Explaining News Bias Detection: A Comparative SHAP Analysis of Transformer Model Decision Mechanisms

Anonymous ACL submission

Abstract

Automated bias detection in news text is heavily used to support journalistic analysis and media accountability, yet little is known about how bias detection models arrive at their decisions or why they fail. In this work, we present a comparative interpretability study of two transformer-based bias detection models: a bias-detector model and a domain-adapted Robustly optimized BERT pretraining approach (RoBERTa) model (DA-RoBERTa- BABE-FT), both fine-tuned on the Bias Annotation for Bias Experiments (BABE) dataset, using SHapley Additive exPlanations (SHAP)-based explanations. We analyze word-level attributions across correct and incorrect predictions to characterize how different model architectures operationalize linguistic bias. Our results show that although both models attend to similar categories of evaluative language, they differ substantially in how these signals are integrated into predictions. The bias detector model assigns stronger internal evidence to false positives than to true positives, indicating a misalignment between attribution strength and prediction correctness and contributing to systematic over-flagging of neutral journalistic content. In contrast, DA-RoBERTa-BABE-FT exhibits attribution patterns that better align with prediction outcomes and produces 63% fewer false positives. We further demonstrate that model errors arise from distinct linguistic mechanisms, with false positives driven by discourse-level ambiguity rather than explicit bias cues. These findings highlight the importance of interpretability-aware evaluation for bias detection systems and suggest that architectural and training choices critically affect both model reliability and deployment suitability in journalistic contexts.

Keywords: news bias detection; model interpretability; SHAP analysis; transformer models; linguistic bias;

1 Introduction

Detecting bias in news text is crucial to maintaining journalistic standards and helping readers assess media credibility. Natural language processing models trained to identify biased language in news articles promise to support these goals, but understanding how they make decisions is essential for trust and effective deployment (Bolukbasi et al., 2016; Caliskan et al., 2017). As automated bias detection tools enter newsrooms and fact-checking workflows, transparency about their decision-making becomes critical (Blodgett et al., 2020).

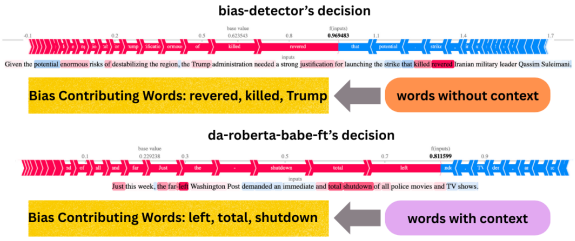

Figure 1: SHAP Reveals the Word Attributions for the model's decision. Red words increase bias, Blue words decrease bias in the sentence. This reveals, how the two models look differently into the sentence to make the decision.

Model interpretability helps validate bias detection systems and diagnose failures. While accuracy metrics show overall performance, interpretability reveals which linguistic signals drive predictions that enable validation against journalistic principles and identification of problematic shortcuts (Lipton, 2017; Doshi-Velez and Kim, 2017). For journalism applications, interpretability is especially important: incorrect flagging can affect credibility, and stakeholders need to understand what constitutes bias from the model's perspective (Rudin, 2019).

We apply SHAP (SHapley Additive exPlanations) to compare how two bias detection mod-

els: bias-detector (Ghosh et al., 2025) and DA-RoBERTa-BABE-FT (Krieger et al., 2022) interpret news text on the BABE dataset (Spinde et al., 2021). By analyzing word-level attributions across true positives, false positives, and true negatives, we investigate: (1) how models identify bias in news articles, (2) what linguistic patterns distinguish correct predictions from errors, and (3) how architectural differences affect interpretability and error patterns. Our analysis reveals: (1) a reversal in attribution strength where false positives show higher SHAP magnitude than true positives in one model, indicating strong but misleading signals; (2) that models conflate emotional or evaluative language with social bias, suggesting limitations in contextual understanding relevant to journalism; and (3) significant differences between architectures in false positive rates and interpretability patterns, with DA-RoBERTa-BABE-FT showing 63% fewer false positives and more intuitive SHAP alignments.

This work contributes: (1) a comparative interpretability framework using SHAP for bias detection models in news text; (2) a word-level attribution analysis revealing how different architectures operationalize bias in journalistic contexts; (3) identification of systematic misalignment between attribution strength and prediction correctness, particularly in false positive errors, and its implications for model reliability; and (4) insights into how bias detection models interpret linguistic signals, informing the development of more reliable systems for journalistic applications.

