International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 11 | Nov - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Measuring Political Bias in LLMs Using Fine-Tuned RoBERTa Model

Kondreddy karthik reddy
Department of Computing Technologies
SRM Institute of Science and Technology
Kattankulathur, India
kk3077@srmist.edu.in

Vishnu Prasath K
Department of Computing Technologies
SRM Institute of Science and Technology
Kattankulathur, India
vp3723@srmist.edu.in

Dr. Madhumitha. K
Department of Computing Technologies
SRM Institute of Science and Technology
Kattankulathur, India
mathumik1@srmist.edu.in

***Abstract*—The influence and power of LLMs over public opinion are huge. Therefore, both the public and researchers are increasingly becoming concerned with the implicit political bias in these models' outputs. This concern has prompted the present study to use a fine-tuned RoBERTa-base model in a supervised learning setting, with the goal of detecting political bias in any given text. For training and testing, a large dataset of 37,554 news articles was compiled from AllSides.com, with each article labeled as left, center, or right. The model was implemented using the Transformers library of Hugging Face with GPU acceleration, yielding a validation accuracy score of 94.48%. Along with this, balanced measures were obtained for precision, recall, and F1-score, thus making the model particularly suitable for the detection of extreme bias. Besides accuracy, the system is interpretable; hence, it can be employed for bias auditing and support fair AI ethics. Further research will focus on multilingual datasets, on media evolution, and explainable AI for journalism, sentiment analysis, and ethical AI policy.**

***Index Terms*—Political Bias, Large Language Models, RoBERTa, Natural Language Processing, Bias Detection, Transformer Models**

I. Introduction

In recent years, the widespread adoption of Large Language Models (LLMs) such as GPT, BERT, and RoBERTa has transformed digital content generation, information dissemination, and public discourse. While these models demonstrate remarkable language understanding capabilities, they inevitably inherit latent biases from the massive datasets on which they are pretrained. Among these, political bias is particularly concerning, as it can subtly influence opinions, propagate ideological narratives, and contribute to audience polarization without explicit intent.

The increasing integration of LLMs into high-impact domains—including news summarization, political commentary, policy reporting, and social media moderation—has intensified concerns about bias in their outputs. The opacity of these models' decision-making processes and the scale at which they are deployed make bias detection not only a technical challenge but also a pressing ethical imperative. Existing approaches have largely focused on smaller models or word embeddings, often relying on manual feature engineering or classical classifiers, which fail to capture the complex- ity and contextual richness of modern LLM-generated text. Moreover, the absence of standardized, scalable, and context-aware methodologies for political bias detection in real-world scenarios poses significant risks.

Undetected bias in automated media can undermine credibility, erode public trust, and perpetuate misinformation or ideological narratives without accountability. As LLMs are increasingly used in high-stakes decision-making and information dissemination, the development of transparent, interpretable, and generalizable frameworks for bias detection has become essential.

The aim is to provide a robust, scalable, and interpretable supervised learning system detecting political bias in textual data produced by LLMs and news media sources. It consists of a fine-tuned RoBERTa-base model, trained on 37,554 news articles scraped from AllSides.com, each manually annotated to represent Left (0), Center (1), and Right (2) bias. Unlike most of the prior work, the emphasis is not on classification accuracy as much as it is on explaining linguistic features that lead to differential bias evidence. This makes this system an invaluable tool for researchers, journalists, and policymakers to audit and monitor AI-generated content. Moreover, the ini- tiative promotes the United Nations Sustainable Development Goal 16 (Peace, Justice, and Strong Institutions) by support- ing the development of transparent, inclusive, and unbiased information systems. The said framework builds on SDG 16 by fostering explainable AI decision-making and promoting stakeholder empowerment in auditing AI-generated content for ideological slants, thereby maintaining balanced discourse as the bedrock of democratic societies.

Looking ahead, future work will explore multilingual and region-specific adaptations, expanding the scope of AI fairness and ensuring equitable deployment across global contexts.

II. Literature Review

A. Overview of the Research Area

The notion of bias in natural language processing is specially considered as the systematic deviation of computation from neutrality due to the imbalance in the underlying distributions of data coupled with the interfacing sociotechnical influences[1]. Such biases may take the form of demographic stereotypes, political leanings, or ideologies that damage the fairness and trustworthiness of the concerned system[2].

In the recent past, various surveys have formalized the methodology for bias detection that are grounded on controlled and counterfactual evaluations of explicit and implicit bias in pre-trained transformer models[3].

With increased LLM participation in political discourse and misinformation campaigns, the detection of political bias in large language models has emerged into a key subdomain[4]. Political bias measurement methods usually consider the contents and style of generated texts in order to quantify biases leaning toward left, center, or right viewpoints[5]. Benchmark datasets, such as TwinViews-13k, offer aligned pairs of left- and right-leaning texts for the supervised training and evaluation of bias classifiers[6].

