Understanding Political Bias in News Text Classification

R. Batista
University of Porto, Porto, Portugal

(Dated: January 6, 2026)

Political bias detection in news text is an increasingly relevant topic at the intersection of Artificial Intelligence and society, raising concerns about transparency, bias, and the societal impact of automated decision-making systems. Rather than redefining political ideologies, this work treats political bias labels as given and investigates how text classification models learn and exploit different types of signals when performing political bias classification.

We study this problem through a comparative analysis of two datasets with complementary characteristics: one composed of full news articles, enabling the analysis of structural and source-related cues, and another consisting of short sentences annotated with biased words, allowing the study of semantic cues in isolation. Using TF-IDF-based models and BERT, combined with interpretability techniques and masking strategies, we analyze model behavior beyond standard performance metrics.

Our results show that models often rely on structural or source-related patterns in article-based datasets, while semantic cues alone are insufficient for robust classification in short texts. These findings highlight the importance of interpretability and dataset design when deploying political bias classification systems in socially sensitive contexts.

I. INTRODUCTION

The use of Artificial Intelligence (AI) for analyzing political content has grown rapidly in recent years, particularly in applications such as political bias detection, content moderation, and media analysis. While such systems promise scalability and efficiency, they also raise important socio-technical concerns related to transparency, accountability, and the societal impact of automated interpretations of political discourse.

Political bias classification is especially sensitive in this context. Automated systems that label news content as left, right, or center may influence public perception, reinforce existing biases, or be misused in content filtering and recommendation pipelines. From an AI and Society perspective, it is therefore not sufficient to assess these systems solely based on predictive performance. Understanding how models arrive at their decisions and which signals they rely on can be as important as the predictions themselves. Moreover, assigning political bias labels to news content is inherently complex. Political positioning is often contextual, nuanced, and dependent on cultural, temporal, and annotator-specific factors. This complexity makes political bias classification a particularly challenging task and further motivates the need to critically examine how automated systems learn from annotated data.

In this work, we focus on analyzing the behavior of text classification models trained on political bias annotations, rather than attempting to redefine or validate political ideologies. Specifically, we investigate whether models rely primarily on semantic content, such as ideological language and political entities, or whether they exploit alternative signals, including structural patterns and source-related cues present in news articles.

To this end, we conduct a comparative study using two datasets with complementary characteristics. The first dataset consists of full news articles and enables the analysis of structural and source-related signals that may act as shortcuts for classification models. The second dataset is composed of short sentences annotated with biased words, allowing us to study semantic bias cues in isolation, without article-level structure or source information. We evaluate both a traditional TF-IDF-based[1] classifier and a contextual language model (BERT)[2], combining standard performance metrics with interpretability techniques and masking strategies, such as LIME[3] and transformers_interpret library[4]. This approach allows us to move beyond accuracy-focused evaluation and examine the types of patterns learned by different models under varying data conditions. By contrasting structural and semantic signals across datasets and models, this report aims to contribute to a more nuanced understanding of automated political bias classification systems and to highlight the importance of interpretability and dataset design in socially sensitive AI applications.

II. RELATED WORK

Research on automated political bias classification has expanded significantly with the availability of annotated news corpora and the growing interest in computational analysis of political discourse. Several studies frame political bias detection as a supervised text classification problem, leveraging lexical, stylistic, and topical features extracted from news articles[5]. Publicly available datasets with ideological annotations have played a key role in this progress, enabling large-scale experimentation and comparative evaluation across approaches[3,6]. Early analyses demonstrate that political bias can often be inferred from recurring linguistic patterns and framing choices, even without explicit modeling of political ideology.

More recent work has explored neural and representation-based methods for political bias classification. Transformer-based language models, particularly those relying on contextual embeddings, have shown strong performance improvements over traditional bag-of-words approaches[2,7]. These models are capable of capturing richer contextual and syntactic information, allowing them to model subtle distinctions in political language. However, multiple studies caution that such performance gains may be driven by dataset-specific artifacts, topic correlations, or superficial lexical cues rather than a deep understanding of political semantics[8,9].