2 Related Work

2.1 Transformer Models for NLP

Transformer architectures, introduced by Vaswani et al. (2017), have revolutionized NLP through their attention-based mechanism. Devlin et al. (2019) developed BERT, demonstrating the power of bidirectional pre-training for language understanding tasks. Liu et al. (2019) introduced RoBERTa, a robustly optimized BERT variant that improved performance through better training procedures. These transformer-based models have become the foundation for many bias detection systems, including the Bias-Detector and DA-RoBERTa-BABE-FT model analyzed in this work. Understanding how these architectures process and interpret linguistic signals is essential for developing more reliable bias detection systems.

2.2 Bias Detection in News

Automated bias detection in news text has gained attention with the rise of computational journalism. Spinde et al. (2021) introduced BABE (Bias Analysis by Experts), a sentence-level dataset for bias detection in news articles using distant supervision from expert annotations. Their work demonstrated the feasibility of neural models for identifying biased language in journalistic contexts. Building on this foundation, Ghosh et al. (2025) developed the bias-detector model, which we analyze in this work, focusing specifically on detecting bias in news articles. Krieger et al. (2022) advanced the field by proposing domain-adaptive pre-training approaches for language bias detection in news, showing that specialized pre-training improves performance on media bias tasks. Beyond news-specific applications, research on bias detection has expanded to measure and identify various forms of social bias in language models. Nadeem et al. (2021) introduced StereoSet, a benchmark for measuring stereotypical bias in pretrained language models, while Nangia et al. (2020) created CrowS-Pairs, a challenge dataset for measuring social biases in masked language models. Powers et al. (2025) proposed the GUS Framework for benchmarking social bias classification using both discriminative and generative language models, providing a comprehensive evaluation framework. Hutchinson et al. (2020) examined social biases in NLP models as barriers for persons with disabilities, highlighting the importance of inclusive bias detection. Borkan et al. (2019) developed nuanced metrics for measuring unintended bias with real data for text classification, while Röttger et al. (2021) introduced HateCheck, a functional test suite for hate speech detection models. These works collectively establish the landscape of bias detection and measurement in NLP, with our work focusing specifically on interpretability of bias detection models in news text.

2.3 Interpretability Methods in NLP

Understanding how NLP models make decisions has become increasingly important as these systems are deployed in high-stakes applications. Interpretability methods can be broadly categorized into model-agnostic approaches that work with any model and model-specific approaches that leverage internal model representations. Ribeiro et al. (2016) introduced LIME (Local Interpretable

Model-agnostic Explanations), a model-agnostic method that explains individual predictions by approximating the model locally with interpretable models. Lundberg and Lee (2017) proposed SHAP (SHapley Additive exPlanations), which unifies multiple explanation methods under a unified framework based on Shapley values from cooperative game theory, providing theoretically grounded feature attributions. The interpretability of attention mechanisms in transformer models has been particularly debated. Jain and Wallace (2019) argued that attention weights do not necessarily provide faithful explanations, challenging the common assumption that attention can be directly interpreted. Wiegreffe and Pinter (2019) responded by demonstrating cases where attention can indeed provide meaningful explanations, while Serrano and Smith (2019) provided a more nuanced analysis of when attention can be considered interpretable. These debates highlight the complexity of interpreting neural models and the need for rigorous evaluation of interpretability methods. DeYoung et al. (2020) introduced ERASER, a benchmark for evaluating rationalized NLP models, providing standardized metrics for assessing explanation quality. In this work, we employ SHAP as our primary interpretability method, leveraging its theoretical foundations and model-agnostic nature to compare different bias detection architectures.

2.4 SHAP Applications in NLP

SHAP has been widely adopted for explaining NLP models due to its theoretical grounding and ability to provide consistent, locally accurate explanations. Lundberg et al. (2019) extended SHAP to tree-based models, demonstrating how local explanations can be aggregated to achieve global understanding. Mosca et al. (2022) provided a comprehensive review of SHAP-based explanation methods specifically for NLP interpretability, discussing applications across various NLP tasks and highlighting best practices. Covert et al. (2022) developed a unified framework for model explanation based on the principle of explaining by removing features, which aligns with SHAP's approach of measuring feature contributions through marginalization. Despite the widespread use of SHAP in NLP, relatively little work has applied it to understand bias detection models specifically, particularly in comparative settings. Most existing applications focus on single-model explanations or task-specific interpretability. Our work extends this line of research by applying SHAP to conduct a comparative analysis of two different bias detection architectures, revealing how model design choices affect interpretability patterns and error modes in news bias detection.

