Algorithmic Accountability in Media: Delineating Political Bias through TFIDF and Logistic Regression

Ji Noh
Vanderbilt University
Eun.ji.noh@vanderbilt.edu

ABSTRACT

This study employs Term Frequency-Inverse Document Frequency (TFIDF) to quantify political bias in media, guided by the premise that language use within news content reflects underlying political leanings. Drawing from a comprehensive dataset of U.S. news articles, this paper examines linguistic patterns that differentiate conservative and liberal biases. The methodology involves rigorous data preprocessing to exclude non-informative characters while preserving contextual integrity. The study hypothesizes that TFIDF can reveal distinctive lexical trends corresponding to political orientations. Initial findings validate the hypothesis, showcasing the potential of TFIDF combined with logistic regression in classifying content with an accuracy of 72%. The investigation illuminates the subtle yet significant role of lexical choices in media bias, contributing to the broader dialogue on algorithmic influence in news consumption.

Keywords

Political Bias, Media Content Analysis, TFIDF, Natural Language Processing, Classification, Logistic Regression

1. INTRODUCTION

In the landscape of media content analysis, the quantification of political bias presents a formidable challenge. To address this, researchers have turned to cutting-edge tools such as algorithms, machine learning, and NLP techniques, which shed light on the intricate dynamics of news framing. Diakopoulos' pivotal review, "Algorithmic Accountability: Journalistic Investigation of Computational Power Structures," critically examines the interplay between media narratives and political biases through the lens of computational technology [1]. His work underscores the crucial role of algorithmic scrutiny in dissecting the subtle influences within digital ecosystems. This scrutiny is paramount in an era where news and information are increasingly filtered by computational algorithms, potentially shaping public political orientations and opinions [2].

Term Frequency-Inverse Document Frequency (TFIDF) emerges as a potent analytical method within this sphere. Celebrated for its proficiency in textual analysis, TFIDF assesses the significance of words within a corpus, balancing their frequency against their document-specific uniqueness [3]. This dual consideration illuminates terms that are prevalent in particular texts but not across the dataset, enabling a granular examination of lexical variations that reflect diverse political ideologies [4].

The decision to utilize TFIDF in this study is twofold: It demonstrates not only a sound methodological choice due to its revelatory power in detecting linguistic nuances but also aligns with the principles of algorithmic accountability Diakopoulos advocates [5]. By deploying TFIDF, we seek to contribute substantively to the discourse on media bias, enhancing the collective understanding of how algorithmically driven content construction may influence the ideological landscape [6].

2. PURPOSE STATEMENT AND RESEARCH QUESTION

The political orientation of media outlets in the United States has become a focal point of scholarly inquiry. With divergent political ideologies from liberal to conservative spectrums, these orientations distinctly affect how narratives are framed, topics are selected, and stories are portrayed, ultimately shaping public discourse, influencing policymaking, and affecting the functioning of democracy [7]. In this crucial juncture, it is imperative to quantitatively discern and classify the political bias ingrained within media coverage. The advent of sophisticated natural language processing (NLP) and text analysis techniques, particularly Term Frequency-Inverse Document Frequency (TFIDF), now permits a methodical scrutiny of the content disseminated by news organizations, thereby unveiling patterns suggestive of underlying political biases.

The following research question guide this study:

Can Term Frequency-Inverse Document Frequency (TFIDF) analysis distinguish between liberal and conservative news companies in the United States based on the content of their article?

Employing spaCy, a powerful NLP library, will facilitate nuanced linguistic processing, while scikit-learn's machine learning algorithms will serve to quantify and classify the identified linguistic features. Combined, these technologies offer a robust approach for dissecting the subtle linguistic cues that signal ideological underpinnings within media content.

To this end, the study posits the following hypothesis:

It is hypothesized that TFIDF analysis of news articles will reveal distinct lexical patterns that correlate with the political bias of the news source. Specifically, it is anticipated that conservative news outlets will exhibit a unique set of frequently used terms and narratives that differ significantly from those employed by liberal news organizations. These differential patterns are expected to be indicative of the respective political leanings, thereby providing a quantitative measure of bias within the media landscape.

3. LITERATURE REVIEW

The quantification of political bias within the media is a complex interplay of linguistics, politics, and technology. The burgeoning field of computational journalism has been examined in various scholarly reviews, notably by Diakopoulos, who explores the consequences of algorithmic decision-making within news media contexts [1]. This inquiry is particularly pertinent in an era where content curation is often algorithm-driven, significantly influencing public discourse and political stances [7].

