Giving Control Back to Models: Enabling Offensive Language Detection Models to Autonomously Identify and Mitigate Biases

Anonymous ACL submission

Abstract

The rapid development of social media has led to an increase in online harassment and offensive speech, posing significant challenges for effective content moderation. Existing automated detection models often exhibit a bias towards predicting offensive speech based on specific vocabulary, which not only compromises model fairness but also potentially exacerbates biases against vulnerable and minority groups. Addressing these issues, we propose a bias self-awareness and data self-iteration framework for mitigating model biases.This framework aims to "giving control back to models: enabling offensive language detection models to autonomously identify and mitigate biases" through bias self-awareness algorithms and self-iterative data augmentation method. Experimental results demonstrate that the proposed framework effectively reduces the false positive rate of models in both in-distribution and out-of-distribution tests, enhances model accuracy and fairness, and shows promising performance improvements in detecting offensive speech on larger-scale datasets.

1 Introduction

The rapid development of social media has significantly enhanced the ease with which people can connect, share, and obtain data online, as well as convey emotional messages. However, the convenience of internet technology has concurrently increased the risk of individuals encountering cyberbullying and online attacks. Automatic detection of offensive language is an effective measure to maintain the safety, health, and friendliness of online social platforms (Schmidt and Wiegand, 2017). This technology has broad applications across various internet interaction environments, including social networks, online forums, instant messaging tools, news media platforms, and gaming communities.

By integrating multiple natural language processing (NLP) techniques, numerous models (Zhou et al., 2021a; Fan et al., 2024; Lu et al., 2023a) have been designed and applied to the task of detecting offensive language. However, even the most advanced models tend to overly rely on specific words to predict offensive content(Kennedy et al., 2020), often mistakenly classifying sentences containing these words as offensive(Zhou et al., 2021b). This phenomenon raises concerns about bias in offensive language detection systems, thereby limiting their fairness(Ramponi and Tonelli, 2022). Additionally, it can lead to prejudiced treatment of vulnerable and minority groups, potentially exacerbating racism(Harris et al., 2022).

In offensive language detection, not only identity-related vocabulary such as "gay" or "black"(Waseem and Hovy, 2016) but also non-identity-related vocabulary like "sport" and "football" are often inappropriately associated with offensive content. One of the root causes of this issue lies in the biases present in the data collection process(Wiegand et al., 2019). Because the collected data frequently place these specific vocabulary in offensive contexts, it fosters erroneous statistical associations between these vocabulary and offensive labels, known as spurious statistical correlations. Models learn and make predictions based on these spurious statistical correlations, leading to biases in the models themselves. These incorrectly associated vocabulary are commonly referred to as "spurious artifacts," and their associations with labels are termed "spurious correlations"(Ramponi and Tonelli, 2022).

Regarding the identification of spurious arti-

facts, Ramponi and Tonelli (2022) approached this issue by examining datasets and employing statistical methods such as Pointwise Mutual Information (PMI) to measure the potential association strength between a word and offensive labels. Subsequently, they used manual annotation to identify spurious artifacts. However, this method has two significant drawbacks: 1) Given the vastness of datasets, manual annotation is impractical. 2) The spurious artifacts identified from the dataset may not be universally applicable to all models; for instance, Model A might be misled by a spurious artifact x, while Model B remains unaffected.

To mitigate model biases, Zhang et al. (2023) proposed a data augmentation method that utilizes large language models like GPT-3 to generate sentences and expand negative sample instances, thereby balancing the dataset and reducing model bias. Experimental results indicate that data augmentation is an effective approach for mitigating model bias. However, determining the amount of data augmentation often relies on the researchers' prior experience and lacks objective criteria, making the process largely subjective.

To address the aforementioned issues, we propose a model bias correction framework based on Bias Self-Awareness and Data Self-Iteration (BSADSI), which is founded on the core principle of "giving control back to models." The BSADSI framework incorporates an innovative Model Bias Self-Awareness algorithm (MBSA), enabling the model to autonomously identify and acquire spurious artifacts. Furthermore, BSADSI integrates reinforcement learning strategies, allowing the model to independently determine the content and extent of data augmentation. Our main contributions are as follows:

1. We propose the Model Bias Self-Awareness algorithm framework (MBSA), which automatically identifies spurious artifacts in the dataset, thereby achieving autonomous understanding and identification of biases.