2.5 Model Comparison Methods

Dietterich (1998) proposed statistical tests for comparing supervised classification algorithms, emphasizing the importance of accounting for variance in performance estimates. Dror et al. (2018) provided a comprehensive guide to testing statistical significance in natural language processing, highlighting common pitfalls and best practices in model comparison. While these frameworks are important for rigorous evaluation, our work focuses on interpretability-based comparison rather than purely statistical comparisons, examining how different models interpret the same inputs rather than just comparing their performance metrics.

3 Methodology

3.1 Dataset and Models

We evaluate two bias detection models on the BABE (Bias Analysis by Experts) dataset (Spinde et al., 2021), a sentence-level dataset for bias detection in news articles. BABE contains expert-annotated sentences from news sources, labeled as biased or non-biased, making it suitable for analyzing how models interpret linguistic signals in journalistic contexts.

Models: We analyze two pre-trained state-of-the-art bias detection models: (1) bias-detector[1] (Ghosh et al., 2025), a model specifically designed for detecting bias in news text, and (2) DA-RoBERTa-BABE-FT[2], a domain-adaptive RoBERTa model (Liu et al., 2019) fine-tuned on the BABE dataset (Krieger et al., 2022). Both models are transformer-based architectures (Vaswani et al., 2017; Devlin et al., 2019) that classify sentences as biased or non-biased. We use the test split of BABE for evaluation, ensuring that our analysis reflects model behavior on held-out data.

Preprocessing: For each model, we use the corresponding tokenizer from the HuggingFace Transformers library. Input sequences are tokenized with a maximum length of 256 tokens, following standard practices for sentence classification tasks. La-

[1] https://huggingface.co/himel7/bias-detector
[2] https://huggingface.co/mediabiasgroup/DA-RoBERTa-BABE-FT

bels are normalized to binary format (0 for non-biased, 1 for biased) to ensure consistent analysis across models.

3.2 SHAP-Based Interpretability Framework

We employ SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017) to explain model predictions. SHAP is based on Shapley values from cooperative game theory (Shapley, 1953), which provide a theoretically grounded method for attributing prediction contributions to individual features.

Theoretical Foundation: Shapley values distribute the total prediction value among features according to their marginal contributions across all possible feature coalitions. For a model prediction $f(x)$ on input x with features F, the Shapley value ϕ_i for feature i is:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)] \tag{1}$$

where S represents a subset of features, and $f(S)$ denotes the model output when only features in S are present. This formulation ensures that attributions satisfy desirable properties: efficiency (attributions sum to the prediction difference from baseline), symmetry (symmetric features receive equal attributions), dummy (features with no effect receive zero attribution), and additivity (attributions are additive across model outputs).

SHAP Implementation: We use the SHAP library's TextExplainer with a Text masker, which handles tokenization and masking for transformer models. For each input sentence, SHAP generates token-level attributions by computing Shapley values for each token's contribution to the model's prediction. The explainer uses a background dataset (a subset of the test set) to establish baseline expectations, and computes attributions by marginalizing over token presence/absence.

Prediction Function: For each model, we construct a prediction function that takes a list of text inputs and returns class probabilities. The function tokenizes inputs, passes them through the model, and applies softmax to obtain probability distributions over classes. SHAP then explains these probability outputs, providing attributions for the positive (biased) class probability, which we use as our primary analysis target.

3.3 Word-Level Attribution Aggregation

Transformer-based models operate on subword units, while bias analysis is more interpretable at the word level. To bridge this gap, we aggregate token-level SHAP attributions into word-level explanations using a structured aggregation pipeline, illustrated in Figure 2.

Figure 2: Word-level SHAP aggregation pipeline. Token-level SHAP values produced by a transformer model are grouped into words based on tokenizer boundaries and text alignment, then aggregated to produce interpretable word-level attributions.

