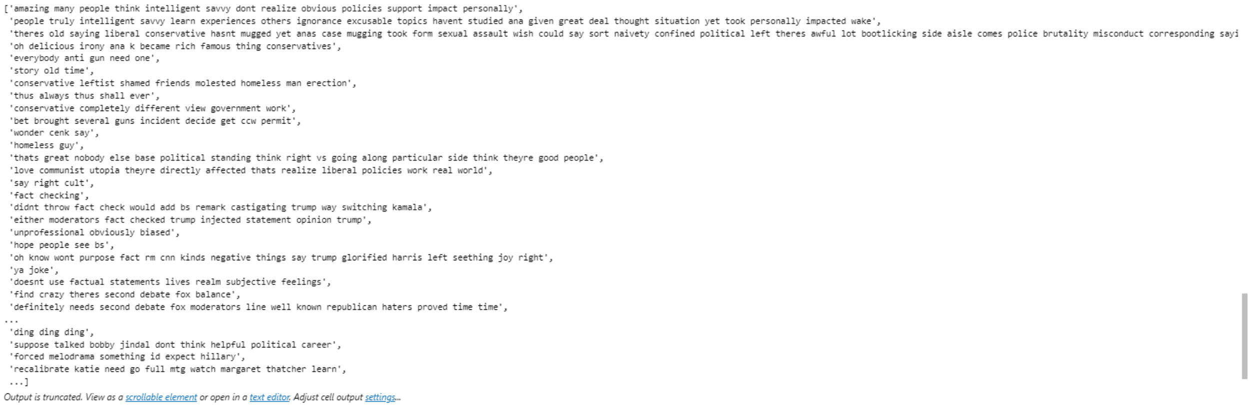
van der Linden, Sander, Costas Panagopoulos, and Jon Roozenbeek. 2020. "You are fake news: political bias in perceptions of fake news." SAGE Publications. *Media, Culture & Society.* Vol. 42, no. 3 (2020): 460-470.

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

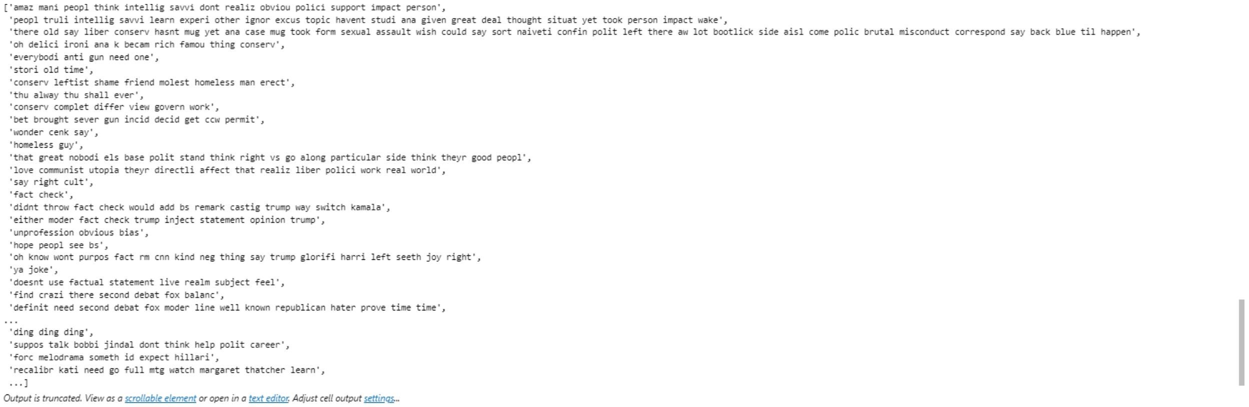
Reddit:



```
['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...

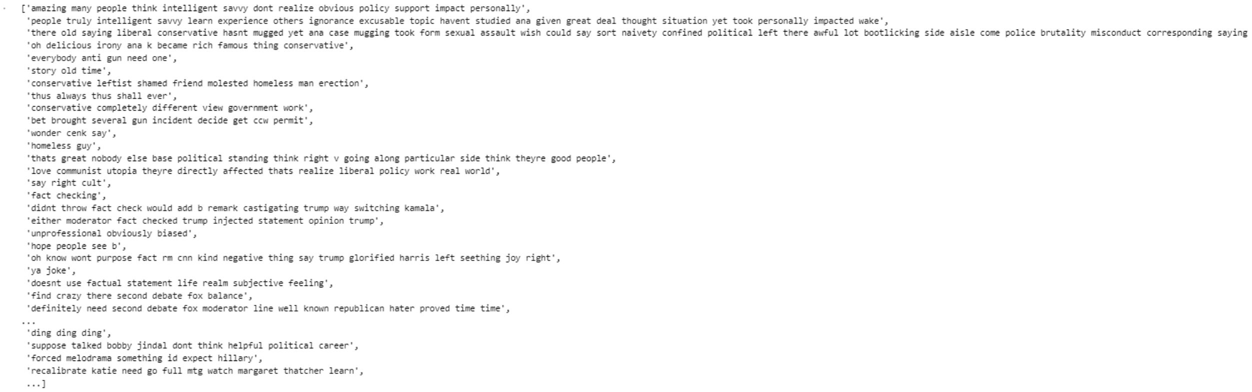
Stemming Output:



```
['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',
 ...]
```

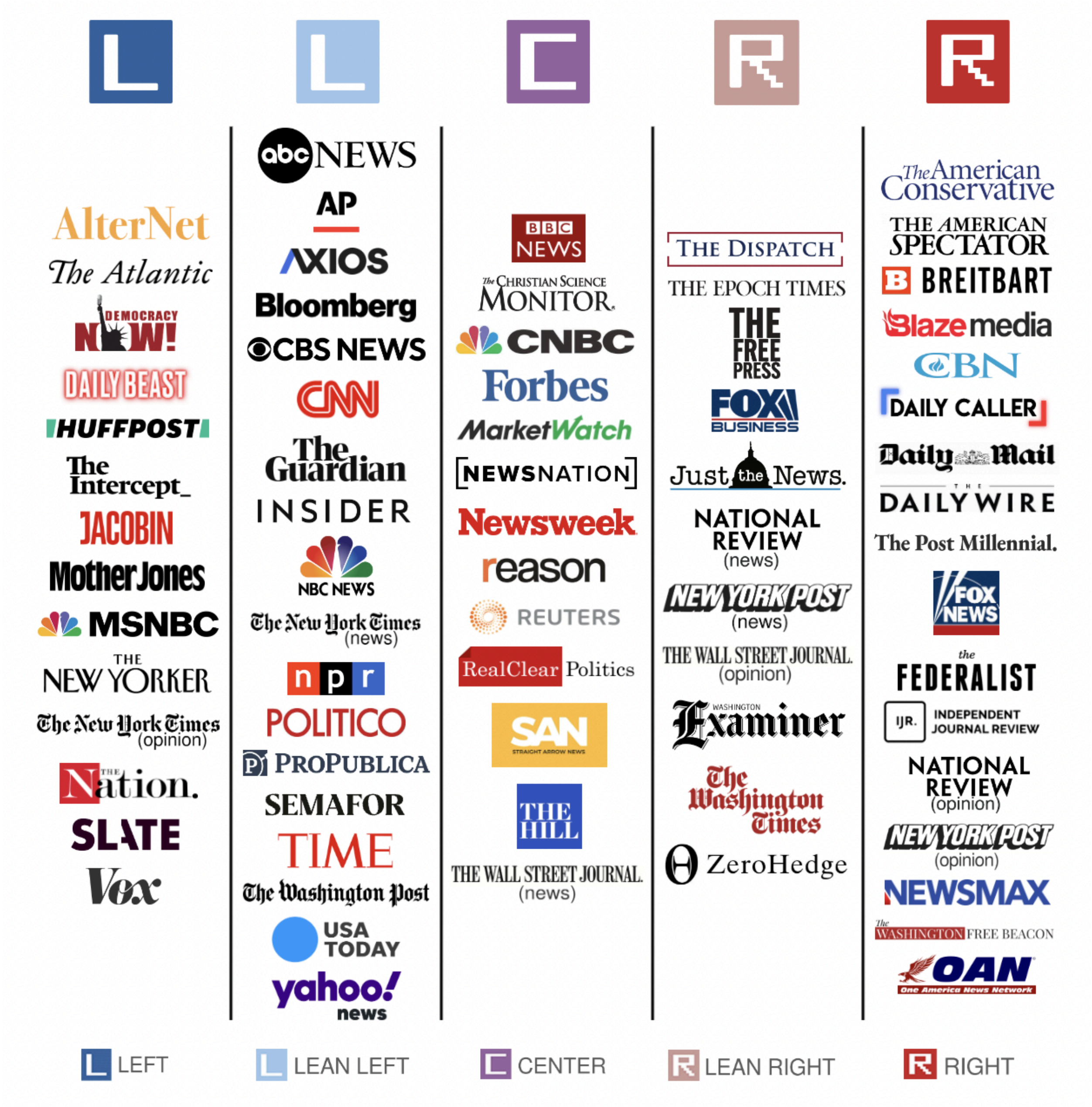
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

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',
    ...]
```

AllSides Media Bias Chart:



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