Research into political bias in media often leverages the capabilities of Natural Language Processing (NLP) and Machine Learning. TFIDF stands out as a salient method within this domain. It evaluates the importance of a word not merely based on frequency but in relation to its uniqueness across documents, illuminating distinct language usage indicative of different political ideologies [8]. TFIDF's effectiveness in identifying keyword trends related to political leanings has been established, marking its critical role in media content analysis [3, 4].

While there is substantial research on algorithmic accountability and the biases of computational systems, Diakopoulos points out the need for more in-depth exploration of how these biases manifest within the linguistic constructs of media content [1]. Furthermore, studies on language influence suggest that lexical choices in media can shape and reflect cultural and political perceptions, emphasizing the importance of examining the nuances of word selection in news reporting [4, 8].

The political bias of media outlets can shape narratives and influence public opinion. Barberá et al. demonstrate the importance of ideological representation in media and its effects on political communication [7]. Their work reveals how political bias can manifest in different linguistic patterns, with implications for the perception and propagation of ideologies through news content [7].

A critical gap in the literature pertains to the dynamic nature of political language and the need for adaptive models that reflect changing political landscapes [9]. Traditional models may not account for the temporal shifts in party politics or societal issues, necessitating more flexible and evolving analytical approaches.

This study's purpose—to apply TFIDF in combination with logistic regression to distinguish political bias from media content—is aligned with the need identified in the literature for refined analytical tools capable of adapting to the intricate dynamics of political linguistics. The methodology employed responds to Diakopoulos' call for a more granular examination of computational influences on media narratives, aiming to enhance the collective understanding of ideological bias in news content. It addresses the research gaps by employing advanced computational techniques to parse through the subtleties of lexical choices and by proposing a model that can be continually updated to reflect current political discourse.

4. METHODS

For transparency and to facilitate reproducibility, the code underlying the methodologies and analyses described in this paper is openly available. Interested parties can access the complete set of Python notebook, detailed methodological explanations, and supplementary materials on GitHub at https://github.com/jinoh0731/NLP-Algorithmic-Accountability-in-Media. This repository includes all custom code utilized for data preprocessing, analysis, and visualization, as well as additional notes on the procedures and parameters selected for the study.

4.1 Raw Data

The dataset employed in this study was sourced from Kaggle.com, specifically the "All the News" dataset compiled by Andrew Thompson. It comprises an extensive collection of 142,570 articles from various news outlets across the United States, spanning the years 2000 to 2016. The data was initially segmented into three distinct sets due to Kaggle's size constraints and was subsequently merged into a single comprehensive dataset for analysis. This merging was facilitated using Panda library from Python.

4.1.1 Data Structure

The consolidated dataset features 142,570 rows and ten columns, encapsulating a significant breadth of data points. Each row represents a unique news article, characterized by several attributes including article ID, title, publication source, author, publication date, with year, month, and date being separate columns, URL, and the article content itself. This structure supports a robust analysis of media content over a substantial temporal span.

4.1.2 Missing Values Analysis

A preliminary assessment of the dataset revealed a generally high completeness across most fields. Specific columns exhibited notable deficiencies: the 'title' column lacked only two entries, whereas the 'author' and 'publication date' columns were missing 15,876 and 2,641 entries, respectively. Most critically, the 'URL' column was missing 57,011 entries, which impacts the ability to trace articles back to their original source directly. However, for the purposes of this study, which focuses primarily on the 'publication' and 'content' columns, the absence of null values in these fields negates the need for imputation measures.

4.1.3 Date Range and Relevance

The articles predominantly cover recent events, with the bulk of the data concentrated in the years 2016 and 2017. This temporal focus is particularly pertinent for examining news trends and media representation leading up to and following the 2016 U.S. presidential election, offering insights into the evolving media landscape during this critical period.

4.1.4 Diversity and Distribution of Publications

The dataset encompasses contributions from 15 distinct news outlets including New York Times, Breitbart, CNN, Business Insider, the Atlantic, Fox News, Talking Points Memo, Buzzfeed News, National Review, New York Post, the Guardian, NPR, Reuters, Vox, and the Washington Post. These outlets represent a wide range of political perspectives, extending from conservative through to liberal, with some maintaining a neutral stance. This diversity allows for a comprehensive analysis of political biases as they are represented across different media platforms.