Given an input sentence, we first compute token-level SHAP values using the model-specific tokenizer and SHAP TextExplainer. Subword tokens are then grouped into words based on tokenizer boundary markers and original text alignment. For each word, we aggregate the SHAP values of all constituent tokens by summation, preserving the additive property of Shapley values. This produces a single attribution score per word, reflecting its overall contribution to the model's predicted probability of the biased class.

To ensure linguistic consistency, we apply normalization steps that correct tokenization artifacts and align word boundaries with the original text. These steps prevent spurious attributions arising from subword fragmentation or punctuation handling. The resulting word-level attributions enable consistent comparison across instances and models while retaining faithfulness to the underlying token-level explanations.

3.4 Stratified Sampling Strategy

To ensure comprehensive analysis across different prediction types, we employ stratified sampling to select instances for SHAP analysis. For each model, we first evaluate all test instances to obtain predictions and ground truth labels. We then categorize instances into: (1) True Positives (TP): correctly predicted as biased, (2) False Positives (FP): incorrectly predicted as biased, (3) True Negatives (TN): correctly predicted as non-biased, and (4) False Negatives (FN): incorrectly predicted as non-biased.

Sampling Procedure: We sample up to 100 instances from each category (TP, FP, TN) to en-

sure balanced representation. For the bias-detector model, we analyze 237 instances total (100 TP, 37 FP, 100 TN), while for DA-RoBERTa-BABE-FT, we analyze 212 instances (100 TP, 12 FP, 100 TN). The sampling is performed with a fixed random seed to ensure reproducibility. This stratified approach helps us to compare SHAP patterns across prediction categories and identify systematic differences between correct and incorrect predictions.

3.5 Analysis Framework

Per-Instance Analysis: For each sampled instance, we compute word-level SHAP attributions and store: (1) the original text, (2) ground truth label, (3) predicted label, (4) predicted probability for the biased class, (5) token-level SHAP values, and (6) aggregated word-level attributions with word frequencies.

Global Word Importance: We aggregate word-level attributions across all instances to compute global word importance metrics. For each unique word, we compute: (1) mean absolute SHAP value across all occurrences, (2) total occurrence count, and (3) mean signed SHAP value (indicating whether the word generally increases or decreases bias probability). This global analysis reveals which words are most influential across the dataset and how they contribute to model predictions.

Category-Specific Analysis: We analyze SHAP patterns separately for TP, FP, and TN instances to understand how model interpretations differ between correct and incorrect predictions. For each category, we compute: (1) mean absolute SHAP value per instance (indicating overall attribution magnitude), (2) top words by mean absolute SHAP value, (3) distribution of positive vs. negative SHAP values, and (4) word frequency patterns. This category-specific analysis enables identification of patterns that distinguish correct predictions from errors.

Comparative Analysis: We compare SHAP patterns between the two models to understand how architectural differences affect interpretability. We examine: (1) differences in top bias indicators, (2) variations in false positive patterns, (3) differences in SHAP magnitude distributions across prediction categories, and (4) model-specific word interpretation patterns.

3.6 Statistical Significance Testing

To assess whether observed differences between models are statistically significant, we employ McNemar's test (Dietterich, 1998). McNemar's test is appropriate for comparing two classifiers on the same test set, as it accounts for the paired nature of predictions (both models make predictions on identical instances).

McNemar's Test: We construct a 2×2 contingency table where: (a) both models are correct, (b) Model 1 is wrong and Model 2 is correct, (c) Model 1 is correct and Model 2 is wrong, and (d) both models are wrong. The test statistic is computed as:

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c} \tag{2}$$

with 1 degree of freedom, where b and c represent the off-diagonal counts. We apply continuity correction for small sample sizes. The null hypothesis is that the two models have the same error rate (i.e., $b = c$). A significant result indicates that the models differ in their prediction patterns, which may reflect differences in how they interpret linguistic signals.

3.7 Evaluation Metrics

We report standard classification metrics: accuracy, precision, recall, and F1 score (both binary and macro-averaged). Additionally, we analyze SHAP-specific metrics: (1) mean absolute SHAP value per instance category (TP, FP, TN), (2) distribution of SHAP values across words, and (3) word-level attribution patterns. These metrics provide both performance and interpretability perspectives on model behavior.

