Identifying and Mitigating Bias in News Articles

Seema Vora, Ananya Sini Achan, Surabhi Gupta

Abstract

Political bias in news articles influences public opinion and reinforces ideological divides. This study develops a two-step NLP pipeline using the BABE dataset to detect and mitigate bias. RoBERTa, achieving 71% accuracy, classifies articles by political leaning, while a fine-tuned LLaMA model rephrases biased text into neutral expressions. Results show reductions in biased language, providing a framework for unbiased news consumption and future advancements in bias mitigation.

1 Introduction

In today's hyper-connected digital age, the news we consume often reflects the ideological biases of its sources, subtly shaping our perspectives and reinforcing existing beliefs. This is particularly evident in political journalism, where media outlets frequently cater to specific political parties, leading to polarized audiences. As our society becomes more divided along political lines, there is a growing need for tools that enable readers to access unbiased, factual information. Without such tools, currently individuals are falling into echo chambers that reinstate their beliefs and are less likely to have constructive dialogues about opposing ideologies.

Political bias in media has been widely studied, and these studies have shown that biased language, selective topic framing, and word choice can influence readers' perceptions and opinions. Political bias often manifests in subtle linguistic choices that align content with specific ideological views, reinforcing stereotypes and misrepresentations. Our research aims to identify these biases and aim to remove these biases by experimenting with various approaches.

So far research in this space focuses independently on classifying bias or removing bias terminology. Our research seeks to develop a model capable of detecting political bias within news articles and subsequently debias politically leaning content. First, we aim to experiment with a few techniques which will help us classify the political leaning of a text as right-leaning, left-leaning or center. Second, we introduced a novel approach by experimenting with text-generation methodologies using LLaMA and compared its performance to the existing DBias methodology (Raza et al., 2022). By combining classification and generative techniques, our research explores a novel pipeline for identifying and reducing political bias in textual data. Ultimately, our work aspires to contribute to a more informed public, and broadening perspectives in today's fragmented media landscape. In order to carry out this task, we have split the problem into two steps and created models at each step:

1. Bias Detection: Detecting the bias in the incoming news text and allocating it to either left, right or center.
2. Bias Removal: Once the bias is detected, we will either remove/replace it to curate unbiased versions of the news text.

First, the model is fine-tuned on a dataset annotated for political leanings to serve as a classifier. For this step we are experimenting with BERT as our control model, and variants of BERT such as AlBERT, DistilBERT, RoBERTa and LLaMA as our treatment models. Then, biased sentences identified are passed through LLaMA, which generates rephrased, neutralized versions using prompt-based guidance. We compare our recreation of an existing technique known as DBias (control model) vs LLaMA (treatment model). This combination of classification and debiasing aims to create a pipeline that can both identify and reduce political bias in text effectively, contributing to ongoing efforts in NLP to promote fairness and objectivity in language generation.

2 Related Work

A. Bias Detection in Text

Past research has sought various methods to quantify and measure bias in texts, including manual annotation, dictionary-based approaches, and machine learning-based classification. Deep learning models, particularly transformer-based architectures, capture subtle biases in word choice and context within text using word embeddings. Studies have shown that BERT-based models can accurately classify content along ideological lines by identifying linguistic cues associated with political biases such as, left-leaning or right-leaning (K. Rakhecha et al, 2023).

Additionally, methods such as SimCSE (Simple Contrastive Learning of Sentence Embeddings) have been developed to identify consistent patterns in political discourse across news sources. SimCSE leverages a logistic regression classifier to effectively identify consistent choices in framing sentences in news articles using contrastive learning (Muhammad Nadeem et al, 2022).

B. Removing Bias in Text

On the other hand, generative models like LLaMA, developed by Meta, have emerged as powerful tools for text generation and rephrasing tasks. By rephrasing or suggesting neutral alternatives to politically charged language, generative models can play a vital role in mitigating bias in text. However, effectively guiding these models to produce neutral output requires careful prompt engineering, fine-tuning, and evaluation (Feng et al., ACL 2023).

The challenge of debiasing language models has gained considerable attention in NLP research. Common approaches include fine-tuning models on curated, neutral datasets, applying adversarial training to discourage biased language, and using dictionary-based or lexicon-based methods to replace biased words with neutral terms. The DBias framework, for example, leverages Transformers to detect, mask, and replace biased words in news articles (Raza et al., 2022). Using named entity recognition (NER) and masked language modeling (MLM), DBias flags bias-bearing words and replaces them with neutral terms, ensuring the semantic integrity of the text while mitigating bias. The model fine-tunes DistilBERT and RoBERTa for bias detection and recognition, achieving high accuracy while ensuring fair representations in news media (Raza et al., 2022).

