ADScAI
Applied Data Science & Artificial Intelligence Symposium

Bias Lens: Systemic Bias Detection with Explainable Analysis

Loganathan Bhavaneetharan
Department of Computing
Informatics Institute of Technology
Colombo, Sri Lanka
loganthan.20201212@iit.ac.lk

Saadh Jawwadh
Department of Computing
Informatics Institute of Technology
Colombo, Sri Lanka
saadh.j@iit.ac.lk

***Keywords*—Bias Detection, Explainable AI, Fairness, Multi-label Classification, Natural Language Processing**

I. Introduction

Natural Language Processing (NLP) models have revolutionized the way information is processed, yet they often inherit and even amplify societal biases present in training data [3]. These biases can lead to unfair, discriminatory outcomes in high-stakes applications. Traditional bias detection methods generally rely on binary classification, which oversimplifies the complex and nuanced nature of biased language. In response, we propose **Bias Lens**—a comprehensive framework that combines fine-grained multi-label token classification with a suite of Explainable AI (XAI) methods to not only detect but also elucidate bias in text. Our approach leverages advanced techniques, most notably Integrated Gap Gradients (IG2), to provide detailed neuron-level attributions, while also incorporating complementary methods (LIME, SHAP, DeepLIFT, Attention, Counterfactual, LRP, Occlusion, and Saliency) to create a multifaceted explanation ecosystem. This dual capability enhances transparency and facilitates trust among users such as content moderators and AI auditors.

II. Literature Review

Recent developments in bias detection have shifted focus from coarse document-level approaches to more refined, token-level analysis. The GUS framework [1] introduced a multi-label token classification strategy using BIO tagging to capture overlapping biases, specifically Generalization, Unfairness, and Stereotype. Similarly, the Nbias framework [5] extended these techniques to detect subtler bias instances. Although these methods achieve competitive performance (with macro F1-scores around 0.80 and low Hamming loss), they lack intrinsic interpretability—a critical factor for ensuring model trustworthiness. In parallel, XAI techniques such as Integrated Gradients [4], SHAP, LIME, and DeepLIFT have been applied to explain model decisions. More recent work by Liu *et al.* [2] introduced Integrated Gap Gradients (IG2), which attributes differences in prediction logits to specific internal neurons, thereby providing deeper insight into the model's inner workings. Bias Lens builds upon these foundations by integrating multiple XAI methods to generate comprehensive, actionable explanations.

III. Materials and Methods

A. Bias Detection Model Architecture

At the core of Bias Lens is a fine-tuned BERT-based model (`bert-base-uncased`) configured for token-level bias detection. We frame bias detection as a multi-label Named Entity Recognition (NER) task using the BIO (Begin, Inside, Outside) tagging scheme. Each input sentence is tokenized via WordPiece tokenization; words may be split into sub-tokens, where the first sub-token receives a B-tag (indicating the beginning of a bias span) and subsequent sub-tokens are tagged with I (inside), while tokens outside any bias span are marked with O. Our bias taxonomy encompasses six categories: *Generalization*, *Assumption*, *Exclusion*, *Unfairness*, *Stereotype*, and *Framing* [1]. The model outputs a k-dimensional probability vector for each token using sigmoid activations, enabling multiple bias labels per token. Training is performed with binary cross-entropy loss augmented with focal loss to address class imbalance. Key parameters include a batch size of 16, a learning rate between 3×10^{-5} and 5×10^{-5}, and 15–20 epochs with early stopping based on validation performance.

B. Explainable AI Integration

To complement the classification process, Bias Lens incorporates a diverse array of XAI methods, including: **LIME**, **SHAP**, **Integrated Gradients**, **Integrated Gap Gradients (IG2)**, **DeepLIFT**, **Attention-based visualization**, **Counterfactual analysis**, **Layer-wise Relevance Propagation (LRP)**, **Occlusion**, **Saliency methods**

1) Expanded Discussion of Integrated Gap Gradients (IG2): While prior work [2] briefly introduced IG2, we expand its explanation here. IG2 extends the traditional Integrated Gradients method by focusing on the difference in prediction

https://doi.org/10.31705/ADScAI.2025.17

logits (i.e., the gap) between two bias classes. For each neuron $w_{(l)j}$ in layer l, IG2 computes an attribution score as:

$$\text{IG2}(w_{(l)j}) = w_{(l)j} \times \int_0^1 \frac{\partial \Delta F(\alpha \cdot w_{(l)j})}{\partial w_{(l)j}} \, d\alpha, \qquad (1)$$

where ΔF is the difference in logits between a pair of bias classes. The integral is approximated via a Riemann sum over a set number of steps. This formulation enables a direct mapping from the internal representations (neurons) to bias predictions, effectively highlighting those neurons that have a disproportionate influence on steering the model towards a biased output. For instance, when comparing the logits for *Stereotype* versus *Neutral*, IG2 identifies the specific neurons that drive the model towards labeling text as stereotyped, offering deeper interpretability beyond input-level attributions.