4 Evaluation Results

4.1 Overall Model Performance

Table 1 summarizes the performance metrics for both models. While bias-detector achieved higher accuracy and F1, DA-RoBERTa-BABE-FT showed better precision. The significant difference in false positive rates (15.6% vs. 5.7%) suggests that architectural differences and training approaches substantially affect precision in bias detection tasks.

4.2 SHAP Magnitude Patterns

Word-level SHAP attribution magnitudes reveal differences in how models interpret linguistic signals. Fig. 3 presents mean absolute SHAP values by prediction category for both models.

Table 1: Model Performance on BABE Test Set

Metric	bias-detector	da-roberta-babe-ft
Accuracy	0.8520	0.8230
Precision	0.9237	0.9704
Recall	0.8014	0.7048
Binary F1	0.8582	0.8166
Macro F1	**0.8517**	0.8228
Weighted F1	0.8525	0.8221

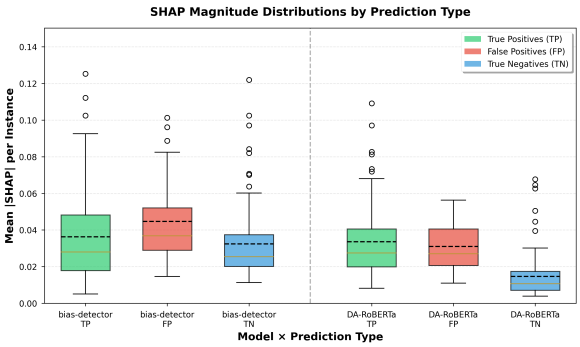

Figure 3: SHAP Magnitude Distributions: False positives in bias-detector exhibit higher attribution magnitude than true positives, while DA-RoBERTa-BABE-FT shows the opposite, indicating better alignment between explanations and predictions.

This pattern of bias-detector indicates that incorrect predictions (false positives) exhibit stronger individual word signals than correct predictions, suggesting a reversal in attribution strength between false and true positives where the model relies on strong but misleading linguistic cues. However, DA-RoBERTa-BABE-FT model exhibited the opposite pattern where correct predictions show stronger signals than errors suggesting that its SHAP attributions better align with actual model decision-making. DA-RoBERTa-BABE-FT showed lower confidence on false positives, indicating less overconfidence on incorrect predictions. This suggests DA-RoBERTa-BABE-FT is less overconfident on incorrect predictions.

4.3 Global Bias Indicators

We aggregated word-level SHAP attributions across all instances to identify global bias indicators. Fig. 4 presents the top 10 words with highest mean absolute SHAP values for each model.

For bias-detector, Top indicators include emotional and evaluative language: "boasted", "bloated", "dubious", "antisemitic", and "skeptics". Many top words are negative descriptors ("bloated", "dubious") or explicit bias terms ("antisemitic"),

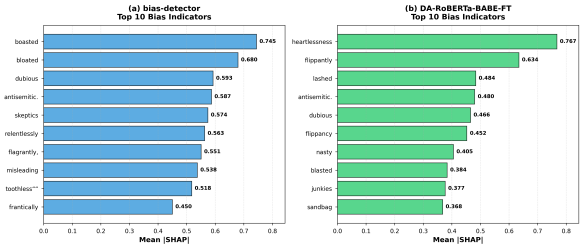

Figure 4: Top Bias Indicators for the bias-detector model.

suggesting the model relies heavily on loaded language and explicit bias markers. While for DA-RoBERTa-BABE-FT, top indicators emphasize dismissive and moral language: "heartlessness", "flippantly", "lashed", "antisemitic", and "dubious". This model shows a stronger focus on dismissive language patterns ("flippantly", "heartlessness") and emotional descriptors compared to the original model. Some overlap exists ("dubious", "antisemitic", "lashed"), suggesting consensus on certain linguistic bias markers.

4.4 Word-Level Patterns

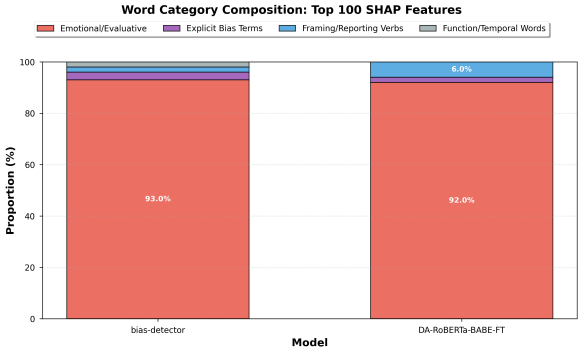

Figure 5: Word category composition of the top 100 SHAP-attributed features for each model.