3 Methods

Dataset

We use the Neural Media Bias Detection Using Distant Supervision With Bias Annotations By Experts (BABE) dataset for this study, which is a comprehensive resource designed to advance research on media bias in the United States (Timo et al, 2021). The dataset includes articles from a diverse range of media outlets, categorized by political leanings—right, center, and left—based on each source's established ideological perspective. Each article is segmented into sentences, and each sentence inherits the political bias label of the article, facilitating detailed sentence-level bias detection. Furthermore, annotations highlight polarized language, making the dataset particularly suitable for both classification and debiasing tasks. The processed training dataset includes 2,026 sentences, balanced across political leanings, with a focus on highly debated topics such as Black Lives Matter, Taxes, Universal Health Care, and Gun Control (Spinde et al., Findings 2021).

Political Leaning	Right	Left	Center
No. of Sentences	759	749	518

Table 1: Dataset Breakdown.

Models

A. Baseline Models

We adopt a two-step approach to address media bias in text: (1) **Bias Classification** to detect and categorize sentences as leaning right, left, or center, and (2) **Bias Removal** to neutralize politically polarized language.

Model 1: Bias Classification

BERT for Classification: We fine-tune a classic BERT model to classify sentences in the BABE dataset with configuration: learning rate (6.78×10^{-5}), batch size (32), dropout (0.150). BERT's contextual embeddings are expected to capture ideological cues from each sentence, allowing the model to learn subtle distinctions in language that indicate political bias.

Model 2: Removing Bias in Text

DBIAS Model: Raza et al. (2022) discuss bias-sensitive language modeling for political ideology assessment called DBias. Using that as reference, we employed a similar model to detect and neutralize polarized words in each sentence. Initially the intention was to utilize and fine tune upon the open-source DBias pip package, but ultimately we decided to build our model from scratch. The DBias model was not successful on our dataset when utilizing NER to detect biased words, so we leveraged the dataset provided biased words to create our base model. Upon the creation of the biased word bank, the model masks polarized terms before passing the text to the classifier. We evaluate the DBIAS model's effectiveness in bias detection and compare it with subsequent treatment models to assess its capacity for generalizing across different political contexts.

B. Treatment Models

To improve the baseline models, we fine-tuned a range of BERT variants and LLAMA models.

Model 1: Bias Classification

1. **Suite of BERT models**: We fine-tune multiple kinds of BERT models to classify each sentence from the BABE dataset as leaning right, left, or center. The results can be seen in Table 2. Hyperparameters for our BERT models were optimized using the Optuna framework with a Successive Halving Pruner.
   a. **ALBERT**: A lightweight variant of BERT with fewer parameters, was fine-tuned for sentence-level political bias classification using the albert-base-v2 architecture. ALBERT leverages shared parameters and factorized embedding parameterization to efficiently capture contextual information (Lan, Z. et al, 2020). The final configuration was: learning rate (4.52×10^{-5}), batch size (16), epochs (5), and weight decay (0.027).
   b. **RoBERTa**: A transformer model optimized with a more extensive pre-training dataset and longer sequences, enhancing its sensitivity to biased language. The following configuration was used: learning rate (7.84×10^{-6}), batch size (64), and dropout rate (0.24).
   c. **DistilBert:** A distilled, smaller version of BERT offering a balance between speed and performance. The smaller size allows for faster training and inference. The initial configurations used were: learning rate (2.36×10^{-5}), batch size (16), epochs (2) and weight decay (0.01).

2. **LLAMA for Classification:** In addition to BERT, we also created a LLAMA model. The specific model used is “Llama-2-7b-hf” and it was hypertuned to work for classification purposes. This was by hypertuning the model to use a task type of “SEQ_CLS” along with adding epoch based strategies to classify rather than generate. Moreover, compute metrics were added to study how the classification prompts performed. It’s also important to note that LLAMA models are more attuned for generation purposes rather than classification (Yang & Rush, 2023).

Model 2: Removing Bias in Text

LLaMA for Neutralization

To neutralize politically biased text, we fine-tuned the LLaMA-2-7b-hf model using QLoRA (Quantized Low-Rank Adaptation), a memory-efficient fine-tuning technique that applies low-rank updates to the pre-trained model's weights. This enabled us to tune LLaMA for bias neutralization within our computational constraints.

