LLM-based Adversarial Dataset Augmentation for Automatic Media Bias Detection

Martin Wessel
CDTM - Technical University of Munich
Munich, Germany
m.wessel@media-bias-research.org

Abstract

This study presents BiasAdapt, a novel data augmentation strategy designed to enhance the robustness of automatic media bias detection models. Leveraging the BABE dataset, BiasAdapt uses a generative language model to identify bias-indicative keywords and replace them with alternatives from opposing categories, thus creating adversarial examples that preserve the original bias labels. The contributions of this work are twofold: it proposes a scalable method for augmenting bias datasets with adversarial examples while preserving labels, and it publicly releases an augmented adversarial media bias dataset. Training on BiasAdapt reduces the reliance on spurious cues in four of the six evaluated media bias categories.

1 Introduction

Automatic media bias detection has gained significant attention with more capable language models. Systems that automatically detect media bias can help media consumers better identify slanted reporting, help journalists uncover overlooked biases, and help researchers evaluate the reporting landscape (Hamborg et al., 2019; Spinde et al., 2021). However, existing models often rely on spurious cues for classification decisions, which can lead to a superficial understanding of bias and compromise their generalization capabilities and objectivity (Wessel and Horych, 2024). Data augmentation techniques can mitigate the reliance on such shortcuts (Wang et al., 2023). Training data for systems that automatically detect media bias originates predominantly from small, manually labeled datasets (Wessel et al., 2023) with associated high labeling costs (Hamborg, 2020; Spinde et al., 2021). Classical data augmentation techniques would require manual relabeling for every augmented sentence, as, for instance, random swaps of words or deletions could alter the bias of a sentence. To minimize the high manual relabeling costs, an adaptation is required. BiasAdapt, a process designed to enhance the robustness of automatic media bias detection systems, aims to address this.[1] BiasAdapt identifies keywords associated with predefined categories such as gender, origin, or political affiliation. It then generates and replaces alternative words from opposing subcategories. In this study, this adversarial augmentation process is performed on the BABE dataset (Spinde et al., 2021). The augmented data serves as training data, reducing reliance on spurious cues in four of the six evaluated media bias categories. However, these modifications also affect classification performance in some categories, requiring further investigation.

The process of augmenting an existing data set with adversarial data using LLMs is transferable to domains beyond the detection of media bias. It allows for label-preserving alterations of predefined dimensions with accurate content exchanges that require an in-depth understanding of the sentence.

2 Related Work

Media bias, a phenomenon where the information presented in the media is skewed, has been the subject of significant research (Hamborg et al., 2019; Baumer et al., 2015; Spinde et al., 2023). Advances in bias detection, mainly through transformer-based methodologies, have notably improved classification accuracy (Spinde et al., 2021, 2023).

Despite these advancements, a persistent challenge is the dependence on small, narrowly focused, manually annotated datasets (Wessel et al., 2023). This limitation often results in models that overfit and generalize poorly. Recent work by Wessel and Horych (2024) highlights that transformer-based models in automatic

[1] The dataset and code are publicly available under https://github.com/martinpwessel/BiasAdapt-Repository.

Proceedings of LaTeCH-CLfL 2025, pages 19–24

media bias detection predominantly target highly connotative words and do not grasp the nuance of context. This leads to reliance on unreliable indicators or spurious cues for classification decisions, manifesting itself as inconsistent bias determinations under stress tests. Spurious cues in this context are superficial lexical features, such as demographic terms or political affiliations, that bias detection models incorrectly rely on to classify bias instead of analyzing the actual linguistic and contextual indicators of bias.

Wessel and Horych (2024) introduce a CheckList-based invariance test (INV) (Ribeiro et al., 2020) to assess the resilience of bias detection models to irrelevant input alterations. They define seven bias categories -gender, origin, religion, political affiliation, occupation, politician names, and disability- based on prior literature and practical observations of bias-related word associations. Their CheckList-based invariance test systematically examines whether altering terms within these categories (e.g., replacing a male-associated name with a female-associated one) changes the model's classification. If the model's bias determination fluctuates despite maintaining sentence semantics, it suggests reliance on spurious cues rather than true contextual understanding. Wessel and Horych (2024) report significant disparities in model behavior across datasets. For example, words linked to gender or origin frequently influence bias predictions, implying that classifiers are using these cues instead of analyzing how bias is actually expressed. Such findings emphasize the necessity of model refinement for more robust detection methods.