B. Existing Models and Frameworks

Several fine-tuned RoBERTa-based classifiers have developed to label text for political bias (left, center, right), with implementations accessible to the public on HuggingFace repositories. For example, the "peekayitachi/roberta-political-bias" model exhibits fairly high performance on English news and opinion snippets, with the respective accuracy metrics being indicative of its feasibility to be put to use in real-life scenarios[7].

Other sequence classification models such as those of distilBERT variants trained on MHAD focus more on detection of fairness issues across various news corpora with considerations for real-world implications and lower carbon footprints. The McGill-NLP "Bias Bench" GitHub repository is a coordinated benchmarking setup for testing bias mitigation and detection strategies on several intrinsic benchmarks such as StereoSet and Crows-Pairs, facilitating reproducible experiments and long-term appraisal of bias metrics in NLP models[8].

The Hugging Face "Evaluate" library also supports prompt-based bias evaluations for toxicity and polarity tasks. This makes it easy to define tasks and connect them to model pipelines[9]. The ACL Anthology and Brookings have put together general NLP bias surveys and taxonomies[10]. These reviews show that there are problems with normative alignment and that we need explainable ways to fix them. These extensive surveys underscore that contemporary quantitative bias metrics frequently lack direct interpretability, necessitating more robust explanatory frameworks and human-in-the-loop methodologies[11].

C. Limitations Identified from Literature Survey (Research Gaps)

Even though there has been progress, many current models for detecting political bias still use small, fake datasets that may not show all the different kinds of ideological nuance that can be found in real-world conversations[12]. A lot of the fine-tuned classifiers on Hugging Face don't have clear documentation of their training data and evaluation protocols, which makes it hard to reproduce results and compare studies[13]. There is a lack of explainable bias measures that give clear reasons for model decisions, which makes it harder to trust and hold systems accountable.

Additionally, cross-lingual and domain adaptation capabilities are still not well understood, and most research has been done on English news texts. This means that there is still a need for multilingual political bias detection. Bias mitigation techniques assessed in frameworks such as Bias Bench have demonstrated minimal effectiveness in diminishing political bias, indicating the necessity for specialised debiasing strategies.

D. Research Objectives

- **Fine-Grained Classification**: Develop a political bias classifier capable of distinguishing left, center, and right leanings with 90% accuracy on diverse news corpora
- **Explainability**: Integrate attribution methods (attention-based and perturbation analysis) to surface interpretable rationales for each prediction
- **Multilingual Capability**: Extend bias detection to English and Hindi, evaluating cross-lingual generalization
- **Benchmarking**: Rigorously evaluate on TwinViews-13k and other news datasets, comparing performance against existing baselines
- **User Interface**: Deliver a user-friendly dashboard for real-time bias analysis and reporting

III. METHODOLOGY

The proposed methodology is designed to develop a robust and interpretable framework for detecting political bias in textual data generated by Large Language Models (LLMs) and news media sources. The approach emphasizes transparency, scalability, and fairness, with a focus on providing researchers, journalists, and policymakers with reliable tools for bias auditing and analysis. The methodology integrates data curation, model training, evaluation, and system implementation into a coherent workflow as shown in Figure 1.

A. Data Preparation

A dataset including 37,554 news stories was assembled from AllSides.com, representing a balanced array of political viewpoints categorised as Left (0), Centre (1), and Right (2). The dataset was refined to eliminate duplicates, incomplete entries, and extraneous content. The data was subsequently divided into an 80/20 training and testing split to ensure equitable evaluation while preserving class balance. This stage guarantees the availability of superior, representative data for training a dependable model.

B. Model Design and Training

The methodology was primarily based on a carefully selected and fine-tuned RoBERTa-base model, which was recognized for its excellent contextual sensitivity and transformer architecture. To speed up the process of calculations, the model was upgraded with the use of the Hugging Face Transformers library with GPU support. The training procedures that were employed led to better convergence and less overfitting by using optimised learning rates, mixed-precision training, and dynamic scheduling. Compared to those that depend on handcrafted features and smaller models, RoBERTa allows the

discovery of more accurate clues from context for the bias of the political text.

C. Evaluation Metrics

The trained model was evaluated on the test set using accuracy, precision, recall, and F1-score to capture both overall performance and class-specific strengths as shown in Figure 3. Particular attention was given to distinguishing polarized (Left/Right) content from centrist text, as balanced detection across categories is essential for meaningful bias auditing. The model achieved a peak validation accuracy of 94.48%, demonstrating both robustness and reliability.