2. We introduce a self-iterative data augmentation method that utilizes large language model to enhance datasets. We integrate reinforcement learning strategies to enable the model to autonomously determine the amount of data augmentation based on MBSA feedback, automatically expanding negative sample instances, thereby enhancing its self-learning and adaptation capabilities through iterative improvements.

3. Experimental results demonstrate that the BSADSI framework we proposed effectively reduces the false positive rate of models in offensive language detection tasks, improves model robustness, and enhances fairness in the recognition process.

2 Related Work

In this chapter, we systematically review research findings in two aspects: identifying spurious correlations in detecting offensive language and methods for mitigating model biases.

2.1 Identifying Spurious Correlation in Offensive Language Detection

Previous research has extensively explored strategies to identify spurious correlations in detecting offensive language. Manerba and Tonelli (2021) manually crafted test templates and replaced identity attributes within them to observe how model predictions vary with these changes, thereby identifying biases in specific identity features. Röttger et al. (2021), based on relevant literature and informal interviews, designed 29 functional tests, constructing test cases and validating them effectively to reveal biases in models like BERT. Ramponi and Tonelli (2022) employed Pointwise Mutual Information (PMI) to assess the potential strength of correlations between vocabulary and offensive labels. They then used manual annotations to remove authentic artifacts and identify spurious artifacts. Building on this literature, Zhang et al. (2023) introduced the Relative Spuriousness (RS) method to verify the spurious correlation between words and labels. Despite these methods achieving some success in identifying spurious correlations in offensive language detection, they generally fail to fully consider the variability between models and often overlook the importance of the model's own role in the identification process and its potential impact.

2.2 Methods for Mitigating Model Bias

In the realm of offensive language detection, various methods have been widely employed to mitigate model biases. Sen et al. (2021, 2022) explored the impact of Counterfactually Augmented Data on offensive language detection models, utilizing techniques such as inserting irrelevant information and synonym substitution to construct counterfactual data. Bose et al. (2022) employed regularization techniques on Spurious Artifacts to alleviate model biases. Many researchers have mitigated model biases by expanding negative sample instances. Wullach et al. (2021) leveraged the pre-trained GPT-2 model to generate large-scale text sequences, expanding manually annotated hate speech datasets to balance the dataset and reduce model biases. Hartvigsen et al. (2022) used GPT-3 to generate the TOXIGEN dataset, aiming to balance the distribution of offensive language and mitigate biases against minority groups. Previous studies demonstrate that data augmentation is an effective approach to mitigate model biases. However, determining the required amount of data often heavily relies on researchers' intuition and experience, lacking objective methods to quantify the necessary data scale for reducing model biases.

3 Methodology

The Model Bias Correction Framework BSADSI we proposed is illustrated in Figure 1. This framework primarily consists of two processes: Bias Self-Awareness (MBSA) and self-iterative data augmentation method. In the MBSA process, we initially use an offensive speech detection model to classify the data from the validation set, identifying instances with high confidence but incorrect judgments to construct a bias dataset. Subsequently, we extract vocabulary from this bias dataset, conduct filtering and validation to obtain a set of spurious artifacts. Finally, we compute a bias coefficient for each spurious artifacts to determine the scale of generated data. During the self-iterative data augmentation process, we introduce reinforcement learning strategies where the offensive speech detection model acts as an agent. Through interactions between MBSA and a Reward Function feedback loop, the large language model iteratively generates sample data containing spurious artifacts, thus expanding the training set contrapuntally. This iterative process dynamically adjusts the quantity of newly added data, optimizing the model's ability to dentify and correct biases.

