Solutions

There are some recent works trying to solve this bias technically by merging the general bias mitigating methods into the model and using more inclusive datasets to train the model, as well as proposing new educational policies to minimize its negative impacts.

Adjusting Threshold

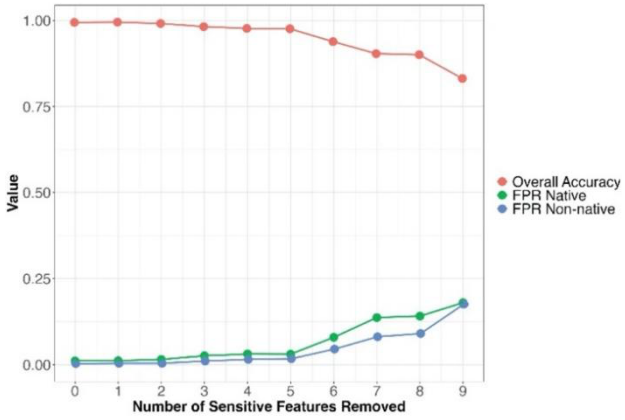
Adjusting the decision threshold is a powerful technique in machine learning, and it helps to improve models' accuracy and fairness, especially when dealing with biases (Esposito et al., 2021). The threshold determines at which point a model classifies data into one category or another. Adjusting the threshold works by shifting the point where a model classifies an input, giving it the flexibility to account for biases inherent in the data. This is especially useful in reducing bias against minority groups or in cases where there is a disproportionate representation in the dataset.

Studies by Jiang et al. (2024b) explored the implementation of adjusting decision thresholds to try to mitigate bias in AI detectors. In their study, 75,567 human-written essays were collected, where 30.6% were submitted by self-identified NS and rest were submitted by self-identified NNS. Finally, a dataset consisted of 10,000 randomly selected essays from 75,567 human-written responses and 10,000 GPT-4-generated essays was used for training and validating the AI essay detectors, which is relatively balanced. Jiang et al. (2024b) created separate decision thresholds for NS and NNS to optimize fairness between the two groups. They tested various threshold combinations and selected the one that best minimized the False Positive Rate Parity (FPRP), an assessment for whether the false-positive rate is similar across different groups. To implement this, they randomly assigned half of the AI-generated essays to the NS group and the other half to the NNS group, in order to make sure the threshold is consistent through the experiment. As the result, the model with adjusting threshold only has a 0.03% lower FPRP than regular models, without making any difference in overall accuracy (Jiang et al. 2024b).

While adjusting the decision threshold in Jiang et al.'s study provides an improvement in fairness, the reduction in FPRP by just 0.03% suggests that this technique alone has a very limited impact on mitigating bias. An alternative approach to change the threshold is dynamic thresholding, which can offer more flexibility. Instead of using a fixed threshold, dynamic thresholding adjusts the decision point for every input based on data characteristics, such as text complexity (Xu et al., 2021). This method allows the model to be more sensitive to the nuances within each category, enabling it to handle diverse writing styles more effectively. More specifically, the English levels of NNS are varied, using fixed decision threshold cannot represent all the NNS fairly (Xu et al., 2021). Future studies can focus on dynamically adjusting the threshold based on real-time feedback, which can better balance false positive rates and improve fairness across both NS and NNS. As a result, this methodology can generates a more refined and adaptable solution to mitigate the bias.

Removing sensitive features

Removing sensitive features is a bias mitigation technique often used in machine learning to ensure that models do not make decisions based on attributes that could lead to unfair outcomes. This method involves identifying and excluding features like race, gender or language proficiency that may disproportionately influence a model's predictions. For instance, the sensitive features in a model used for evaluating job candidates are age and gender (Zhu et al., 2023). To apply the removal of sensitive features in solving the bias in AI detectors towards NNS, Jiang et al. (2024b) conducted statistical tests to firstly identify which linguistic features showing significant differences between NS and NNS. They used Mann-Whitney U tests to compare the linguistic features between these two groups and determine which ones were significantly different. Once these sensitive features were identified, the effect of each was measured by using Cohen's D. After ranking these features based on their effect sizes, they removed the most sensitive features to observe how their removal impacted detection accuracy and bias in the model. The result was shown by Graph 1, as more sensitive features were removed, the overall accuracy started to decline. Both the FPR for NS and NNS showed little change initially but began to increase after about six features were removed (Jiang et al. 2024b). This indicates that while removing a few sensitive features does not make a huge difference in the result, removing too many will lead to a notable rise in false positives for both groups and a drop in accuracy, suggesting that this strategy is not effective for this specific application.