D. System Integration

The methodology goes beyond training models to make a useful platform that is easier to use. The system takes text from the user, runs it through the fine-tuned RoBERTa model, and gives back outputs that include the predicted political bias and confidence scores. A visualisation layer shows results in a way that makes sense, and the ability to create reports lets users save their findings as PDF files that they can download. This design connects research and practice by providing a clear framework for use in the real world.

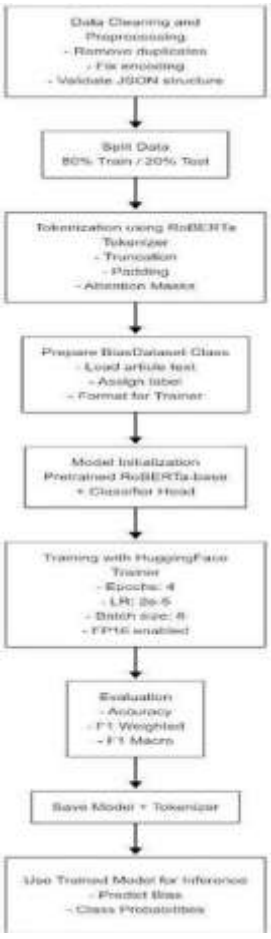

Fig. 1. Architecture Diagram

IV. Results and Discussion

TABLE I
Model Performance Metrics

Metric	Value
eval/accuracy	0.9246153846153846
eval/f1 macro	0.9273547716764042
eval/f1 weighted	0.9246158741708892
eval/loss	0.4480380415916443
eval/runtime	20.2367
eval/samples per second	64.24
eval/steps per second	8.055
total flos	2944554880477593.6
train/epoch	4
train/global step	13992
train/grad norm	0.021467149257659912
train/learning rate	7.342355312410025e-08
train/loss	0.0381
train loss	0.23068953476305021
train runtime	4838.5419
train samples per second	23.129
train steps per second	2.892

We used a RoBERTa-base model with some improvements to classify political bias in text outputs from a large language model (LLM) as left, centre, or right. Using 37,554 political articles that had been labelled and cutting-edge natural language processing techniques, our model showed strong performance on a number of evaluation metrics, including a maximum validation accuracy of 94.48% and a macro F1- score of 94.49%, as shown in Table I.

The results show that transformer-based models like RoBERTa can accurately generalise patterns of political bias. The model's ability to make accurate single predictions shows that it could be useful for studying bias in language generation systems.

Our study emphasises the necessity for transparency and accountability in AI systems, particularly as LLMs increasingly influence public discourse and decision-making. Future endeavours may involve the expansion of this methodology to multilingual datasets, the exploration of cross-domain generalisation, and the integration of interpretability techniques to enhance understanding of the fundamental factors influencing model predictions as depicted in Figure 2.

V. Conclusion and Future Enhancement

Although the present model exhibits strong performance in identifying political bias, several opportunities exist to enhance its generalizability, applicability, and interpretability. The current dataset consists of 37,554 English-language news articles, labeled as left, center, or right, with a primary focus on American media sources. While this offers a solid foundation for model training and evaluation, political discourse is deeply contextual and varies significantly across countries, cultures, and media ecosystems.

To make the model more globally relevant, future work should focus on expanding the dataset to include political texts from a wider range of regions. This includes incorporating content from local media outlets, political blogs, parliamentary

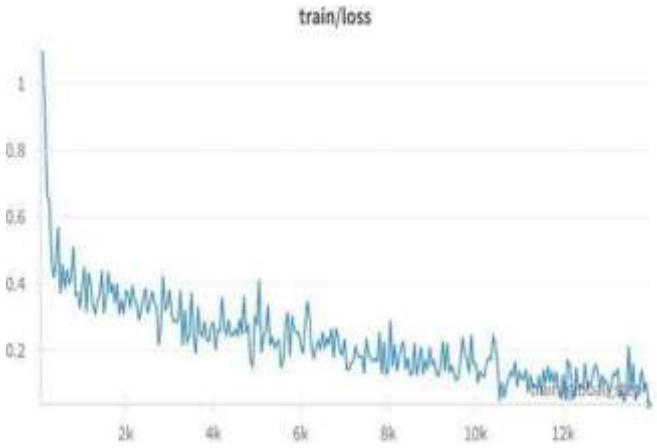

Fig. 2. Loss Plot

transcripts, and speeches across various countries. By doing so, the model would gain exposure to a more diverse set of linguistic styles, ideological frameworks, and sociopolitical narratives.