Algorithm (Appendix A) outlines the iterative process of the BSABSI framework. Initially, the model undergoes initial fine-tuning on the unaugmented base dataset. Subsequently, the model's performance is evaluated using a reward function, recording this initial score. The MBSA module analyzes the spurious artifacts set generated by the model in this round and determines the demand for negative example samples. This information guides the large language model to generate negative example samples, which are then integrated into the training dataset, completing the initial augmentation.As the process proceeds to the N-th iteration, the model undergoes further fine-tuning on the dataset expanded from the previous N-1 rounds. After adjustments, the model is re-evaluated using the scorer, comparing its score with that of the N-1 rounds. If no score improvement is observed for T consecutive rounds, the model is deemed optimal, and the iteration process terminates. Conversely, if performance continues to improve, MBSA intervenes again to analyze the spurious artifacts set identified by the model in this round and determine the scale of additional negative example samples to be added.It is noteworthy that if MBSA in a particular round fails to discover new spurious artifacts, the iteration will also terminate. If the termination condition is not met, the iterative process described above is repeated.

3.1 MBSA algorithm framework

The MBSA framework consists of three main components: bias data acquisition, spurious artifacts acquisition, and bias coefficient calculation.

(1) Bias data acquisition

To tackle the problem of model bias resulting from data imbalance, we start by evaluating the validation set to quantify the extent of the bias in the model. Initially,

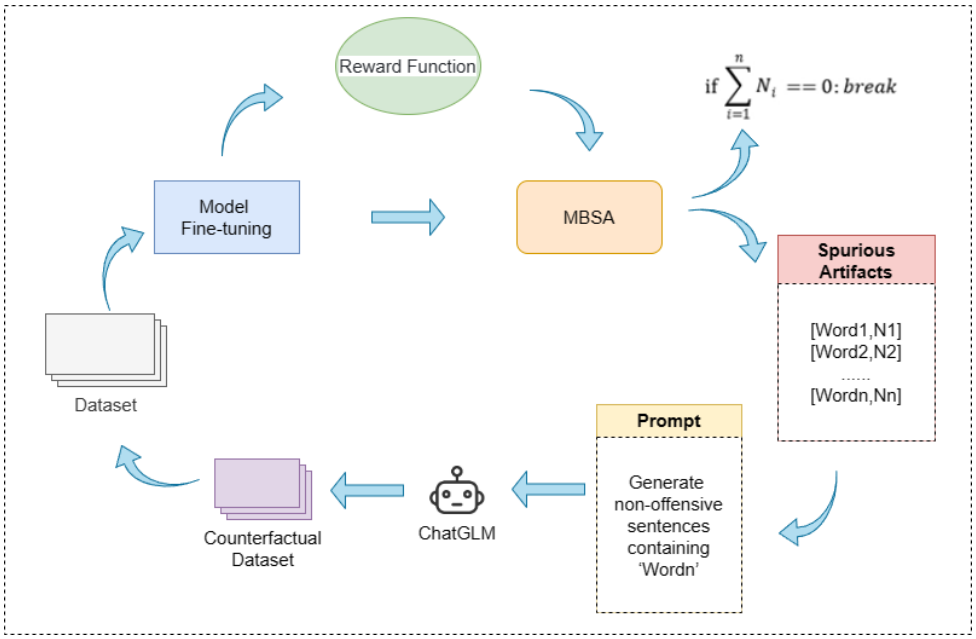

Figure 1: BSADSI.

a threshold, represented by θ, is established as the standard for bias identification.When the difference between the positive and negative class probabilities for a sample in the validation set exceeds a predefined threshold θ , and the model's prediction contradicts the actual label of the sample, we deem it highly likely that the sample contains spurious artifacts that induce model misclassification. Using a fine-tuned model, we systematically examine the entire validation set, employing the aforementioned bias identification criteria to automatically screen and gather samples exhibiting bias characteristics. These aggregated samples constitute our bias data set, which is a critical input for further bias understanding and model optimization.

(2) Spurious artifacts acquisition

After acquiring the bias data set, the primary task shifts to identifying spurious artifacts contributing to model bias. Initially, we perform word segmentation on the Chinese data, removing stop words and words with strong negative sentiment to reduce noise. Subsequently, we employ the Pointwise Mutual Information (PMI) method to select words that are highly correlated with the offensive speech label, creating a candidate set of spurious artifacts. We then utilize a masking validation strategy, where each candidate spurious artifact is individually masked within the sentence. If the model' s prediction changes from incorrect to correct upon masking the word, it indicates that the word significantly impacts the model' s ability to identify offensive speech, and it is added to the spurious artifacts set.