Graph 1. The change of Accuracy and False Positive Rate (FPR) among different numbers of removed features (Jiang et al., 2024b).

Unlike sensitive features such as age and gender in the evaluation of job candidates, the most crucial sensitive features in AI detectors are usage patterns, grammar and text structure, which also exhibit significant differences between AI-generated and human-writing content (Jiang et al., 2024b). More importantly, these sensitive features play an important role in model's ability to accurately classify texts. Therefore, model's capacity to correctly distinguish between AI-generated and human-written articles would be compromised without these features, which eventually leads to a significant drop in classification accuracy as Graph 1 shows. In the end, the removal of such core linguistic features poses a risk to the reliability of AI detectors. An alternative approach to reduce the impact of sensitive features is using Fairness-Constrained Optimization to limits the models' reliance on certain sensitive features rather than removing them entirely (T. Kamishima et al., 2011). In the studies of T. Kamishima et al. (2011), the model with Fairness-Constrained Optimization got 10% improvement in fairness in the task to predict the income with genders as sensitive feature. Thus, this method has been proved to be effective to mitigate the bias.

$$\min(L_{task}(y, f_\theta(x)) + \lambda I(s;\ f_\theta(x)) \quad (1)$$

The most important formula of Fairness-Constrained Optimization is shown in equation 1, it adds an additional term $I(s;\ f_\theta(x))$ to model's original loss, which quantifies how much information shares with $f_\theta(x)$. This value is higher if the model predictions are closely related to the sensitive feature 's'. Therefore, the model will get a higher loss, and then try to make the predictions less rely on 's' (T. Kamishima et al., 2011). In addition, since the sensitive features in the AI detection models play an important role in model's function, λ can helps the models to control the trade-off between the main task's accuracy and the fairness constrain. More specifically, a higher λ increases the model's focus on reducing $I(s;\ f_\theta(x))$, thus prioritizing fairness, while a lower λ allows the model to prioritize accuracy by focusing more on the original loss. In a word, Fairness-Constrained Optimization with an appropriate λ that reach the balance between fairness and accuracy is capable for improving model's fairness without sacrificing accuracy.

Balancing data

Training a model with a balanced dataset is essential for ensuring fairness and accuracy, especially in binary classification tasks. According to Wei & Dunbrack. (2013), balanced datasets can help models to learn more effectively by exposing them to equal representations of each class. When datasets are imbalanced, models can become practical towards the majority class, leading to curved performance metrics, such as high overall accuracy but poor performance for the minority class. In contrast, a balanced dataset enables better understanding for features across the different categories, thus improving models' accuracy and reliability in varied scenarios.

Jiang et al. (2024b) implemented this method to solve the bias by selecting an equal number of essays from both NS and NNS. They carefully chose 100 essays from both NS and NNS for each prompt, resulting in a total of 5,000 essays from each group. These human-written essays were then paired with an equal number of AI-generated essays for training and evaluating the AI detection model. This approach ensures equal representation of both groups, hopefully can reduces the bias during the model's training process and improves fairness in its predictions. As a result, the model trained with datasets that implemented balancing data strategy has a slightly lower FPRP value compared to the baseline models, where FPRP decreased for 0.21%. In addition, the FPRP for models with both balanced dataset and separate threshold is 0.22% lower than the regular models.

The findings indicate that while balancing the dataset helps to reduce bias, it is not sufficient to fully address it. The root cause lies in the variety of English proficiency among NNS. Some NNS, especially those who have lived in English-speaking environments for a long period, tend to have a stronger use of vocabulary and sentence structure compared to beginners. Consequently, models will fail to identify patterns typically used by beginner-level NNS and make misclassification, if they were trained predominantly on essays from advanced NNS. To address this issue, the dataset needs to be balanced not only between NS and NNS, but also across different English proficiency levels within the NNS group. One approach would be to classify NNS' writing into three proficiency levels, Beginner, Intermediate, and Advanced, and then sample an equal number of essays from each group to the dataset. This would

enable the model to learn the language features of NNS at varying levels of proficiency, ultimately leading to fairer and more inclusive decision-making.