An exploratory data analysis (EDA) of the dataset revealed a varied distribution among the contributing news organizations. Breitbart emerges as the most prolific contributor, furnishing the dataset with 23,781 articles. It is closely followed by the New York Post, which accounts for 17,493 articles, underscoring its significant influence within the dataset. Other prominent contributors include NPR and CNN, providing 11,992 and 11,488 articles, respectively. Both The Washington Post and Reuters also maintain a strong presence, each contributing over 10,000 articles. Publications such as The Guardian, The New York Times, The Atlantic, Business Insider, National Review, and Talking Points Memo offer a moderate range, between 5,214 and 8,681 articles. On the lower spectrum, Vox, Buzzfeed News, and Fox News contribute between 4,354 and

4,947 articles each, yielding a substantial base for analytical review.

4.2 Categorization of Political Bias

The categorization of these outlets into political biases—Conservative and Liberal—draws upon their editorial stances, the types of stories they prioritize, and the perspectives they commonly present. This classification is guided by the widely recognized AllSides Media Bias Chart, depicted in Figure 1, which serves as a benchmark for assessing media sources on their political leanings [10].

Figure 1. A diagram categorizing media outlets according to their political orientation.

4.2.1 Conservative Outlets

Within the dataset publications list, Breitbart and Fox News are classified under the Right category, noted for their conservative editorial perspectives. The New York Post and National Review are categorized as Lean Right, displaying a slightly less pronounced conservative bias. These outlets frequently highlight themes such as limited government, individual liberties, free markets, and traditional values, aligning with conservative ideologies.

4.2.2 Liberal Outlets

The Atlantic and Vox are positioned in the Left category, with a clear liberal editorial focus. CNN, The Guardian, NPR, The New York Times, The Washington Post, and Business Insider, categorized as Lean Left, often engage with topics such as social equality, environmental activism, and progressive social policies. Their narratives typically critique conservative positions and advocate for liberal viewpoints.

4.2.3 Moderate and Undefined Categories

Reuters is recognized for its neutral stance, not significantly leaning towards either conservative or liberal poles. Meanwhile, Buzzfeed News and Talking Points Memo are not explicitly categorized due to insufficient information on their political biases and are excluded from this specific analysis.

4.2.4 Data Segmentation Strategy

To ensure a balanced analysis, the dataset is divided into two primary categories: Conservative and Liberal. Each category includes an equal number of outlets from the Right and Lean Right for Conservatives, and from the Left and Lean Left for Liberals. This methodological approach, incorporating two sources from each of these four political subgroups, allows for a fair comparison while maintaining a diversity of sources within each political grouping.

4.3 Data Sampling Strategy

Upon segmenting the data according to political orientation, the dataset comprised 51,831 articles from conservative news outlets and 31,417 articles from liberal sources. To facilitate a focused and manageable analysis within the constraints of available computing resources, a decision was made to sample a subset of the data. Specifically, a random sampling method was employed to select 250 articles from each of the eight prominent news publications included in the study — Breitbart, Fox News, New York Post, and National Review for conservative group, and Atlantic, Vox, CNN, and New York Times for liberal group. This approach yielded a balanced dataset consisting of 2,000 articles, with an equal distribution of 1,000 articles each from conservative and liberal news outlets.

The rationale behind sampling only 2,000 articles was driven primarily by limitations in computational power. The computational resources available dictated a smaller, more manageable dataset to ensure the feasibility of conducting detailed text analysis using Term Frequency-Inverse Document Frequency (TFIDF) from scikit-learn and other natural language processing techniques using spaCy, without compromising the performance of the computing system. This sampling strategy, while reducing the volume of data analyzed, was designed to maintain a representative balance between the different political spectrums, thereby allowing for an effective comparison and analysis of linguistic patterns associated with each political bias.

4.4 Mitigation of Potential Bias in Textual Analysis

In the course of preparing data for analysis, several challenges pertaining to bias mitigation were addressed. This involved the systematic removal of emojis using regex patterns to reduce noise, and the excision of special characters and numerals that could impair machine learning models. Additionally, we manually corrected spacing anomalies in abbreviations and addressed isolated characters resulting from numeral removal, ensuring a pristine dataset for unbiased analysis.