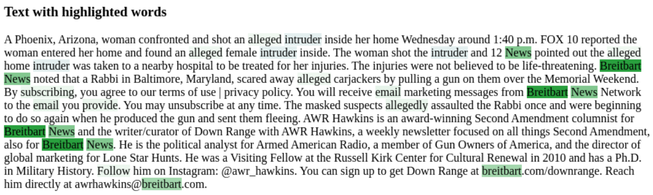
To address these concerns, recent research emphasizes interpretability and critical inspection of political bias classifiers. Explainability methods have been applied to analyze which words or textual patterns most strongly influence model predictions, revealing potential reliance on proxies or spurious correlations[10]. This line of work highlights the socio-technical risks associated with automated political bias analysis, including the reinforcement of annotation biases and the oversimplification of complex political positions[11]. Consequently, political bias classification is increasingly viewed not only as a predictive task, but also as an interpretability and accountability challenge.

III. EXPERIMENTAL SETUP

This section describes the datasets, models, evaluation procedure, and interpretability strategies used in our experiments.

Datasets. We use two publicly available datasets[6,9], selected to isolate different families of signals that can drive political bias classification: article-level structural and source-related cues, and sentence-level semantic cues with explicit span annotations.

Our first dataset is the Political Bias dataset by Santana[6], hosted on Kaggle, which contains news articles annotated with political bias labels. The original dataset provides five bias categories (left, slightly left, center, slightly right, and right). In this work, we merge these categories into a three-way classification setting (left, center, right) in order to reduce label sparsity and align the task with common political bias classification formulations. After this consolidation, the dataset contains 6,396 article instances. The instances correspond to full article text, which naturally includes artifacts of online news formatting and outlet-specific markers. During exploratory inspection of the raw text, we observed recurring structural and source-related cues (e.g., "READ MORE", cookie or banner snippets, and outlet names) that are not directly related to ideological content but may act as shortcuts for classification models (Figure 1). Importantly, each news source in the dataset is consistently associated with a single political bias label, meaning that all articles from the same outlet share the same target class.



Text with highlighted words

A Phoenix, Arizona, woman confronted and shot an alleged intruder inside her home Wednesday around 1:40 p.m. FOX 10 reported the woman entered her home and found an alleged female intruder inside. The woman shot the intruder and 12 News pointed out the alleged home intruder was taken to a nearby hospital to be treated for her injuries. The injuries were not believed to be life-threatening. Breitbart News noted that a Rabbi in Baltimore, Maryland, scared away alleged carjackers by pulling a gun on them over the Memorial Weekend. By subscribing, you agree to our terms of use | privacy policy. You will receive email marketing messages from Breitbart News Network to the email you provide. You may unsubscribe at any time. The masked suspects allegedly assaulted the Rabbi once and were beginning to do so again when he produced the gun and sent them fleeing. AWR Hawkins is an award-winning Second Amendment columnist for Breitbart News and the writer/curator of Down Range with AWR Hawkins, a weekly newsletter focused on all things Second Amendment, also for Breitbart News. He is the political analyst for Armed American Radio, a member of Gun Owners of America, and the director of global marketing for Lone Star Hunts. He was a Visiting Fellow at the Russell Kirk Center for Cultural Renewal in 2010 and has a Ph.D. in Military History. Follow him on Instagram: @awr_hawkins. You can sign up to get Down Range at breitbart.com/downrange. Reach him directly at awrhawkins@breitbart.com.

FIG. 1: Example of LIME-based word importance visualization.