Wang et al. (2023) propose adversarial training and data augmentation to enhance model robustness. Jia and Liang (2017) showcase the utility of adversarial examples in evaluating and enhancing the robustness of natural language processing models, a key consideration in detecting and mitigating media bias. Additionally, Shafahi et al. (2019) highlight the significance of adversarial data augmentation in addressing the subtleties of language, suggesting its essential role in refining models tasked with understanding nuanced biases.

This study refines media bias detection through adversarial data augmentation, addressing the limitations of existing methods. Techniques like frequency-guided word substitution (FGWS) (Mozes et al., 2021) and adversarial text modifications (Samanta and Mehta, 2017) often fail to preserve bias labels, requiring costly human re-annotation (Sabou et al., 2012) when biases are unintentionally altered. When, for instance, words are randomly added or deleted, a previously unbiased sentence might now be biased. The strategy proposed in this study offers key improvements:

- Label Preservation: Maintains label integrity, reducing the need for manual re-labeling (Zhang and Wallace, 2015).
- Contextual Sensitivity: Ensures coherent augmentations by considering keyword context, which prevents misplaced examples (Wei and Zou, 2019).
- Bias Specificity: Targets bias mitigation, avoiding reinforcement of existing biases (Dixon et al., 2018).

3 Methodology

The BiasAdapt augmentation process expands the dataset to improve bias detection within text-based content. The process begins with an existing annotated dataset. In this case, the BABE (Spinde et al., 2021) data set consists of sentences that are binary labeled for bias. The next step identifies keywords within each sentence by predefined categories. A keyword is any word that can clearly be attributed to one category. For instance, for gender, every gender-associated word is a keyword; for religion, every religion-associated word, and so on. For the context of media bias, these categories are gender, origin, religion, political affiliation, occupation, and politician names as defined by Wessel and Horych (2024).[2] As these categories need to be predefined before the annotation, prior knowledge of where spurious cues may arise in the specific context is necessary. BiasAdapt identifies keywords by individually querying each sentence to a generative language model. For all prompts, GPT-3.5 Turbo (Brown et al., 2020) is used. The language model returns the identified keywords and the associated category (gender, origin, etc.). Once more, these words are queried using the same language model with instructions to generate alternative words for each keyword. The process queries the same language model again, instructing it to generate alternative words for each keyword. These alternatives

[2] Wessel and Horych (2024) also include the category disability. This category was excluded from this analysis because the BABE data set contains only a few words associated with disability, leaving too few permutations for meaningful effects.

must come from opposing categories, ensuring they are associated with, for instance, an opposite political affiliation, gender, or a different religion.
The alternative words then substitute the original terms in the sentence to create new instances, each maintaining the initial bias label. The bias label remains unchanged because the substituted keywords belong to the same predefined category, ensuring that the sentence's bias, whether introduced through framing or word choice, is preserved. Bias can arise from how a sentence is structured but also from the connotations of specific words. For example, replacing 'he' with 'she' in 'He lacks the toughness for leadership' retains gender bias because the stereotype about leadership remains intact. Similarly, swapping 'Christian' with 'Muslim' in 'Policy unfairly favors Christian values' maintains religious bias by preserving the critical framing of the sentence. In political contexts, replacing 'left-wing politician' with 'right-wing politician' in a sentence about corruption does not alter the underlying bias, as the negative framing remains the same. Likewise, in occupation-based bias, exchanging 'artist' with 'construction worker' in 'Artists contribute little to the economy' preserves bias against certain professions. Since these substitutions maintain the same bias patterns, the augmentation process ensures that the dataset's labels remain consistent. This only works for predefined bias categories with predefined opposing subcategories that substitutions can be taken from. In some cases, substitutions may interact with the sentence structure in ways that subtly alter the perceived bias. For example, in 'She is caring and nurturing,' substituting 'she' with 'he' could challenge the stereotype that these traits are inherently feminine, as men are less commonly associated with these characteristics in traditional gender roles. This demonstrates that substitutions in certain contexts may shift or reinforce bias depending on the societal associations linked to the words involved. While the augmentation process follows strict category-based substitutions, potential context-dependent bias shifts are a limitation of this method.