We utilized the Wiki Neutrality Corpus (WNC) dataset (Pryzant et al., 2019) which comprises 180,000 sentence pairs extracted from Wikipedia revisions. These sentence pairs were created as part of Wikipedia editors' adherence to the Neutral Point of View (NPOV) policy, where subjective and biased language was rewritten into neutral phrasing. The corpus spans multiple domains, including political speeches and news headlines. The biased-neutral pairs were used for fine-tuning LLaMA to learn the transformation from subjective to neutral language. To focus on the most informative pairs:

1. **Random Sampling**: Reduced dataset to 20,000 sentence pairs for manageability.
2. **Similarity Filtering**: Cosine similarity between biased (source) and neutral (target) sentences was computed using TF-IDF vectorization. The 2,000 sentence pairs with the lowest similarity scores (indicating significant edits) were selected for fine-tuning.
3. **Prompt and Response Construction**: Each sentence pair was transformed, e.g.: Prompt: *"Rewrite the following sentence to be neutral and unbiased: <biased sentence>"* , Response: *"Rewritten sentence: <neutral sentence>".*

The fine-tuning process employed 4-bit quantization to optimize memory usage. The LoRA configuration included a rank of 8, a LoRA alpha of 16, and a dropout of 0.05, targeting the q_proj and v_proj layers. Training was conducted with a batch size of 1, gradient accumulation over 16 steps, 3 epochs, and a learning rate of 2×10^{-4}, ensuring computational efficiency. After fine-tuning, the LoRA weights were merged with the base LLaMA-2-7b-hf model to produce a standalone debiasing model.

The final model was deployed to generate neutralized sentences using structured prompts. The generation process was configured with a temperature of 0.25 to ensure low randomness and controlled output variability. To enhance diversity, top-k sampling was set to 50, while top-p sampling with a nucleus value of 0.9 allowed for dynamic token selection. A repetition penalty of 1.2 was applied to discourage redundant outputs and prevent the model from merely repeating the input sentence.

Multiple prompts were experimented with to optimize the model's performance such as: *"Rephrase the text to eliminate political bias while preserving its factual meaning: <biased sentence>."* Through iterative refinement, the final input prompt chosen was: *"Rewrite the following sentence to have neutral sentiment, remove any political bias while maintaining its original meaning. The rewritten sentence should be neutral and fact-based: <biased sentence>."*

Model	Accuracy	F1 Score	Precision	Recall
BERT (Baseline)	0.61	0.61	0.62	0.61
ALBERT (Treatment)	0.60	0.61	0.61	0.60
RoBERTA (Treatment)	0.71	0.71	0.71	0.72
DistilBert (Treatment)	0.68	0.68	0.70	0.67
LLaMA (Treatment)	0.10	0.19	1	0.10

Table 2. Bias Classification Results

4 Results and Discussion

The evaluation for the classification and generation models were done using separate performance metrics. The sections below cover the metrics in detail for both processes.

Bias Classification

The BERT (Baseline) model demonstrated a modest performance, achieving an accuracy of 0.6130 and an F1 score of 0.6091. These results establish a baseline for evaluating the effectiveness of the treatment models.

Treatment Models

1. **ALBERT**'s performance was similar to BERT, with slightly lower accuracy (0.6037) and F1 score (0.6052). This suggests that ALBERT may not capture ideological nuances as effectively as other transformer models in this task.
2. **RoBERTA** outperformed other models significantly, achieving a near-perfect accuracy, F1 score, precision, and recall (~0.71 across all metrics). This highlights its ability to detect political bias with high reliability, likely due to its robust pretraining on extensive datasets.
3. **DistilBERT** achieved moderate performance with an accuracy of 0.68 and an F1 score of 0.68. While it is computationally efficient, its performance does not match that of RoBERTa, indicating a trade-off between speed and accuracy.
4. **LLaMA** fine-tuned for bias neutralization, struggled in classification tasks, achieving an accuracy of only 0.103 and an F1 score of 0.187.

The results indicate that **RoBERTa** is the most effective model for this task, leveraging its extensive pretraining to achieve superior results. The poor performance of LLAMA classification highlights its limitations as a classifier, likely due to its design as a generative model rather than a discriminative one. LLaMA is pre trained as a general-purpose language model, not specifically optimized for classification tasks, which might require fine-tuning on labeled data to align with task-specific objectives. In fact, it is usually deemed that BERT models outperform major LLMs in classification tasks (Yang & Rush, 2023). This raises issues such as tokenization mismatches, improper loss function setup, or class imbalance in the dataset could hinder learning and cause incorrect fitting. In this case, even after changing parameters, we notice a low accuracy (Yang & Rush, 2023).