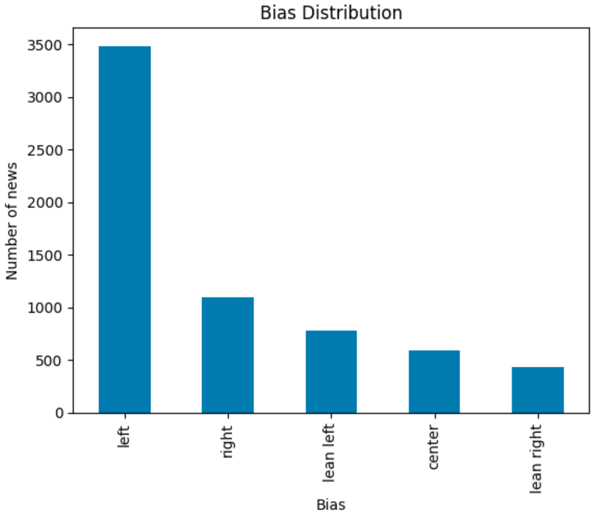


FIG. 2: Distribution of Bias (Dataset 1).

Preprocessing was applied to ensure data consistency. Instances with missing values in the link field were removed, as the absence of source information prevents verification of the news outlet. For a small number of instances with missing article text or placeholder values (e.g., "error fetching article"), the article title was used as a fallback textual input. Although this reduces the length of the input text, it preserves semantically meaningful content and avoids discarding otherwise valid instances. The class distribution after label consolidation is imbalanced (Figure 2), with approximately 66.7% left-leaning articles, 24.0% right-leaning articles, and 9.3% center articles. We intentionally do not apply resampling or class-balancing techniques. Beyond avoiding the introduction of synthetic data, this choice reflects the inherent ambiguity of the center category, which often overlaps semantically with the political extremes and may be artificially distorted by balancing strategies.

Our second dataset is the *Media Bias Including Characteristics* (MBIC) dataset, introduced by Spinde et al.[9]. MBIC is designed to support fine-grained analysis of media bias and explicitly includes annotator background information. The dataset comprises approximately 1,700 short textual statements, each annotated by multiple annotators (ten per statement in the initial release), and provides political bias labels at both the word level (biased spans) and the sentence level. In our experiments, we focus exclusively on the subset of statements anno-

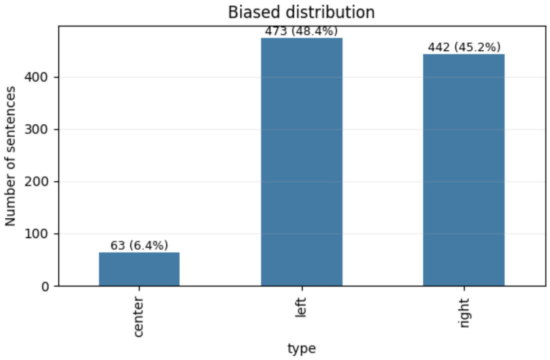
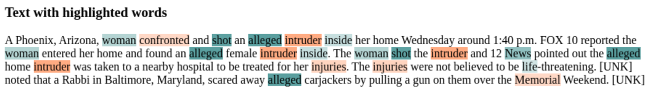


FIG. 3: Distribution of Bias (Dataset 2).

tated as biased, resulting in 978 instances. Each statement is treated as a single text instance with a three-way political bias label (left, center, right). We leverage the provided word-level biased annotations to perform semantic masking experiments, while also evaluating models on the original, unmasked text to assess the impact of these annotations. Preprocessing consisted of removing empty or null statements, which did not contain textual content relevant for classification. Unlike the article-based dataset, MBIC instances contain substantially less contextual information and no explicit source or structural cues. The class distribution (Figure 3) in this subset is also imbalanced, with approximately 48.4% left-labeled instances, 45.2% right-labeled instances, and 6.4% center-labeled instances. As with the first dataset, we preserve the original class imbalance.

Models. We evaluate two fundamentally different text classification approaches in order to contrast their behavior, representational capacity, and reliance on different types of signals. As a strong and widely used baseline for text classification, we employ a Logistic Regression classifier operating on TF-IDF representations[1]. This model provides a transparent and interpretable decision function, making it particularly suitable for analyzing which lexical and structural features contribute to political bias predictions. Text is represented using a TF-IDF vectorizer with unigrams and bigrams, lowercasing enabled, English stopword removal, a minimum document frequency of 2, and a maximum document frequency of 0.95. These settings reduce noise from rare tokens while limiting the influence of overly frequent terms. Logistic Regression is trained with a maximum of 2,000 iterations and uses class-balanced weights to mitigate the impact of class imbalance during optimization. This choice allows the model to remain sensitive to minority classes without altering the original data distribution.