Figure 1 displays the augmentation process with an example sentence from the BABE dataset.
Each sentence may contain multiple identified keywords, each with a list of alternative words, resulting in numerous possible permutations. When substituting these keywords, the rest of the sentence and its label remain unchanged. That is why generating too many permutations can lead to overfitting when the data is used for model training. For this study, three permutations per original sentence are found to be the best trade-off between introducing adversarial examples and the prevention of overfitting. However, this might vary depending on the dataset size, sentence complexity, and length. The three permutations are chosen by randomly sampling alternative words from the word lists. The training setup ensures no data leakage between the training, test, and validation data, as original and altered sentences are always in the same set.

This process creates an Adversarial BABE dataset, which is then used to train a language model to automatically detect media bias. Its detection capabilities are compared to that of a model trained solely on BABE. The performance of the models is evaluated using the test sets from Wessel and Horych (2024). The test set consists of 1,900 binary-labeled sentences distributed over the categories (50% of which are classified as biased). Within each category, variance serves as a metric for spurious cues: Higher values suggest that the model relies on shortcuts rather than general language understanding. For example, if the model does not use gender as a factor in classification, accuracy should remain consistent across sentences containing male, female, and non-binary keywords. Both models are based on a pre-trained RoBERTa model to ensure comparability with Spinde et al. (2021). The model training ends based on an early stopping criteria.

4 Results

The augmentation of the BABE dataset through BiasAdapt significantly increases the dataset size to 14,659 entries, adding 10,986 entries to the original collection. Not for every original sentence permutations can be constructed, as not all sentences contain words that are identified as keywords being associated with one of the predefined categories. While the distribution within categories remains equal, the occurrence of relevant keywords differs between categories depending on their occurrence in the BABE dataset. In the initial step, a total of 4,906 keywords are identified and replaced. The most frequently modified categories are gender (1,469 identified keywords) and politician names (1,232), followed by origin (609), political affiliation (464), religion (97), and occupation (35). This distribution is primarily influenced by the topic choices of the BABE dataset. This study's eval-

Figure 1: Examplary augmentation process using BiasAdapt. The sentence is biased because the phrase "dug his heels in" conveys a negative subjective judgment about the politician's stance.

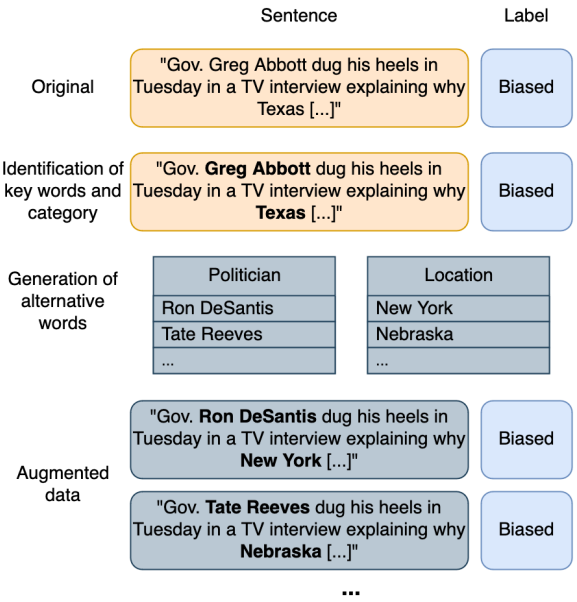

Table 1: The detection results (F1-Scores) on the INV test set by subcategories. Variance values are shown in brackets behind the average scores.