(3) Bias coefficient calculation

Spurious artifacts can interfere with the model' s ability to accurately identify offensive speech. To quantify the misleading effect of each spurious artifact on the model, we introduce Equation 1.

$$R = \frac{N_{w,FP}}{N_{w,neg}} \quad (1)$$

Here, R denotes the bias coefficient, $N_{w,FP}$ is the number of sentences in the validation set containing the spurious artifact w that the model has incorrectly classified as offensive speech, and $N_{w,neg}$ is the number of non-offensive sentences in the validation set that also contain the spurious artifact w.

The greater the bias coefficient R, the more misleading the spurious artifact is, which suggests the need to augment the

training set with more non-offensive (negative) samples containing this spurious artifact to balance the data and mitigate model bias. We have formulated a strategy for determining the number of additional negative samples required based on each spurious artifact's bias coefficient. The specific quantification method is illustrated in Equation 2:

$$a = R \times (N_{w,Off} - N_{w,NonOff}) \quad (2)$$

Here, a represents the number of additional negative samples required. $N_{w,Off}$ is the number of offensive sentences in the training set that contain the spurious artifact, and $N_{w,NonOff}$ is the number of non-offensive sentences in the training set that contain the spurious artifact.

3.2 Self-iterative data augmentation method

The self-iterative data augmentation method introduces reinforcement learning strategies, enhancing data systematically through a continuous iterative process. Its core components include a reward function and a data generator based on a large-scale language model.

(1) Reward function

False positive rate (FPR) emphasizes the proportion of negative samples that are incorrectly classified as positive instances. This is particularly critical in scenarios involving the detection of offensive speech, where a high false positive rate can lead to innocent users or information being wrongly labeled or restricted, thus compromising system fairness and user experience. (Ramponi and Tonelli, 2022) highlights false positive rate as a key metric for assessing bias in offensive speech detection models. Hence, we utilize false positive rate as the criterion for the reward function (RF), quantified specifically as shown in Equation 3.

$$RF = 1 - \frac{D_{FP}}{D_{neg}} \quad (3)$$

Here, D_{FP} is the number of sentences in the validation set that the model incorrectly classifies as offensive speech, and D_{neg} is the number of non-offensive sentences in the validation set.

(2) Data generator

The ChatGLM (Zeng et al., 2023) model has been extensively customized and trained for the Chinese language context, enabling it to achieve higher accuracy and fluency in handling Chinese natural language tasks. Compared to other large language models, ChatGLM demonstrates better understanding and generation of text that aligns with Chinese cultural backgrounds and linguistic norms. The model implements stringent generation constraints, effectively suppressing the generation of potentially offensive or inappropriate content. Additionally, aided by prompt templates designed in Appendix B, ChatGLM can generate targeted high-quality Chinese examples more effectively. Therefore, we utilize ChatGLM as a generator to enhance the data by generating negative examples containing spurious artifacts.

4 Experiment and Analysis

In this section, we first introduce the dataset, model and evaluation metrics. Next, we compare the model after correction with the uncorrected model using BSADSI. Finally, detailed analysis is provided.

4.1 Dataset,Model and Evaluation metrics

During the experiment, three publicly available Chinese offensive speech datasets were used in this article: COLD (Deng et al., 2022); TOXICN(Lu et al., 2023b); SWSR(Jiang et al., 2022).

To compare and analyze the performance of different models in identifying spurious artifacts and correcting biases, we utilize BERT[1] and RoBERTa[2].

During the evaluation phase, we use F1 score and false positive rate (FPR) as the core evaluation metrics to comprehensively assess the performance of the models.

[1] https://huggingface.co/bert-base-chinese
[2] https://huggingface.co/hfl/chinese-roberta-wwm-ext

Table 1: The differences in how different models identify spurious artifacts.