To contrast the linear model with a contextual representation, we fine-tune a BERT-based sequence classification model[2]. BERT encodes text using deep bidirectional self-attention, allowing it to capture contextual and semantic relationships beyond surface-level lexical patterns. This makes it well suited for analyzing whether richer representations reduce reliance on structural or shortcut cues observed in simpler models. We use a pre-trained BERT model with a classification head adapted to a three-class political bias prediction task. The model is fine-tuned for five epochs using a learning rate of 2×10^{-5} and a batch size of 8 for both training and evaluation. Weight decay is set to 0.01 to regularize training, and class weights are computed from the training data to address class imbalance during optimization. Model selection is based on macro F1-score, ensuring balanced performance across classes rather than favoring majority labels.


Text with highlighted words

A Phoenix, Arizona, woman confronted and shot an alleged intruder inside her home Wednesday around 1:40 p.m. FOX 10 reported the woman entered her home and found an alleged female intruder inside. The woman shot the intruder and 12 News pointed out the alleged home intruder was taken to a nearby hospital to be treated for her injuries. The injuries were not believed to be life-threatening. [UNK] noted that a Rabbi in Baltimore, Maryland, scared away alleged carjackers by pulling a gun on them over the Memorial Weekend. [UNK]

FIG. 4: Example of an masked LIME-based word importance visualization.

For both datasets, models are trained and evaluated using a fixed 80/20 train/test split. Performance is assessed using standard classification metrics, including precision, recall, and F1-score for each class (left, center, right), as well as macro-averaged and weighted F1-scores. Given the strong class imbalance present in both datasets, macro F1-score is used as the primary metric for model selection and comparison, as it assigns equal importance to all classes. In addition to standard training and evaluation, we perform masking-based experiments. For the first dataset, as we can see in Figure 4, expressions related to article structure or source-specific artifacts (e.g., formatting elements) are manually identified and replaced with a special `[UNK]` token. For the MBIC dataset, annotated biased words are replaced with `[UNK]` to explicitly remove lexical bias cues. Masking is applied either at inference time on the first dataset, and also during training on the second dataset. All experiments are conducted under identical evaluation conditions to allow for a fair comparison between models and datasets.

Interpretability. To analyze model behavior beyond aggregate performance metrics, we combine local and global interpretability methods with masking-based probing strategies.

For the TF-IDF-based model, we use LIME[3] to generate local explanations, highlighting the most influential tokens for individual predictions. In addition, we analyze global feature importance by inspecting model coefficients aggregated at the class level, providing an overview of which features are consistently associated with each political bias label. For the BERT-based model, we employ the transformers_interpret library[4] to obtain local, token-level attribution scores for individual predictions. Interpretability analyses are conducted on a curated set of representative instances selected to cover different prediction scenarios. Specifically, we analyze correctly and incorrectly classified examples from each class (left, center, right), spanning different confidence regimes. These include high-confidence predictions (confidence > 0.75), low-confidence predictions (confidence in the range $[0.5, 0.55]$), as well as misclassified instances with both high and low confidence. This selection allows us to examine not only successful predictions but also model uncertainty and failure modes. Regarding the masking experiments, first models are evaluated on the original, unmasked text, and both predictions and explanations are analyzed. Next, masking is applied and the same instances are re-evaluated to assess changes in predicted class, confidence, and explanatory patterns. As said in the previous section, the masking strategy differs across datasets according to their characteristics. In addition to inference-time masking, we conduct further experiments on the MBIC dataset by training models directly on masked text. This allows us to assess whether models can adapt to the absence of explicit lexical bias cues during learning, as opposed to merely reacting to their removal at inference time. And finnaly we did a final experiment which consisted on training the TF-IDF-based model on only the biased words.