4.4.1 Emoji Removal

Upon initial data inspection, a significant presence of emojis was noted within the article texts. While emojis enrich communication by conveying emotions and contextual cues, they can introduce substantial noise in datasets intended for natural language processing (NLP) tasks. To address this, a regular expression (regex) pattern was designed and employed to systematically identify and remove a comprehensive range of emojis, including

emoticons, symbols, pictographs, transport and map symbols, and flags, as delineated by Unicode standards. The application of this regex pattern effectively cleansed the dataset of these characters, reducing potential skew in machine learning model outcomes.

4.4.2 Special Character and Numeric Value Removal

Analysis revealed that certain special characters and numeric values embedded within the text could adversely impact the classification processes. These elements again, though occasionally conveying important information, predominantly introduced noise. A regex-based approach was utilized to remove these characters, retaining only alphabetic characters and essential punctuation that forms part of standard expressions (e.g., "Mr.", "Mrs.", contractions). This step was crucial in enhancing the clarity of the dataset, thereby facilitating more accurate recognition and classification of textual patterns by the analytical models employed.

4.4.3 Exclusion of Identifiable Keywords

To ensure the objectivity of the analysis, particularly in learning unbiased models, it was imperative to remove any text that could inadvertently hint at the origin of the news publication. The examination of the dataset revealed that a significant proportion, specifically 993 out of the 2000 sampled articles, contained explicit references to the source publication within the article text. These references were diligently expunged to ensure the neutrality of the dataset. This precautionary step was crucial to ascertain that the computational detection of bias was predicated entirely on the substantive content of the articles, rather than on any identifiable markers of their origins.

4.4.4 Manual Text Cleaning

Further preprocessing involved manual cleaning to correct errors introduced by earlier text cleaning stages. This included the correction of improperly spaced abbreviations such as "U.N." which appeared as "U. N." and similar issues with other initialisms like “F.B.I.” Additionally, the removal of numerals led to isolated alphabetic characters (e.g., the 's' in "1960s"), which were also addressed. This meticulous cleaning process was vital for maintaining the semantic integrity of the dataset.

4.5 Text Preprocessing and Tokenization Methodology

The preliminary stage of our analysis involved a comprehensive preprocessing routine to prepare the corpus for machine learning tasks. Texts were tokenized into constituent words, with punctuation and extraneous whitespace removed to isolate meaningful lexical items. Commonly occurring stop words were excluded, and the remaining tokens were lemmatized to their base forms and converted to lowercase to ensure uniformity. This process preserved the original index of each text, enabling seamless traceability to the original corpus.

The tokens were then reconstituted into a cohesive, processed string for each document, leading to a Data Frame with each row representing the refined version of the initial text. This standardization is critical in NLP, rendering the data apt for comprehensive analysis and modeling.

4.6 Analysis of Term Frequency Distribution

A term frequency analysis conducted on the corpus exhibited a Zipfian distribution, a linguistic pattern characterized by the high occurrence of a few words while the majority are rarely used [11]. To optimize our analytical process and enhance machine learning model performance, a critical parameter was applied: the 'min_df' within the scikit-learn's TfidfVectorizer. This parameter was set to a value of 4, to focus the analysis on words that appeared in at least four documents. This threshold was chosen based on the rationale that it would exclude the least frequent terms which are less likely to contribute to a reliable analysis due to their sparse occurrence, and instead concentrate on terms with enough presence across the dataset to suggest a meaningful pattern or trend.

The application of the TfidfVectorizer was pivotal in addressing the observed frequency disparities. It was set to interpret preprocessed tokens, delimited by spaces, bypassing the need for a customized tokenizer. The strategic 'min_df' parameter was pivotal in ensuring the inclusion of only those terms with considerable frequency, thereby sharpening the pertinence and integrity of the features selected for in-depth analysis.

5. RESULTS

5.1 Implementation of Logistic Regression for Classification

With the preprocessing phase concluded, the study proceeded to apply machine learning techniques to classify news articles into conservative and liberal categories based on their TF-IDF scores. The corpus was split into training and test sets, with an 80:20 ratio, with random sampling, ensuring both robust training and thorough evaluation of the machine learning model.