Model	Bert	Roberta
Spurious Artifacts	日本	暴力
	外地人	男人
	中国	素质
	白人	反感
	四川人	

4.2 The comparison of different models in identifying spurious artifacts.

To compare the differences in how different models autonomously identify spurious artifacts, we conduct a statistical analysis of spurious artifacts perceived by BERT and RoBERTa. Table 1 presents the unique spurious artifacts perceived by each model. BERT autonomously identified 5 unique spurious artifacts, accounting for approximately 28% of the total, while RoBERTa identified 4 unique spurious artifacts, accounting for about 24%. BERT appears to be more sensitive to vocabulary indicating geographical or ethnic references, which it may interpret as potential markers of offensive speech. On the other hand, RoBERTa's biases tend towards gender and certain non-identity-related vocabulary.

4.3 Comparison of model bias correction experiments

To validate the performance of BSADSI, we followed the testing methodology proposed by (Ramponi and Tonelli, 2022). We conducted in-distribution testing on the COLD dataset and out-of-distribution testing on the TOXICN and SWSR datasets. The experimental results are shown in Table 2. For in-distribution testing, we trained the baseline model on the COLD training set and evaluated it on the test set. For out-of-distribution testing, COLD was used as the training set, and the model was evaluated on the test sets of TOXICN and SWSR datasets.

From Table 2, it can be observed that both BERT and RoBERTa models, when using the BSADSI framework for bias identification and correction, show improvements in all evaluation metrics during in-distribution testing on the COLD dataset. Particularly notable is the significant decrease in false positive rate (FPR). For out-of-distribution testing, the BSADSI framework also demonstrates effective results, maintaining or slightly improving F1 score and accuracy (ACC) while effectively reducing the false positive rate.

It is noteworthy that Bert-BSADSI shows a slight decrease in precision on TOXICN and SWSR. This is because models not employing the BSADSI framework sometimes misclassify negative examples containing spurious artifacts by erroneously associating them with offensive content without understanding their semantic meaning. BSADSI effectively eliminates such false associations, necessitating a reassessment of previously misclassified samples, resulting in minor declines in ACC and F1 on small-scale datasets. However, the BSADSI framework significantly reduces false positive rates, suggesting potential improvements in model performance on a broader range of data scenarios while enhancing fairness.

To further investigate potential biases in the model or its excessive sensitivity to specific vocabulary, we quantified the improvement in reducing spurious artifacts by comparing the false positive rates of spurious artifacts before and after applying the BSADSI framework. The experimental results are presented in Appendix C.

The experimental results shown in Appendix C indicate that after applying the BSADSI framework, the false positive rates of spurious artifacts significantly decreased for both Bert and RoBERTa models across the COLD, TOXICN, and SWSR datasets. Specifically, for the Bert model, there was a notable reduction in false positive rates when handling offensive statements involving vocabulary like ” 黑人” and ” 恐怖”, demonstrating that the BSADSI framework effectively mitigates inappropriate responses to specific sensitive vocabulary. Additionally, the false positive rates for frequently mentioned keywords such as ” 警察”, ” 女性” and ” 暴力” also declined, reflecting an improvement in the models' fairness and accuracy when addressing gender and violence-related topics. However, some spurious artifacts like ” 井盖”, ” 河南人” and ” 东北人” showed only a minor decrease in false positive rates, suggesting that erroneous asso-

Table 2: In-distribution and out-of-distribution results(↑: greater the better; ↓: lower the better.)

Model	COLD			TOXICN			SWSR		
	ACC↑	F1↑	FPR↓	ACC↑	F1↑	FPR↓	ACC↑	F1↑	FPR↓
Bert	82.1	79.2	20.8	66.2	**61.5**	16.7	67.5	**60.9**	35.5
Bert-BSADSI	**82.9**	**79.2**	**16.8**	**66.2**	59.7	**12.8**	**69.2**	58.9	**28.1**
Roberta	82.5	79.5	20.9	66.9	61.9	15.4	67.2	58.3	32.5
Roberta-BSADSI	**83.2**	**80.2**	**18.6**	**67.7**	**63.2**	**13.9**	**68.9**	**59.5**	**29.6**

ciations triggered by such data are more challenging to rectify.