Figure 5 presents the lexical category composition of the top 100 most influential SHAP features for both models. For both the bias-detector and DA-RoBERTa-BABE-FT, evaluative and emotionally charged language dominates the highest-attribution features, reflecting the fact that bias across political, social, and cultural domains is often linguistically realized through subjective or judgment-laden expressions. At this aggregate level, the two models appear superficially similar in terms of the lexical categories they attend to.

However, this similarity conceals important differences in how such features are operationalized during prediction. As shown by earlier attribu-

tion magnitude analyses, the bias-detector tends to treat evaluative language as sufficient evidence for bias, often producing high-confidence predictions even in neutral reporting contexts. In contrast, DA-RoBERTa-BABE-FT assigns relatively greater importance to framing and reporting verbs, suggesting that evaluative expressions are more frequently interpreted in relation to attribution, quotation, or narrative context rather than as standalone bias indicators.

Figure 6: Word category composition of the top 100 SHAP-attributed tokens for true positive and false positive predictions, separated by model.

Figure 6 further disambiguates these behaviors by conditioning the analysis on prediction correctness. For the bias-detector, the category distribution remains nearly unchanged between true positives and false positives, with evaluative language dominating both. This invariance indicates that the model relies on similar lexical cues regardless of whether its prediction is correct, explaining its tendency toward systematic false positives driven by emotional tone rather than contextual bias.

By contrast, DA-RoBERTa-BABE-FT exhibits a pronounced shift between correct and incorrect predictions. While true positives are primarily driven by evaluative language, false positives show substantially increased reliance on function and temporal words, alongside weaker evaluative signals. This pattern suggests that DA-RoBERTa-BABE-FT's errors arise in structurally or discourse-level ambiguous cases where explicit semantic bias cues are limited, rather than from indiscriminate overgeneralization of emotional language. Taken together, these findings indicate that although both models attend to similar lexical categories, they differ fundamentally in how those features are integrated into bias judgments.

4.5 Statistical Significance Testing

We applied McNemar's test to assess whether the observed difference in error patterns between the two models is statistically significant. On the test set, we constructed a 2×2 contingency table comparing correct and incorrect predictions for both models. McNemar's test yielded $\chi^2 = 5.723$ with $p = 0.0167$, indicating a statistically significant difference in model errors. This result suggests that the difference in false positive rates (15.6% vs. 5.7%) is unlikely to be due to chance, supporting the conclusion that DA-RoBERTa-BABE-FT achieves higher precision, while the bias-detector attains higher overall F1 score.

5 Conclusion

In this work, we conducted a comparative interpretability analysis of two transformer-based bias detection models for news text using SHAP-based explanations. While both models achieve competitive classification performance, the bias-detector assigns stronger internal evidence to false positive predictions than to correct detections, indicating a misalignment between attribution strength and prediction correctness. Such behavior leads to systematic over-flagging of neutral journalistic content, raising concerns for deployment in real-world editorial settings.

By contrast, DA-RoBERTa-BABE-FT produces significantly fewer false positives and exhibits attribution patterns that align more closely with prediction outcomes, with stronger evidence supporting correct detections and weaker, more distributed signals associated with errors. Although both models rely heavily on evaluative language, they operationalize bias differently: the bias-detector emphasizes explicit negatively valenced descriptors, whereas DA-RoBERTa-BABE-FT more consistently contextualizes evaluative expressions through framing and attribution patterns.

Overall, our results highlight the importance of interpretability-aware evaluation when developing and deploying bias detection models for journalistic applications. Beyond raw accuracy, understanding how and why models make decisions is essential for ensuring responsible use, minimizing false positives, and maintaining trust in automated media analysis tools. Future work should explore context-sensitive modeling approaches and calibration-aware training strategies to further improve the robustness of bias detection systems.