In the constructed pipeline architecture, TfidfVectorizer was employed for text vectorization while Logistic Regression served as the classification algorithm, establishing a vigorous framework for text data analysis. Logistic Regression, commonly utilized for binary classification tasks, computes the probability of a binary response based on one or more predictor variables [12]. It operates on the logistic, or sigmoid function, providing a model to estimate the probability that a given input falls into one of two categorical outcomes [12]. This process transformed the corpus into a numerical format amenable to interpretation by the Logistic Regression model. Once the model was trained on the designated training set, it was then applied to the test set to predict the political bias of the news articles therein.

5.2 Model Evaluation and Outcomes

For the purposes of this investigation, model evaluations, including accuracy, precision, and recall, along with the derived confusion matrix, are extracted strictly from the test set. This subset is separate from the data used to train the model, ensuring that the evaluation metrics reflect the model's performance on previously unseen data, which is critical for validating the model's predictive capabilities.

Table 1. Classification report from logistic regression model.

	Precision	**Recall**	**F1-Score**	**Support**
Conservative	0.71	0.74	0.72	199
Liberal	0.73	0.70	0.72	201
Accuracy			0.72	400
Macro Avg.	0.72	0.72	0.72	400
Weighted Avg.	0.72	0.72	0.72	400

The Logistic Regression model's evaluation, as presented in Table 1, illustrates an accuracy of 72%, signifying that the model was able to correctly identify the political bias of the news articles in 72% of the cases within the test set. The precision metric, calculated at 0.72 on a weighted average, reflects the model's high degree of exactitude in predicting article biases. When looking more closer, the model has a precision of 0.71 for conservative, and 0.73 for liberal biases, suggesting it is slightly more precise in identifying liberal articles. Moreover, the recall rate, also at 0.72 for weighted average, conveys the model's competency in detecting all pertinent instances of bias. The f1-score, which amalgamates precision and recall, corroborates the model's balanced proficiency in these metrics, with a consistent score of 0.72.

Figure 2 elucidates the Confusion Matrix from the Logistic Regression Model, which details the true positive, true negative, false positive, and false negative counts. The matrix showcases that conservative articles were correctly identified 147 times but were misclassified as liberal 52 times. Conversely, liberal articles were accurately classified 141 times and incorrectly labeled as conservative on 60 occasions.

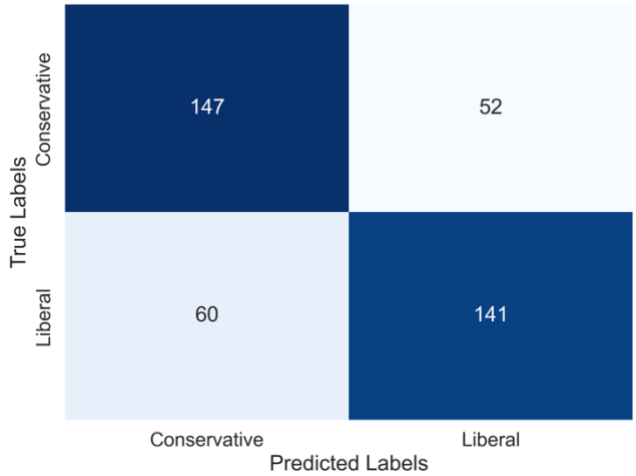

Figure 2. Confusion matrix from logistic regression model

The empirical results, as detailed in Table 1 and Figure 2, substantiate the hypothesis that TFIDF analysis can indeed detect distinct lexical patterns that are indicative of political bias in news content. The Logistic Regression model's ability to classify news articles into conservative and liberal categories with an accuracy of 72%—as evidenced by the precision, recall, and f1-scores provided in Table 1's Classification Report—demonstrates a significant predictive capability. This performance is notably higher than random chance, which would yield an accuracy of 50%. Figure 2's Confusion Matrix further reinforces this point, showing that the model distinguishes between biases with considerable reliability, substantially outperforming the baseline of random guessing. Thus, the integration of TFIDF with logistic regression emerges as a potent tool for media bias classification.