Category	Subcategory	Augmented BABE	BABE
Gender	Male	0.54	0.68
	Female	0.54	0.75
	Non-binary	0.54	0.69
	Average	0.54 (3.0e-6)	0.71 (0.001)
Origin	European	0.92	0.94
	African	0.94	0.99
	Asian	1.00	1.00
	Average	0.95 (0.001)	0.98 (0.001)
Religion	Christian	0.87	0.89
	Islam	0.90	0.89
	Atheism	0.79	0.80
	Average	0.86 (0.002)	0.85 (0.002)
Politician names	Conservatives	0.95	0.97
	Liberals	0.91	0.91
	Socialists	0.92	0.89
	Average	0.93 (2.0e-4)	0.92 (0.001)
Political Affiliation	Left-wing	0.96	0.91
	Right-wing	0.91	0.80
	Centrist	0.96	0.88
	Average	0.94 (6.0e-4)	0.86 (0.002)
Occupation	Services	0.65	0.70
	Creative Arts and Media	0.67	0.68
	Trades and Manual Labor	0.67	0.64
	Average	0.66 (7.0e-5)	0.67 (0.0005)

uation, detailed in Table 1, employs F1-scores to compare performance across six bias categories using the INV test set established by Wessel and Horych (2024). The results are displayed by subcategory and then averaged for a category score. Furthermore, the variance among subcategory results is calculated per category.

The comparison reveals two principal findings: Firstly, in four out of six categories, the model trained with the BiasAdapt-augmented dataset displays a lower classification performance variance (remaining the same in the remaining categories). As the variance is the primary measure for reliance on spurious cues, this indicates that BiasAdapt contributes to a more consistent classification performance across different subcategories and reduces reliance on spurious cues. Secondly, the overall performance in the gender category dropped significantly after training on Augmented BABE, improved for political affiliation, and remained relatively stable for all other categories.

5 Discussion

BiasAdapt successfully identifies and replaces relevant keywords though there is still an underrepresentation of certain categories with little occurrence in the original dataset. The observed decrease in variance for a majority of categories due to the BiasAdapt augmentation underscores the method's effectiveness in diminishing the model's dependence on predefined bias-indicative keywords. The reduced reliance on keywords suggests that augmentation helps the model analyze the text holistically rather than fixating on specific terms. However, this does not work for all categories, and intra-category differences remain.

The decrease in performance observed in the gender category raises important questions about the role of spurious cues in automated bias detection. Unlike political affiliation or origin, where bias is often directly linked to framing, gender bias tends to involve more implicit associations tied to societal roles or traits. The reliance on these implicit cues might have served as a shortcut, aiding model performance in some cases. In the context of gender, keyword substitutions can interact with these subtleties, potentially altering the strength or direction of bias in ways that are difficult to predict.

The relative stability in F1-Scores across the other categories suggests that the model's ability to detect bias in these areas is less disturbed by reducing reliance on spurious cues. This could indicate that the model's prior results in these categories were less dependent on problematic shortcuts or, alternatively, that the augmentation process more effectively preserves the essential signals of bias within these contexts.

6 Future Work

Several avenues for research emerge from the findings of this study. Further investigations into why the performance changed for two categories, as well as why the variance did not decrease for two, is necessary. Expanding the scope of model architectures tested, including a diverse array of language models, could provide a more comprehensive understanding of BiasAdapt's applicability and effectiveness. This would enable a broader assessment of the augmentation process across different computational frameworks for bias detection.

To mitigate potential shifts in bias, future work could explore filtering mechanisms that detect when a keyword replacement significantly alters a sentence's framing. Additionally, human evaluation of augmented sentences could help assess whether bias labels remain appropriate after substitution, particularly in the gender category.

Addressing the limitation related to the requirement for predefined bias categories, future research could explore developing more adaptive, exploratory methods for identifying potential biases. Such approaches could leverage unsupervised learning techniques or advanced content analysis methods to uncover hidden or emergent bias categories, thereby broadening the scope and applicability of the BiasAdapt method. Moreover, an important direction for future work is investigating whether methods like BiasAdapt can indirectly contribute to improving models' contextual understanding of texts by reducing models' reliance on spurious cues. This could involve integrating techniques to enhance semantic comprehension and inferential reasoning within models, thereby addressing one of the fundamental challenges in automatic bias detection.