Methods that have not been applied

Since the bias in the AI detectors towards non-native speakers is a relatively new topic and has not been recognized by everyone, some commonly used bias mitigating methods in general deep learning models have not gotten a chance to justify their effectiveness in this field.

Adversarial debiasing

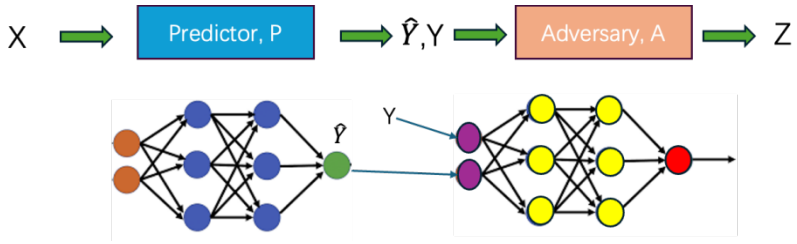
Adversarial debiasing is a machine learning approach proposed by Zhang et al. (2018), aims to reduce bias in models by forcing the model not to rely on sensitive attributes. Adversarial debiasing works by setting up two models competing against each other: the predictor model and the adversary model (Pagano et al., 2023). As equation 2 shows, the predictor model means to minimize the prediction loss L_p, which measures the performance for the main task and determines whether a text is AI-generated or human-written. In addition, $f_\theta(X)$ is the prediction from the model, Y is the actual label, and ℓ is a loss function**.**

$$L_p = \mathbb{E}_{(X,Y)}[\ell(f_\theta(X), Y)] \tag{2}$$

On the other hand, as equation 3 shows, the function of adversary model is to predict the sensitive attribute S, and its objective is to minimize its own loss L_a. $A_\emptyset(X)$ denotes the adversary's prediction of the sensitive feature, and S is the true sensitive attribute**.**

$$L_a = \mathbb{E}_{(X,S)}[\ell(A_\emptyset(X), S)] \tag{3}$$

These models are trained simultaneously in a game-like setting, where the predictor learns to make accurate predictions while trying to "fool" the adversary. More specifically, the predictor, aims to maximize L_a as part of its adversarial role and tiring to confuse the adversary. By doing so, the predictor removes biases related to the sensitive features from its predictions, leading to a fairer and less biased model. The whole process is shown by Graph 2, the **predictor** receives the input data X and produces a prediction $\hat{Y}$. Then, the predictor's output is fed into the **adversary**, which tries to detect sensitive attributes (e.g., linguistic background) from both the input data X and the predictor's output $\hat{Y}$ (Zhang et al. 2018). The adversary aims to maximize its ability to predict the sensitive attributes, while the predictor is trained to confuse the adversary by minimizing its predictive power. This adversarial training forces the predictor to focus on task-relevant features and reduce dependence on sensitive attributes, leading to fairer and less biased outcomes. What's more, Zhang et al. (2018) applied adversarial debiasing in a model used for analogy task (i.e., fill in the gender for occupations), where gender is the sensitive features. As the result, the model with adversarial debiasing has fewer biased analogies then baseline models, which suggest that adversarial debiasing can mitigate the bias in deep learning effectively.



Graph 2. Architecture for Adversarial debiasing (Graph credit: Original)

By directly targeting the linguistic features that can unfairly influence model predictions, adversarial debiasing will be an efficient technique to address bias in AI detectors towards NNS. NNS often use different sentence structures, simpler grammar or less diverse vocabulary, which can lead AI models to misclassify their writing as AI-generated. By training the predictor model to make accurate predictions while competing against an adversary that tries to detect sensitive attributes (e.g., native or non-native status), adversarial debiasing forces the predictor to reduce its reliance on these features. With these structures, adversarial debiasing will grant models a higher fairness level by intelligently reducing the influence of these features in a nuanced way. What's more, this method will not affect the models' ability to retain the useful aspects of linguistic features for the main task, which means models can still have an accurate but fairer results. In a word, it is worthy to apply adversarial debiasing in solving bias in AI detectors, because it offers a systematic way to reduce bias without sacrificing model performance.