6 Limitations

This study has several limitations that should be considered when interpreting the results. First, our analysis focuses on two transformer-based models evaluated on a single dataset (BABE). While BABE is an expert-annotated benchmark for sentence-level news bias detection, it reflects specific annotation guidelines and normative assumptions about bias. Consequently, the observed attribution patterns and failure modes may not fully generalize to other datasets, domains, or cultural definitions of bias.

Second, our interpretability analysis relies on SHAP as a post-hoc explanation method. Although SHAP provides theoretically grounded, additive attributions, it approximates feature contributions under independence assumptions that may not fully capture complex interactions in transformer architectures. As a result, word-level attributions should be interpreted as explanatory signals rather than definitive causal explanations of model behavior.

Finally, while we compare attribution patterns across correct and incorrect predictions, we do not evaluate alternative explanation methods or formally assess explanation faithfulness beyond consistency with model behavior. Future work could extend this analysis by incorporating multiple interpretability techniques, calibration metrics, or human evaluation of explanations to further validate reliability and usability.

7 Ethical Considerations

This work analyzes bias detection models and their explanations, but does not propose automated systems as substitutes for human judgment. Bias in journalism is inherently contextual and normatively defined, and model predictions along with their explanations may reflect annotation guidelines, cultural assumptions, or training data biases present in the BABE dataset. As such, outputs from these models should not be treated as objective or authoritative assessments of bias.

Our analysis highlights the risk of false positives, which may incorrectly label neutral reporting as biased and potentially undermine journalistic credibility if deployed without oversight. We therefore emphasize that bias detection models should be used as assistive tools within human-in-the-loop workflows, where explanations can support critical inspection rather than automated decision-making.

Finally, while interpretability techniques such as SHAP improve transparency, they do not guarantee correctness or fairness. Explanations should be understood as diagnostic aids for model behavior, not as justifications for deploying models in high-stakes or adversarial settings without careful evaluation and governance.

References

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Preprint*, arXiv:1607.06520.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. *Preprint*, arXiv:1903.04561.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Ian Covert, Scott Lundberg, and Su-In Lee. 2022. Explaining by removing: A unified framework for model explanation. *Preprint*, arXiv:2011.14878.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. Eraser: A benchmark to evaluate rationalized nlp models. *Preprint*, arXiv:1911.03429.

Thomas G. Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *Preprint*, arXiv:1702.08608.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing.

In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Himel Ghosh, Ahmed Mosharafa, and Georg Groh. 2025. To bias or not to bias: Detecting bias in news with bias-detector. *Preprint*, arXiv:2505.13010.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. *Preprint*, arXiv:1902.10186.

Jan-David Krieger, Timo Spinde, Terry Ruas, Juhi Kulshrestha, and Bela Gipp. 2022. A domain-adaptive pre-training approach for language bias detection in news. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, JCDL '22. ACM.

Zachary C. Lipton. 2017. The mythos of model interpretability. *Preprint*, arXiv:1606.03490.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Preprint*, arXiv:1705.07874.

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2019. Explainable ai for trees: From local explanations to global understanding. *Preprint*, arXiv:1905.04610.

Edoardo Mosca, Ferenc Szigeti, Stella Tragianni, Daniel Gallagher, and George Louis Groh. 2022. Shap-based explanation methods: A review for nlp interpretability. In *International Conference on Computational Linguistics*.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Maximus Powers, Shaina Raza, Alex Chang, Umang Mavani, Harshitha Reddy Jonala, Ansh Tiwari, and Hua Wei. 2025. The gus framework: Benchmarking social bias classification with discriminative (encoder-only) and generative (decoder-only) language models. *Preprint*, arXiv:2410.08388.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. *Preprint*, arXiv:1602.04938.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Preprint*, arXiv:1811.10154.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. Hatecheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? *Preprint*, arXiv:1906.03731.

Lloyd S Shapley. 1953. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker, editors, *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press, Princeton.

Timo Spinde, Hendrik Westermann, and Georg Rehm. 2021. Neural media bias detection using distant supervision with babe. In *Proceedings of the 2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 3518–3530.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. *Preprint*, arXiv:1908.04626.

A Appendix: Word-Level Attribution Aggregation in Detail

Transformer models operate on subword tokens (e.g., BPE tokens in RoBERTa), but interpretability for bias detection benefits from word-level analysis. We aggregate token-level SHAP values to word-level attributions through a multi-step process.