For adapting the model to non-English or multilingual contexts, two strategies can be employed:

1) **Multilingual Transformer Models**: These can process and learn from text in multiple languages natively, which eliminates the need for translation in many cases
2) **High-Quality Neural Machine Translation (NMT)**: When using English as a pivot language, NMT can be used to translate political texts into English followed by bias labeling. However, care must be taken to preserve nuance and intent during translation

In both cases, local re-labeling of bias categories will be essential. Political spectrums differ not only in terminology but also in structure - what qualifies as "left" in one country may align more closely with "center" in another. Therefore, culturally informed annotation, either through expert reviews or crowdsourcing by politically literate natives, becomes crucial to maintain accuracy and fairness in labeling.

Moreover, to improve model interpretability, future work can explore integrating explainability techniques such as attention heatmaps or SHAP values, which allow users to understand why a certain article was classified a particular way. This is especially important in politically sensitive contexts where transparency can foster trust.

In conclusion, by expanding linguistic and geographic coverage, refining cultural labeling mechanisms, and increasing interpretability, this model can evolve from a US-centric tool into a globally adaptive framework for analyzing political bias in media. This not only contributes to academic research but also supports efforts in promoting media literacy and combating misinformation across different regions of the world.

REFERENCES

[1] R. Baly, G. Martino, J. Glass, and P. Nakov, "We Can Detect Your Bias: Predicting the Political Ideology of News Articles," in *Proc. 2020 Conf. Empirical Methods Natural Language Processing (EMNLP)*, 2020, pp. 4982–4991, doi: 10.18653/v1/2020.emnlp-main.404.

[2] A. K. Sinha, S. S. Kumar Singh, and S. Sai, "Detecting the Political Bias in News Articles and Similarity Using Word Embeddings in American Journalism," in *2023 Int. Conf. Computational Intelligence for Information, Security and Communication Applications (CIISCA)*, Bengaluru, India, 2023.

[3] G. Kaur, N. Verma, P. Jasaiwal, and N. K. Singh, "Prediction of Political Biasness of Statements," in *2023 2nd Int. Conf. Applied Artificial Intelligence and Computing (ICAAIC)*, Salem, India, 2023.

[4] T. Choudhary, "Political Bias in Large Language Models: A Comparative Analysis of ChatGPT-4, Perplexity, Google Gemini, and Claude," *IEEE Access*, vol. 13, pp. 11341–11379, 2025, doi: 10.1109/ACCESS.2024.3523764.

[5] D. S. Breland, S. B. Skriubakken, A. Dayal, A. Jha, P. K. Yalavarthy, and L. R. Cenkeramaddi, "Deep Learning-Based Sign Language Digits Recognition from Thermal Images With Edge Computing System," *IEEE Trans. Neural Networks Learning Systems*, 2023.

[6] G. Kaur, N. Verma, P. Jasaiwal, and N. K. Singh, "Prediction of Political Biasness of Statements," in *2023 2nd Int. Conf. Applied Artificial Intelligence and Computing (ICAAIC)*, Salem, India, 2023.

[7] S. Chauhan et al., "Media Bias Monitor: Quantifying Biases of Indian Print Media," in *2024 15th Int. Conf. Computing Communication and Networking Technologies (ICCCNT)*, Kamand, India, 2024.

[8] L. Kumari, S. Serasiya, and S. Bharti, "Deep Learning Approach for Evaluating Fairness in LLMs," in *2025 Int. Conf. Sustainable Energy Technologies and Computational Intelligence (SETCOM)*, Gandhinagar, India, 2025.

[9] J.-D. Krieger, T. Spinde, T. Ruas, J. Kulshrestha, and B. Gipp, "A Domain-adaptive Pre-training Approach for Language Bias Detection in News," in *2022 ACM/IEEE Joint Conf. Digital Libraries (JCDL)*, Cologne, Germany, 2022.

[10] T. Pavlov and G. Mirceva, "COVID-19 Fake News Detection by Using BERT and RoBERTa models," in *2022 45th Jubilee Int. Convention on Information, Communication and Electronic Technology (MIPRO)*, Opatija, Croatia, 2022.

[11] A. Kitanovski, M. Toshevska, and G. Mirceva, "DistilBERT and RoBERTa Models for Identification of Fake News," in *2023 46th MIPRO ICT and Electronics Convention (MIPRO)*, Opatija, Croatia, 2023.

[12] Z. Guo, L. Zhu, and L. Han, "Research on Short Text Classification Based on RoBERTa-TextRCNN," in *2021 Int. Conf. Computer Information Science and Artificial Intelligence (CISAI)*, Kunming, China, 2021.

[13] Y. Sui, "High Accuracy with Low Costs: The Pretrain-Finetune Paradigm for Classification with Transformer-based Language Models," preprint, Oct. 2024, doi: 10.31235/osf.io/w7943.