7 Conclusion

This study presents BiasAdapt, a data augmentation strategy aimed at improving the robustness of media bias detection systems through adversarial examples. By leveraging prior knowledge of spurious cue dependencies, BiasAdapt demonstrates that data augmentations utilizing large language models (LLMs) can have a measurable impact on improving bias detection performance. Making a significant corpus available for public use lays the groundwork for further exploration in the field. While the focus on a single model and a select number of bias categories limits the generalizability of the findings, this work demonstrates the potential of leveraging LLMs for dataset augmentation and increased robustness in media bias detection. Despite the demonstrated benefits, further investigations to better understand model behavior is necessary. Still, it encourages expanding the scope and transfer to other areas of text analysis with prerequisites similar to media bias.

Limitations

Primarily, the analysis is confined to using a single model architecture, specifically a RoBERTa model. Though beneficial for ensuring comparability with prior work such as Wessel and Horych (2024), this choice restricts understanding how the proposed BiasAdapt augmentation might perform across a broader spectrum of model architectures. Another limitation arises from the reliance on GPT3.5 to generate alternative words. Manual inspections have revealed instances where GPT3.5 may incorrectly identify keywords or suggest inappropriate alternatives. While these errors are infrequent and do not significantly detract from the overall efficacy of the augmentation, they underscore the need for caution and oversight in using generative language models for data augmentation tasks. Furthermore, the replacement can lead to generic or contextually inconsistent replacements, where sentences remain grammatically correct but become unrealistic or lose their meaning.

Additionally, the BiasAdapt approach assumes a priori knowledge of bias categories and subcategories, necessitating predefined taxonomies for media bias. This requirement could constrain the method's applicability, as it presupposes theoretical or empirical insights into potential sources of spurious cues. While this study addresses the issue of over-reliance on specific cues for bias detection, it does not tackle the broader challenge of enhancing models' contextual understanding. This limitation points to an inherent constraint in the scope of the current methodological approach. Lastly, querying an LLM for each sentence and generating multiple permutations can be computationally intensive and time-consuming, particularly for large datasets.

References

Eric Baumer, Elisha Elovic, Ying Qin, Francesca Polletta, and Geri Gay. 2015. Testing and Comparing

Computational Approaches for Identifying the Language of Framing in Political News. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1472–1482, Denver, Colorado. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Felix Hamborg. 2020. Media bias, the social sciences, and nlp: Automating frame analyses to identify bias by word choice and labeling. In *Proceedings of the 58th annual meeting of the association for computational linguistics: student research workshop*, pages 79–87.

Felix Hamborg, Karsten Donnay, and Bela Gipp. 2019. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, 20(4):391–415.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Maximilian Mozes, Pontus Stenetorp, Bennett Kleinberg, and Lewis Griffin. 2021. Frequency-guided word substitutions for detecting textual adversarial examples. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 171–186, Online. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Marta Sabou, Kalina Bontcheva, and Arno Scharl. 2012. Crowdsourcing research opportunities: lessons from natural language processing. In *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies*, pages 1–8.

Suranjana Samanta and Sameep Mehta. 2017. Towards crafting text adversarial samples. *arXiv preprint arXiv:1707.02812*.

Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. 2019. Adversarial training for free! *Advances in neural information processing systems*, 32.

Timo Spinde, Smilla Hinterreiter, Fabian Haak, Terry Ruas, Helge Giese, Norman Meuschke, and Bela Gipp. 2023. The media bias taxonomy: A systematic literature review on the forms and automated detection of media bias. *arXiv preprint*.

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.

Xuezhi Wang et al. 2023. Identifying and mitigating spurious correlations for improving robustness in nlp models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Martin Wessel and Tomáš Horych. 2024. Beyond the surface: Spurious cues in automatic media bias detection. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 21–30, St. Julian's, Malta. Association for Computational Linguistics.

Martin Wessel, Tomas Horych, Terry Ruas, Akiko Aizawa, Bela Gipp, and Timo Spinde. 2023. Introducing MBIB - The First Media Bias Identification Benchmark Task and Dataset Collection. In *Proceedings of 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'23)*, New York, NY, USA. ACM. ISBN 978-1-4503-9408-6/23/07.

Ye Zhang and Byron Wallace. 2015. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*.