Figure 2 illustrates the changes in attention weights of the offensive language detection model before and after bias correction. The depth of color in the rectangles visually represents the magnitude of the attention weights. As shown in Figure 2, before bias correction, the attention weight assigned to the term ” 黑人” was significantly higher than that for other words in the sentence. This disproportionate attention might cause the model to be overly sensitive to the term ” 黑人” leading to biased interpretations of the overall meaning of the sentence. After applying the BSADSI framework for bias correction, the attention weight for the term ” 黑人” significantly decreased. This change reflects the effectiveness of the BSADSI framework in reducing model bias.

4.4 Comparison of data augmentation methods

To conduct an in-depth analysis and comparison of the effects of different data augmentation strategies on the performance of offensive language detection, we evaluate the effectiveness of the proposed BSADSI framework in enhancing model accuracy and reducing false positives. Comparative experiments were conducted, maintaining consistency with previous methodologies, and employing both in-distribution and out-of-distribution testing methods. The experimental results are presented in Table 3. In this table, ”Raw Data” indicates the use of unaugmented data, while ”1:0.5” and ”1:1” represent the positive-to-negative sample ratios with spurious artifacts included after data augmentation. ”BSADSI” denotes the application of the proposed framework.

The experimental results indicate that for in-distribution testing, compared to fixed-ratio data augmentation methods, BSADSI significantly reduces the false positive rate while maintaining comparable performance in other evaluation metrics. When extended to out-of-distribution testing, fixed-ratio augmentation methods may encounter an increase in false positive rates, whereas BSADSI continues to effectively reduce false alarms. It is noteworthy that the BSADSI framework does not exhibit significant advantages in terms of ACC and F1 scores on out-of-distribution testing across the two datasets. This is primarily due to the presence of spurious artifacts in the COLD dataset, which challenges the model’s ability to identify offensive language when the test set encompasses a broader range of data sources with inconsistent distributions, thereby impacting overall performance.

The BSADSI framework enhances data dynamically and purposefully through multi-iteration processes. Experimental data indicates that achieving a 1:0.5 augmentation ratio requires adding 1,314 new instances, whereas a 1:1 ratio necessitates 6,917 new instances. In contrast, the BSADSI framework only requires an additional 3,629 instances. Furthermore, experimental results demonstrate that the BSADSI framework not only reduces dependency on a large volume of extra data but also mitigates the risk of overfitting that can arise from excessive augmentation.

5 Conclusion

The BSADSI framework we proposed demonstrates significant effectiveness in mitigating biases in offensive speech detection models. At its core, this framework aims to give control back to the model itself to correct biases by employing bias self-awareness algorithms and self-iterative data augmentation method. The

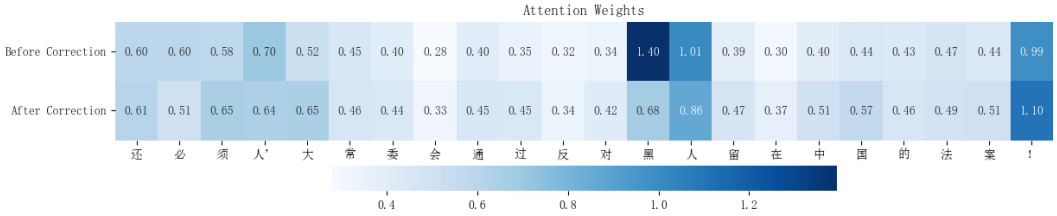

Figure 2: Comparison of Attention Weights Before and After BSADSI.