Bias Removal

Initially our aim was to determine the efficacy of our bias removal models by sending our results to popular LLMs such as GPT 4.0. However, although these large language models are being used quite often, there is ample room for improvement. Large language models (LLMs) are typically "liberal-leaning about political topics" (Bang et al., 2024). The paper *"Measuring political bias in LLMs"* analyzed 14 politically divisive topics on various models, demonstrated that a larger language model did not mean that the models were unbiased, and models from the same family did not have the same political leaning. We chose not to do manual review as well since the definition of bias for different individuals can vary. As a result of these findings we assessed the effectiveness of the baseline DBIAS model and the treatment LLAMA model for bias removal using a combination of classifier-based analysis and, sentiment evaluation:

1. **Classifier Neutrality**: Using the most effective classifier from the classification models, we used the fine-tuned RoBERTa bias classifier to measure the proportion of debiased sentences classified as center, indicating successful neutralization.
2. **Sentiment Analysis**: Sentiment shifts were analyzed using the Twitter-RoBERTa sentiment model (Francesco et al., 2020), comparing distributions and mean scores before and after debiasing.
3. **D4Data Bias Detection** (Raza et al., 2022) evaluated changes in political bias (right, left, center) pre- and post-neutralization.

Table 3. **RoBERTA Classifier (Model 1)**

Category	**Original**	**After DBIAS**	**After LLAMA**
Center Leaning	0%	14%	13%

Table 4. **Twitter-RoBERTa Sentiment Model**

Category	**Original**	**After DBIAS**	**After LLAMA**
Neutral Sentiment	38%	46%	44%

Table 5. **D4 Data Bias Detection Model**

Category	**Original**	**After DBIAS**	**After LLAMA**
Non-Biased	43%	38%	50%

The results of the bias removal task show progress and challenges in neutralizing politically biased text. For instance, the right-leaning sentence "It's a mafia state. There's no control. There's no regulation…" was debiased by LLaMA to "Some critics accuse the government of lax control over organized crime and lack of transparency in elections." While LLaMA demonstrated potential in debiasing, QLoRA's fine-tuning limitations reduced its ability to fully adapt to complex bias nuances. The subjective nature of bias definitions posed challenges, as perceptions of neutrality vary. Despite some improvements (6–7%) in metrics like classifier neutrality and sentiment shifts, the findings highlight the need for more robust fine-tuning techniques with greater computation abilities to address the complexities of political bias effectively.

5 Conclusion

In today's world, it is prevalent to come across extremely biased and one-sided news sources, which can be problematic as it doesn't provide a holistic and informative view of the actual news. As such, this paper explored ways to detect bias and then remove bias from news articles with the goal of creating unbiased summaries of news articles. We created a series of models for bias detection followed by models for bias removal. In summary, we successfully classified bias but had limited success in effectively removing it from the text.

6 References

K. Rakhecha, S. Rauniyar, M. Agrawal and A. Bhatt, "A Survey on Bias Detection in Online News using Deep Learning," 2023 2nd

International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2023, pp. 396-403, doi: 10.1109/ICAAIC56838.2023.10140917.

CS224N, Stanford, Custom Project, Muhammad Nadeem, Sarah Raza, • Mentor and Lucia Zheng. “Detecting Bias in News Articles using NLP Models.” (2022).

Yang, B., & Rush, A. M. (2023). LLaMA Adaptation: Efficient Fine-Tuning of Language Models with Causal Masking. Retrieved from https://arxiv.org/pdf/2310.01208.

Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 2021. Neural Media Bias Detection Using Distant Supervision With BABE - Bias Annotations By Experts. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1166–1177, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pryzant, R., Martinez, R. D., Dass, N., Kurohashi, S., Jurafsky, D., & Yang, D. (2019). Automatically Neutralizing Subjective Bias in Text. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2109–2115. https://aclanthology.org/N19-1221

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. arXiv preprint arXiv:1909.11942. https://doi.org/10.48550/arXiv.1909.11942

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1644–1650, Online. Association for Computational Linguistics.

Raza, S., Reji, D. J., & Ding, C. (2022). Dbias: Detecting biases and ensuring fairness in news articles. *International Journal of Data Science and Analytics*, 1–21. https://doi.org/10.1007/s41060-022-00359-4

Bang, Y., Chen, D., Lee, N., & Fung, P. (2024). *Measuring political bias in large language models: What is said and how it is said*. Centre for Artificial Intelligence Research (CAiRE), The Hong Kong University of Science and Technology.