C. Integration and Visualization Pipeline

The outputs from the XAI methods are integrated into the bias detection pipeline. After the model produces token-level bias predictions, the IG2 (along with other XAI methods such as LIME, SHAP, and DeepLIFT) attribution scores are mapped back to the original text. Tokens are color-coded based on their attribution magnitudes, and interactive dashboards display aggregated neuron-level contributions alongside traditional token attributions. For example, a sample visualization (see Fig. 1) illustrates how words like "lazy" and "uneducated" are highlighted for their strong influence on bias predictions. Complementary charts and heatmaps further summarize the contributions from various XAI methods, providing a multifaceted explanation that helps users, including content moderators and auditors, understand both the input features and internal model activations responsible for bias detection.

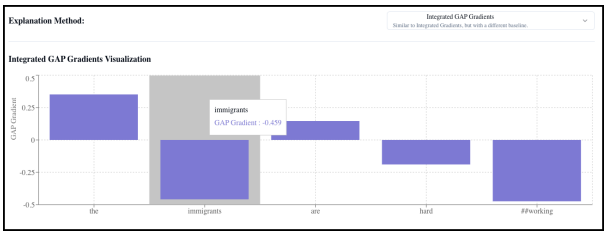

Fig. 1. Illustrative visualization of token-level attributions and IG2-based neuron contributions.

IV. Results and Discussion

Our experiments demonstrate that Bias Lens achieves robust performance, with an overall token-level accuracy of approximately 93% and a macro-averaged F1-score between 0.79 and 0.80. The Hamming loss is maintained at 0.0692. For comparative benchmarking, Table I summarizes the key performance metrics of Bias Lens against two state-of-the-art frameworks.

The results indicate that while Bias Lens and GUS-Net achieve similar macro F1-scores, Bias Lens offers additional benefits in interpretability via its comprehensive XAI integration. In particular, IG2 provides neuron-level insights that complement token-level attributions from methods such as LIME and SHAP. For example, in a test case involving the sentence "Immigrants are lazy and uneducated," the IG2 method clearly identified key neurons that contributed to labeling the sentence as biased (specifically under Unfairness and Stereotype). The accompanying visualizations (see Fig. 1) enable end users to readily understand the basis of the model's predictions. These qualitative findings were further supported by domain expert evaluations, which highlighted improved trust and transparency compared to conventional methods.

TABLE I
Benchmarking Metrics

Model	Hamming Loss	Macro F1	Macro Precision	Macro Recall
Bias-Lens	0.0592	0.80	0.83	0.78
GUS-Net	0.0528	0.81	0.82	0.77
Nbias	0.06	0.68	0.93	0.63

V. Conclusion

Bias Lens advances the field of bias detection in NLP by integrating a robust multi-label token classification model with a comprehensive suite of XAI techniques. Our approach, centered on an extended bias taxonomy and BIO tagging, accurately identifies overlapping bias spans. The expanded discussion of IG2 demonstrates how neuron-level attributions offer deeper interpretability, while complementary XAI methods such as LIME, SHAP, DeepLIFT, and others create a multifaceted explanation framework. Although the quantitative performance of Bias Lens is comparable to that of GUS-Net and Nbias, the enriched transparency and actionable insights provided by our XAI integration significantly enhance model trustworthiness. Future work will focus on extending the bias taxonomy, refining interactive visualization tools, and exploring mitigation strategies (e.g., bias neuron suppression) to further reduce model bias. Ultimately, Bias Lens contributes to the development of more responsible and trustworthy AI systems.

Acknowledgment

The authors gratefully acknowledge the foundational contributions of the GUS framework [1] and the work on Integrated Gap Gradients by Liu *et al.* [2]. We also thank our advisors and domain experts for their valuable feedback.

References

[1] M. Powers *et al.*, "The GUS Framework: Benchmarking Social Bias Classification," 2025.

[2] Y. Liu *et al.*, "The Devil is in the Neurons: Interpreting and Mitigating Social Biases in Pre-Trained Language Models," in Proc. ICLR, 2024.

[3] R. Zhao *et al.*, "On Bias in NLP Models," *Journal of AI Ethics*, vol. 15, no. 2, pp. 123–135, 2023.

[4] A. Manerba and F. Guidotti, "Explainable AI for Bias Detection in NLP," *IEEE Trans. Neural Netw. Learn. Syst.*, 2021.

[5] S. Raza *et al.*, "Advanced Multi-label Bias Detection in NLP," in Proc. ACL, 2024.