Table 3: Comparison of experimental results using different data augmentation methods

Model: Bert	COLD			TOXICN			SWSR		
	ACC↑	F1↑	FPR↓	ACC↑	F1↑	FPR↓	ACC↑	F1↑	FPR↓
Raw Data	82.1	79.2	20.8	66.2	61.5	16.7	67.5	**60.9**	35.5
1:0.5	82.3	79.2	19.5	66.2	60.8	15.39	68.0	59.7	32.3
1:1	82.4	**79.5**	20.1	**67.5**	**63.7**	17.5	65.1	58.3	37.8
BSADSI	**82.9**	79.2	**16.8**	66.2	59.7	**12.8**	**69.2**	58.9	**28.1**

bias self-awareness algorithm automates bias data acquisition, identifies spurious artifacts, and calculates bias coefficients, thereby enhancing efficiency in recognizing spurious associations and ensuring that the model can identify and understand the sources of bias based on its own characteristics. The self-iterative data augmentation method introduces reinforcement learning strategies, allowing the model to autonomously determine the content and scale of data expansion based on feedback from MBSA, thereby achieving dynamic optimization of data augmentation. Experimental results indicate that the BSADSI framework not only effectively reduces the false positive rate in both in-distribution and out-of-distribution tests but also enhances model accuracy and fairness. Moreover, it shows promising potential to significantly improve the performance of offensive speech detection on larger-scale datasets.

6 Limitations

Our research aims to mitigate biases in offensive speech detection models. However, we are aware of several limitations. Firstly, our work primarily focuses on analyzing Chinese language corpora, and our experiments have not yet encompassed non-Chinese language resources. In future work, we plan to expand our framework to evaluate its performance on multilingual offensive speech datasets.Additionally, the bias correction capability of our framework needs enhancement when dealing with implicit offensive speech that employs rhetorical devices such as metaphors, irony, and puns. Future research will concentrate on addressing model biases in detecting implicit offensive speech within complex linguistic contexts.

7 Ethics Statement

Due to the nature of this work, some examples include offensive text and language. However, these examples do not reflect the values of the authors; rather, our research aims to mitigate biases in offensive language detection models and to detect and prevent the spread of harmful content. Furthermore, the Chinese datasets used in our study are publicly available, and we did not anticipate any specific ethical concerns related to this work.

References

Tulika Bose, Nikolaos Aletras, Irina Illina, and Dominique Fohr. 2022. Dynamically refined regularization for improving cross-corpora hate speech detection. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 372–382, Dublin, Ireland. Association for Computational Linguistics.

Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. 2022. COLD: A benchmark for Chinese offensive language detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11580–11599, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xiaochao Fan, Jiapeng Liu, Junjie Liu, Palidan Tuerxun, Wenjun Deng, and Weijie Li. 2024. Identifying hate speech through syntax dependency graph convolution and sentiment knowledge transfer. *IEEE Access*, 12:2730–2741.

Camille Harris, Matan Halevy, Ayanna Howard, Amy Bruckman, and Diyi Yang. 2022. Exploring the role of grammar and word choice in bias toward african american english (aae) in hate speech classification. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 789–798, New York, NY, USA. Association for Computing Machinery.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Aiqi Jiang, Xiaohan Yang, Yang Liu, and Arkaitz Zubiaga. 2022. Swsr: A chinese dataset and lexicon for online sexism detection. *Online Social Networks and Media*, 27:100182.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442, Online. Association for Computational Linguistics.

Junyu Lu, Hongfei Lin, Xiaokun Zhang, Zhaoqing Li, Tongyue Zhang, Linlin Zong, Fenglong Ma, and Bo Xu. 2023a. Hate speech detection via dual contrastive learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2787–2795.

Junyu Lu, Bo Xu, Xiaokun Zhang, Changrong Min, Liang Yang, and Hongfei Lin. 2023b. Facilitating fine-grained detection of Chinese toxic language: Hierarchical taxonomy, resources, and benchmarks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16235–16250, Toronto, Canada. Association for Computational Linguistics.

Marta Marchiori Manerba and Sara Tonelli. 2021. Fine-grained fairness analysis of abusive language detection systems with checklist. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 81–91.

Alan Ramponi and Sara Tonelli. 2022. Features or spurious artifacts? data-centric baselines for fair and robust hate speech detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3027–3040, Seattle, United States. Association for Computational Linguistics.