5.3 Coefficient Analysis

The coefficient analysis generated from the Logistic Regression model illuminates the lexical features most indicative of political bias within the news articles classified. The model leverages a vocabulary of terms, each weighted by coefficients that quantify their relative influence on the classification outcome. From Figure 3, the "coefficients_1" column delineates terms significantly correlated with conservative content; these terms, including "cruz", "percent", "report", "cop", "twitter", "immigration", "abortion", and "hillary" have negative coefficients, signaling their propensity to associate an article with conservative bias. This could be reflective of the conservative media's focus on policy areas such as immigration reform and pro-life advocacy, as well as the tendency to often reference political figures like Hillary Clinton in critical discourse.

	coefficients_1	vocabulary_1	coefficients_2	vocabulary_2
0	-1.316182	cruz	4.133149	mr.
1	-1.174077	percent	1.844995	ms.
2	-1.117921	report	1.480747	like
3	-1.083050	cop	1.193452	people
4	-1.047913	twitter	1.076182	united
5	-1.015983	claim	1.071965	story
6	-0.978560	follow	1.063716	update
7	-0.906623	immigration	1.034198	partner
8	-0.900817	yankees	1.019156	episode
9	-0.891165	abortion	0.998240	white
10	-0.878421	left	0.997538	ad
11	-0.871587	click	0.965188	country
12	-0.855822	de	0.955962	question
13	-0.853896	news	0.929334	change
14	-0.851496	associated	0.903500	likely
15	-0.837008	prediction	0.896181	series
16	-0.833874	rubio	0.874415	say
17	-0.830554	hillary	0.873296	young
18	-0.827638	clinton	0.869042	world
19	-0.804704	channel	0.868804	sponsor

Figure 3. Logistic Regression Model Coefficient for Political Bias Classification

Conversely, the "coefficients_2" column features terms with positive coefficients, such as "mr.", "ms.", "united", "white", "young", and "change" which, when present, elevate the probability of an article being categorized as liberal. Appearance of these words could possibly be due to these terms frequently surface in liberal media in discussions about social respect and change, youth activism, and progressive movements, which are commonly associated with liberal ideologies. The presence of these terms within the liberal news discourse could reflect an editorial emphasis on personal titles that denote formality and respect, societal evolution, and the involvement of younger generations in political activism [13].

Figure 3 illustrates the two ends of the coefficient spectrum. The leftmost columns, "coefficients_1" and "vocabulary_1," present the twenty terms most negatively associated with a liberal classification—effectively, they are predictors of a conservative classification. Inversely, the rightmost columns, "coefficients_2" and "vocabulary_2," encapsulate the twenty terms with the highest positive association with liberal content, thus serving as strong predictors of a liberal classification.

The interplay of these terms and their respective weights within the model's architecture provides a window into how language usage in the media can serve as a barometer for political leanings. For example, the negative weighting of "clinton" in relation to liberal bias might seem counterintuitive, yet it could exemplify how conservative outlets frequently reference opposition figures to critique or frame discussions, which is a common rhetorical strategy in politically charged reporting. Similarly, the use of honorifics such as "mr." and "ms." in articles that the model associates with a liberal bias could indicate a norm of formality in coverage of individuals, a stylistic choice that may resonate more with publications on the liberal end of the spectrum.

6. CONCLUSION AND FUTURE WORK

The investigation set out to detect if Term Frequency-Inverse Document Frequency (TFIDF) can effectively differentiate between liberal and conservative news organizations in the U.S. based on article content. The hypothesis posited that TFIDF would expose distinctive linguistic patterns corresponding with each political bias, as conservative and liberal outlets typically utilize varying lexicons and narratives.

6.1 Conclusion

The findings from logistic regression modeling, articulated in the form of coefficients and classification matrices, reveal a clear distinction between conservative and liberal biases in news content, thus validating the research hypothesis. With an accuracy surpassing random classification, the model's ability to correctly predict political bias in 72% of cases illustrates the predictive power of TFIDF when coupled with logistic regression techniques. Terms such as "cruz" and "immigration" showed a strong negative correlation with liberal content, hence indicating conservative bias, whereas terms like "mr." and "people" carried positive coefficients, thereby aligning more with liberal content.

The coefficients' analysis in Figure 3 provides a nuanced insight into the linguistic attributes that the logistic regression model deemed significant in demarcating the political orientation of news articles. This model's interpretability facilitates a deeper understanding of how specific lexical choices are weighted differently within the media's political landscape.

6.2 Limitation

This research encountered several notable limitations that could impact the robustness and generalizability of the findings. The absence of explicit labels for political bias within the dataset necessitated reliance on external sources for bias inference, which introduces a level of subjectivity. The delineation of news sources into specific political categories is inherently nuanced and not strictly binary; hence, the assignment of political leanings is prone to influence the outcomes of the model.