IV. RESULTS AND DISCUSSION

This section presents and discusses the experimental results obtained across both datasets, combining quantitative evaluation with interpretability and masking-based analyses to better understand model behavior. Due to space and readability constraints, detailed instance-level interpretability outputs are provided in the accompanying notebooks, as including them in this format or in the appendix would render them largely illegible.

A. Dataset 1: Structural and Source-Related Cues

TABLE I: Classification performance on Dataset 1 (article-based).

Model	Class	Precision	Recall	F1-score
TF-IDF + LR	Left	0.895	0.953	0.923
	Center	0.670	0.636	0.652
	Right	0.942	0.789	0.859
	Macro Avg	–	–	**0.811**
BERT	Left	0.918	0.953	0.935
	Center	0.764	0.686	0.723
	Right	0.913	0.851	0.881
	Macro Avg	–	–	**0.846**

Table I reports the classification performance of the TF-IDF + Logistic Regression model and BERT on the article-based dataset. Overall, both models achieve strong performance, particularly for the left and right classes. In contrast, the center class consistently exhibits lower recall and F1-score. This behavior can be attributed both to higher semantic ambiguity and to the strong class imbalance present in the dataset, as center instances represent a substantially smaller portion of the data compared to left- and right-leaning articles. The TF-IDF-based model achieves a macro F1-score of 0.811, with strong performance on left-leaning articles (F1 = 0.923) and slightly weaker results for right-leaning articles (F1 = 0.859). Performance on the center class is substantially lower (F1 = 0.652), suggesting that surface-level lexical features are less effective at capturing centrist positioning. BERT improves overall performance, reaching a macro F1-score of 0.846. Gains are most noticeable for the center class (F1 = 0.723), indicating that contextual representations help mitigate, but do not fully resolve, ambiguity in centrist content. Despite this improvement, center remains the most challenging class for both models.

Local explanations obtained with LIME for the TF-IDF-based model reveal a strong reliance on non-semantic cues, as illustrated in the notebooks. Frequently highlighted tokens include structural expressions such as “READ MORE” and “Enable Social Cookies”, as well as source identifiers like “Breitbart” and “Getty Images”. While political entities (e.g., “Trump”, “Republicans”) also appear in explanations, especially for left-classified instances, structural and source-related tokens often dominate the decision process.

This behavior is further confirmed by global feature importance analysis. The most influential features across classes correspond primarily to formatting-related or source-specific terms rather than ideological language. The consistency between local and global explanations suggests that shortcut learning based on publication-specific patterns is embedded in the model’s overall decision strategy, rather than being limited to isolated predictions.

To assess the impact of these cues, structural and source-related expressions are masked and the same instances are re-evaluated. The effects of masking vary substantially across classes. For left-leaning instances, masking leads to a modest reduction in prediction confidence (e.g., from 0.94 to 0.88), while the predicted class remains unchanged. Explanations shift toward political entities and content-related terms, indicating that semantic cues are sufficient to sustain the prediction. In contrast, right-leaning instances exhibit a substantial drop in confidence (e.g., from 0.98 to 0.51), accompanied by increased probability mass assigned to other classes. Local explanations become dominated by generic action-related words without clear ideological meaning, providing strong evidence that structural or source-related shortcuts play a central role in these predictions. Center instances are particularly unstable: masking often leads to drastic confidence reductions and, in some cases, class switches (e.g., from center to left), highlighting the sensitivity of centrist predictions to small semantic variations once structural signals are removed. A summary of confidence changes be-

TABLE II: Prediction probabilities before and after structural masking for representative instances (Dataset 1).

Instance	**Original**			**Masked**		
	Center	Left	Right	Center	Left	Right
Left	0.030	**0.937**	0.032	0.057	**0.883**	0.059
Right	0.011	0.011	**0.978**	0.169	0.325	**0.506**
Center	**0.933**	0.031	0.036	0.345	**0.358**	0.297

fore and after masking is reported in Table II.

Interpretability analysis for BERT shows a different behavior. Local explanations are generally more diffuse and harder to interpret, but predictions for clearly left- or right-leaning instances remain almost unchanged after masking, with confidence remaining close to 1.0. However, for center instances, structural masking can still induce high-confidence class changes. These results suggest that BERT is more robust to overt structural shortcuts than the TF-IDF-based model, but remains sensitive in borderline and ambiguous cases.