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)*, pages 41–58, Online. Association for Computational Linguistics.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Indira Sen, Mattia Samory, Fabian Flöck, Claudia Wagner, and Isabelle Augenstein. 2021. How does counterfactually augmented data impact models for social computing constructs? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 325–344, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Indira Sen, Mattia Samory, Claudia Wagner, and Isabelle Augenstein. 2022. Counterfactually augmented data and unintended bias: The case of sexism and hate speech detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4716–4726, Seattle, United States. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. 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 602–608, Minneapolis, Minnesota. Association for Computational Linguistics.

Tomer Wullach, Amir Adler, and Einat Minkov. 2021. Fight fire with fire: Fine-tuning hate detectors using large samples of generated hate

speech. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4699–4705, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130B: an open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Zhehao Zhang, Jiaao Chen, and Diyi Yang. 2023. Mitigating biases in hate speech detection from a causal perspective. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6610–6625, Singapore. Association for Computational Linguistics.

Xianbing Zhou, Yang Yong, Xiaochao Fan, Ge Ren, Yunfeng Song, Yufeng Diao, Liang Yang, and Hongfei Lin. 2021a. Hate speech detection based on sentiment knowledge sharing. 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 7158–7166, Online. Association for Computational Linguistics.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah Smith. 2021b. Challenges in automated debiasing for toxic language detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3143–3155, Online. Association for Computational Linguistics.

A Iterative Model Refinement with BSABSI

Result: Model refined with iterative data augmentation until optimal.
Input : Original dataset D, scoring function R, MBSA module, language generation module for negative samples.

for $t \leftarrow 1$ **to** N **do**
  // Refine model on current dataset
  $M_t \leftarrow TrainModel(D_{t-1})$;
  // Calculate score with reward function
  $Score_t \leftarrow R(M_t, D_{t-1})$;
  if $t == 1$ **then**
    $OriginalScore \leftarrow Score_t$;
    // Get spurious artifacts set and number of negative examples to be generated via MBSA
    $FASet_t, NegSampleCounts_t \leftarrow MBSA(M_t)$;
  else
    // Compare scores to detect improvement
    if $Score_t > OriginalScore$ **then**
      $OriginalScore \leftarrow Score_t$;
      $FASet_t, NegSampleCounts_t \leftarrow MBSA(M_t)$;
      $T \leftarrow 0$;
    end
    else
      // Stop if the 5-wheel does not lift
      if $T > 5$ **then**
        break;
      end
      else
        // continue if the T <= 5
        continue;
      end
    end
  end
  // Check for termination conditions in MBSA feedback
  if $|NegSampleCounts_t| == 0$ *indicates balance* **then**
    break;
  end
  // Generate and augment negative samples
  $NegSamples_t \leftarrow GenerateNegSamples(FASet_t, NegSampleCounts_t)$;
  $T \leftarrow T + 1$;
  $D_t \leftarrow D_{t-1} \cup NegSamples_t$;
end

Algorithm 1: Iterative Model Refinement with BSABSI

B Prompt template

prompt = 生成一个包含 “xx” 这个词，但不具有攻击性的句子, 要求不超过 40 个字。

C The Variation in False Positive Rates of Spurious Artifacts

Bert					
COLD		TOXICN		SWSR	
Spurious Artifacts	FPR Decline(%)	Spurious Artifacts	FPR Decline(%)	Spurious Artifacts	FPR Decline(%)
恐怖	35.3	警察	100.0	黑人	100.0
刻板	33.3	艾滋	57.1	恐怖	33.4
日本	20.8	女人	31.3	男性	14.3
外地人	20.7	白人	28.6	女性	13.3
井盖	20.0	黑人	22.5	警察	12.5
Roberta					
COLD		TOXICN		SWSR	
Spurious Artifacts	FPR Decline(%)	Spurious Artifacts	FPR Decline(%)	Spurious Artifacts	FPR Decline(%)
恐怖	52.9	素质	20.0	反感	100.0
暴力	5.0	女人	18.8	恐怖	33.3
井盖	4.0	男人	15.0	警察	12.5
女人	4.0	女性	10.0	暴力	12.5
河南人	3.7	东北	8.0	女性	5.0