Token-to-Word Mapping: We identify word boundaries using tokenization markers. RoBERTa-style tokenizers use special markers (e.g., Ġ or _)

to indicate tokens that begin new words. We detect these markers and group consecutive subword tokens that belong to the same word. For tokens without explicit markers, we use whitespace information from the original text to determine word boundaries.

Word Boundary Detection: To handle cases where tokens may be incorrectly merged (e.g., "tuesdayquoting" or "nationalismfueled"), we implement heuristic word boundary detection. We split tokens when: (1) a lowercase letter is followed by an uppercase letter (e.g., "tuesdayQuoting" → ["tuesday", "Quoting"]), (2) common word endings (e.g., "day", "ing", "ed", "ly") are followed by lowercase letters that could start new words, or (3) punctuation or hyphens indicate word separation.

Attribution Aggregation: For each word, we sum the SHAP values of all tokens that comprise that word. If a word is split across multiple tokens, the word-level attribution is the sum of all constituent token attributions. This aggregation preserves the additive property of SHAP values while providing interpretable word-level explanations.

Normalization: We apply normalization to handle tokenization artifacts. Merged words (e.g., "dmnboasted" → "boasted") are split, and punctuation artifacts are stripped from word boundaries. This ensures that our word-level analysis reflects actual linguistic patterns rather than tokenization idiosyncrasies.

B Appendix: Token-Level SHAP Examples

B.0.1 True Positives

Refer to Fig. 7 below.

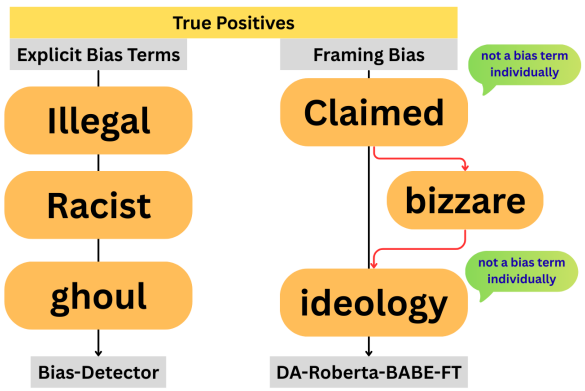

Figure 7: Examples of Biased Terms detected by both the models for the True Positives.

B.0.2 False Positives

Bias-Detector: False positives exhibit the highest SHAP magnitude (0.0354) despite being incorrect predictions. Top words include: "illegal" (0.1400, positive), "said" (0.1304, negative), "tuesday" (0.1128, negative), and "could" (0.1094, negative). Crucially, "illegal" appears in both true positives (0.3489) and false positives (0.1400), indicating context-dependent interpretation where the same word can signal bias or not depending on surrounding context. Many top false positive words have negative SHAP values (indicating they should decrease bias probability), yet the model predicts bias with high confidence (0.796). This suggests non-linear interactions or missing contextual features that word-level SHAP cannot capture. Function and temporal words dominate false positives, with intensifiers ("highly": 0.0624, "very": 0.0599) showing positive signals, indicating the model may conflate emotional intensity with bias.

DA-RoBERTa-BABE-FT: False positives show lower SHAP magnitude (0.0215) and lower confidence (0.673), indicating more reasonable error patterns. Top words include: "claims" (0.2061, positive), function words ("the", "and", "of"), and "abortion" (0.0186). Notably, "claims" appears in both true positives (as "claimed", "claiming", "claim") and false positives, suggesting the same framing language triggers correct and incorrect predictions depending on context. This context-sensitivity is a limitation of both models, though DA-RoBERTa-BABE-FT shows fewer instances.

B.0.3 True Negatives

Refer to Fig. 8 below.

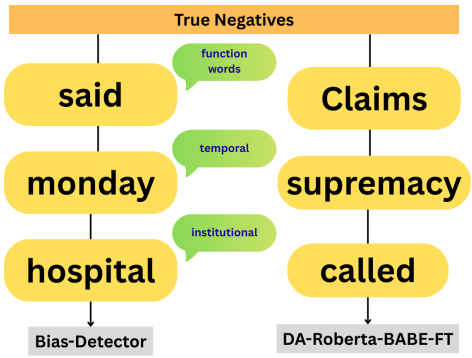

Figure 8: Examples of Biased Terms detected by both the models for the True Negatives.