B. Dataset 2: Semantic Bias Cues

Table III reports the classification performance of the TF-IDF + Logistic Regression model and BERT on the MBIC dataset. Compared to the first dataset, performance is substantially lower for both models across all classes. This reflects the increased difficulty of the task, which is characterized by a smaller number of instances, short sentence-level inputs, limited contextual information, and the absence of structure- or source-related cues.

TABLE III: Classification performance on the MBIC dataset (sentence-level, no masking).

Model	**Class**	**Precision**	**Recall**	**F1-score**
TF-IDF + LR	Left	0.68	0.77	0.72
	Center	0.25	0.08	0.12
	Right	0.73	0.69	0.71
	Macro Avg	–	–	**0.52**
BERT	Left	0.68	0.63	0.66
	Center	0.25	0.08	0.12
	Right	0.64	0.75	0.69
	Macro Avg	–	–	**0.49**

For the TF-IDF-based model, performance is moderate for the left and right classes (F1-scores of approximately 0.72 and 0.71, respectively), while the center class is poorly learned (F1-score of 0.12). BERT does not produce a clear improvement in this setting, achieving a macro F1-score of 0.487, which is comparable to or slightly below the TF-IDF baseline. These results suggest that, in the absence of broader contextual signals, contextual language models do not provide a substantial advantage over surface-level representations.

TABLE IV: Classification performance on MBIC with biased words masked during training.

Model	**Class**	**Precision**	**Recall**	**F1-score**
TF-IDF + LR	Left	0.66	0.69	0.68
	Center	0.25	0.08	0.12
	Right	0.67	0.70	0.69
	Macro Avg	–	–	**0.49**
BERT	Left	0.56	0.46	0.51
	Center	0.17	0.39	0.24
	Right	0.57	0.57	0.57
	Macro Avg	–	–	**0.44**

Local interpretability analysis without masking reveals a markedly different behavior compared to Dataset 1. For the TF-IDF-based model, explanations are dominated by semantic and entity-level cues, such as political actors, groups, or identity-related terms (e.g., references to political parties or social groups). Unlike the article-based setting, no structure-related tokens appear, confirming that predictions are driven primarily by lexical content. Representative LIME explanations illustrating this behavior are provided in the notebooks. For BERT, local explanations obtained with transformers_interpret are more difficult to interpret, with importance distributed across multiple tokens and sub-tokens, making it challenging to identify consistent decision patterns.

When semantic masking is applied at inference time, prediction confidence changes only marginally and the predicted class often remains unchanged. This suggests that bias-related information is not confined to a small set of explicitly annotated words, but rather distributed across the sentence. We further evaluate models trained directly on masked text, where biased words are replaced during training. In this setting, performance degrades for both models, as shown in Table IV. The effect is particularly pronounced for the TF-IDF-based classifier, whose macro F1-score drops to 0.49, indicating that the removal of explicit lexical bias cues limits the model's ability to learn stable decision boundaries.

The most extreme case corresponds to training exclusively on biased words, which results in very poor performance and strong class confusion (Table V). In this scenario, the TF-IDF-based model achieves a macro F1-score of only 0.34, with severe degradation for the left and center classes. This outcome highlights that isolated biased words, without surrounding linguistic context, are insufficient for reliable political bias classification.

Overall, results on the MBIC dataset emphasize the limitations of semantic-only approaches. Sentence-level inputs with minimal context provide weak supervision signals, limiting both classification performance and interpretability, even when explicit bias annotations are available.

TABLE V: Classification performance on MBIC using only biased words as input (TF-IDF + LR).