Furthermore, the data preprocessing phase, which included the removal of special characters and numerical figures, presented challenges. While such cleaning is essential for model training, it must be balanced against the risk of discarding potentially relevant information. The subtleties of political language, which can shift with the changing landscapes of party politics and societal issues, require a dynamic model capable of adapting over time to maintain accuracy.

An additional limitation is the lack of category labels for the articles. The political bias in news reporting can be heavily dependent on the subject matter, whether it concerns politics, economics, sports, or other topics. Without this categorization, the model's ability to discern how topical focus influences political bias is compromised, potentially obscuring the ways in which different issues are framed by media outlets of varying political stances.

6.3 Future Work

Looking ahead, the study opens avenues for further research to refine the classification models and expand their scope. While TFIDF combined with logistic regression has proven effective, the exploration of more complex models and algorithms may offer enhanced accuracy and deeper insights. Additionally, future studies could incorporate larger datasets encompassing a broader array of political ideologies and possibly integrate real-time data to account for the evolving political discourse.

Moreover, further refinement in preprocessing techniques could yield even more granular insights, and experimenting with different text representations, such as word embeddings or deep learning approaches, might provide a richer linguistic analysis. The incorporation of sentiment analysis and contextual features could also augment the detection of subtler forms of bias.

Lastly, interdisciplinary collaborations, integrating insights from political science, sociology, and ethics, would ensure a more holistic approach to understanding media bias. This would also align with the growing need for algorithmic accountability in media representation, thus contributing to the development of more equitable and transparent news dissemination practices.

7. REFERENCES

[1] Diakopoulos, N. (2014). Algorithmic Accountability: Journalistic Investigation of Computational Power Structures. *Digital Journalism*, *3*(3), 398–415. https://doi.org/10.1080/21670811.2014.976411.

[2] Anderson, C.W. (2012). Towards a sociology of computational and algorithmic journalism. *New Media & Society*, *15*(7), 1005–1021. https://doi.org/10.1177/1461444812465137

[3] Jiang, H. & Li, W. (2011). Improved Algorithm Based on TFIDF in Text Classification. *Advanced Materials Research, 403-408*. 1791-1794. 10.4028. https://doi.org/10.4028/www.scientific.net/amr.403-408.1791.

[4] He, X. & Zhou, X. (2015). Contrastive Analysis of Lexical Choice and Ideologies in News Reporting the Same Accidents between Chinese and American Newspapers. *Theory and Practice in Language Studies*, *5*(11). 2356. https://doi.org/10.17507/tpls.0511.21.

[5] Mei, A. L. (2024). How Does Language Influence Our Minds? From a Linguistics Perspective. *Lecture Notes in Education Psychology and Public Media*. *42*. 205-209. https://doi.org/10.54254/2753-7048/42/20240840.

[6] Lim, C. & Kim, S. (2024). Examining factors influencing the user's loyalty on algorithmic news recommendation service. *Humanities Social Science Communications, 11*(1), 10. https://doi.org/10.1057/s41599-023-02516-x

[7] Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber? *Psychological Science, 26*(10), 1531-1542. https://doi.org/10.1177/0956797615594620

[8] Demidov, D. (2023). Political Bias of News Content: Classification based on Individual Articles and Media.

[9] Rawat, S. & Vadivu G. (2022). Media Bias Detection Using Sentimental Analysis and Clustering Algorithms. *Proceedings of International Conference on Deep Learning, Computing and Intelligence*, 485–494. https://doi.org/10.1007/978-981-16-5652-1_43

[10] AllSides Media Bias Ratings™. (2023). AllSides Technologies, Inc. https://www.allsides.com/media-bias/media-bias-ratings. Retrieved April 2024.

[11] Shufaniya, A. & Arnon, I. (2022). A Cognitive Bias for Zipfian Distributions? Uniform Distributions Become More Skewed via Cultural Transmission. *Journal of Language Evolution. 7*(1), 59-80. https://doi.org/10.1093/jole/lzac005.

[12] Maalouf, M. (2011). Logistic regression in data analysis: An overview. *International Journal of Data Analysis Techniques and Strategies, 3*(3), 281-299. https://doi.org/10.1504/ijdats.2011.041335.

[13] Bruchmann, K. & Vincent, S. & Folks, A. (2023). Political bias indicators and perceptions of news. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2023.1078966.