Class	**Precision**	**Recall**	**F1-score**
Left	0.57	0.22	0.32
Center	0.10	0.08	0.09
Right	0.50	0.84	0.62
Macro Avg	–	–	**0.34**

C. Societal and Ethical Implications

The results of this study raise important sociotechnical considerations regarding the deployment of automated political bias classification systems. High predictive performance, particularly in article-based settings, may conceal reliance on superficial cues rather than on substantive ideological analysis. Such behavior can lead to misleading interpretations of political content and undermine trust in AI-driven media analysis tools. The contrast between structural and semantic settings further highlights the role of dataset design in shaping model behavior. Systems trained on datasets where political bias is strongly correlated with news outlets risk encoding and amplifying existing media stereotypes, while sentence-level approaches struggle to capture bias without broader context. In both cases, centrist content emerges as especially vulnerable to misclassification, which may contribute to polarization by implicitly forcing nuanced positions into binary categories.

From an society perspective, these findings emphasize the importance of transparency and interpretability when developing and deploying models for such sensitive tasks. Interpretability tools and targeted analyses, such as masking experiments, are essential for uncovering hidden dependencies and for assessing whether model decisions align with socially acceptable notions of fairness and accountability.

V. CONCLUSION

In this work, we studied political bias classification from a data-centric and interpretability-driven perspective, focusing on how different datasets and modeling choices influence learned decision strategies. By comparing an article-based dataset rich in structural and source-related cues with a sentence-level dataset designed to isolate semantic bias signals, we showed that high predictive performance does not necessarily imply meaningful ideological understanding. Instead, models often rely on shortcuts that emerge from dataset-specific properties. Masking experiments and interpretability analyses further revealed that explicitly annotated biased words are insufficient on their own, and that political bias is distributed across linguistic context rather than localized in isolated terms. Although contextual models such as BERT exhibit increased robustness, they remain sensitive in ambiguous cases, particularly for centrist content.

This study has several limitations. The analysis is restricted to two datasets and a limited set of models, and interpretability methods provide only partial insights into complex decision processes. Future work could explore additional datasets and alternative bias formulations to better assess the societal impact of automated political bias classification systems. Overall, our findings highlight the importance of combining performance evaluation with interpretability and careful dataset design when addressing politically sensitive AI applications.

[1] M. Das, S. K., and P. J. A. Alphonse, "A comparative study on tf-idf feature weighting method and its analysis using unstructured dataset," (2023), arXiv:2308.04037 [cs.CL].

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," (2019), arXiv:1810.04805 [cs.CL].

[3] S. Lim, A. Jatowt, M. Färber, and M. Yoshikawa, in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, edited by N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis (European Language Resources Association, Marseille, France, 2020) pp. 1478–1484.

[4] C. Pierse, J. Hessel, *et al.*, "Transformers interpret: Explainability for transformer models," `https://github.com/cdpierse/transformers-interpret` (2020).

[5] Media Bias Research, "Publications," (2024).

[6] M. Santana, "Political bias dataset," `https://www.kaggle.com/datasets/mayobanexsantana/political-bias` (2020).

[7] R. Baly, G. Da San Martino, J. Glass, and P. Nakov, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, edited by B. Webber, T. Cohn, Y. He, and Y. Liu (Association for Computational Linguistics, Online, 2020) pp. 4982–4991.

[8] W.-F. Chen, K. Al Khatib, H. Wachsmuth, and B. Stein, in *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, edited by D. Bamman, D. Hovy, D. Jurgens, B. O'Connor, and S. Volkova (Association for Computational Linguistics, Online, 2020) pp. 149–154.

[9] T. Spinde, B. Plank, K. Krippendorff, *et al.*, Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (2021).

[10] M. T. Ribeiro, S. Singh, and C. Guestrin, "”why should i trust you?”: Explaining the predictions of any classifier," (2016), arXiv:1602.04938 [cs.LG].

[11] T. Spinde, L. Rudnitckaia, J. Mitrović, F. Hamborg, M. Granitzer, B. Gipp, and K. Donnay, Information Processing Management **58**, 102505 (2021).

VI. ACKNOWLEDGEMENTS

This document was reviewed and refined with the assistance of Large Language Models, which helped mostly on translating, correcting grammar, and enhance clarity. It was also used in the code, mostly to comment and generate the graphics. The overall content and ideas remain solely the responsibility of the author.