Relaxed Equalized Odds Framework

The Relaxed Equalized Odds framework proposed by Pleiss et al. (2017) is a modification of the traditional Equalized Odds fairness criterion that further balancing fairness and calibration in machine learning models. Traditional Equalized Odds is a fairness framework in machine learning that requires a model to have the same FPR and false negative rate (FNR) across different demographic groups (Hardt et al., 2016). This ensures that errors made by the model are not disproportionately distributed among these groups, otherwise, penalties will be applied. This method was also proved to be effective by the studies of Hardt et al.(2016) where the performance across different races more balanced by 19%. As equation 4 shows, the fairness penalty encourages the model to reduce disparities in error rates. The overall loss for the model is shown in equation 5, the goal **minimizes the total loss** L_{total}. By adding the fairness penalty term $P_{fairness}$ and the strength of the fairness constraint λ, the model is not just optimizing for accuracy but is also being incentivized to reduce disparities in FPR and FNR across groups.

$$P_{fairness} = |FPR_A - FPR_B| + |FNR_A - FNR_B| \quad (4)$$

$$L_{total} = L_{prediction} + \lambda * P_{fairness} \quad (5)$$

While traditional Equalized Odds ensures fairness by balancing error rates, Pleiss et al. (2017) demonstrated that Equalized Odds conflicts with providing calibrated probabilities, making it generally impossible to achieve both simultaneously except in trivial cases. Therefore, Pleiss et al. (2017) introduced a relaxed version of Equalized Odds that allows for a trade-off between these two objectives. In the Relaxed Equalized Odds framework, instead of strictly requiring equal false positive and false negative rates across groups, the focus is on relaxing one of these error rates to maintain calibration, while

other parts remain the same. More specifically, once the $|FPR_A - FPR_B|$ less than a specific value, the Relaxed Equalized Odds framework will treat the difference in FPR as zero. With this modification, the Relaxed Equalized Odds framework allows the model to prioritize either FPR or FNR, depending on the context of the task, while ensuring that the predicted probabilities remain aligned with the actual likelihood of events.

For addressing bias in AI detectors toward NNS, minimizing FPR is generally more critical because it reflects the percentage of human-writings been misclassified a as AI-generated. By prioritize the FPR for in $P_{fairness}$, this framework allows models balance the numbers of misclassification for NNS and NS. This focus on minimizing FPR for NNS ensures that the model is actively correctting this bias without sacrificing the overall calibration of predicted probabilities. What's more, since this framework allows for a controlled relaxation of the FPR, the model will not overly penalize itself for small discrepancies in FPR between different groups. However, setting the right thresholds for relaxation might be challenging. If the threshold is too lenient, allowing FPR for NS and NNS to vary may fail to reduce significant biases, while a too strict threshold may causes the model to overly focus on minimizing FPR differences and sacrificing overall accuracy. All in all, this framework with suitable threshold settings ensures the models achieve balanced and nuanced fairness while maintaining predictive accuracy, offering a promising and efficient approach to mitigating bias in AI detectors.

Mitigating the Bias from Educational Aspects

Despite the effort made in technical aspects to solve the bias in AI detectors towards non-native speakers, some institutions also proposed solutions for this problem in educational aspects.

Human-in-the-Loop

Despite the technical methods, there are also some studies trying to mitigate the bias in AI detectors towards NNS from educational aspects. The U.S. Department of Education proposed a "human-in-the-loop" approach, where educators actively participate in AI decision-making rather than relying solely on AI outputs (US Government, 2023). This means human judgment is integrated into a continuous feedback loop where educators monitor, validate, and adjust AI's outputs. During this process, educators will review AI predictions during deployment, especially for flagged outputs like AI-generated work or plagiarism. More specifically, if an AI system classifies NNS' work as AI-generated due to linguistic differences, the educator needs to review the text and may correct unjust flags. In this pattern, educators also provide contextual understanding that AI lacks, considering individual student backgrounds, intent and writing styles. This oversight allows AI systems to benefit from automation while ensuring that biased outputs are mitigated, and fairness is maintained.

However, while Human-in-the-Loop can enhance the fairness, it has notable weaknesses. One significant drawback is its time-consuming nature. Since human reviewers must actively intervene in the decision-

making process, the speed and efficiency of AI systems are diminished. This becomes particularly problematic in large educational settings with a large number of tasks to review, such as grading or plagiarism detection, where requiring human reviewing each flagged output can slow down the overall process. Additionally, educators might unknowingly apply their own biases, especially when reviewing work from diverse cultural and linguistic backgrounds, leading to inconsistent or unfair decisions. This human subjectivity can undermine the neutrality that AI is supposed to bring to the evaluation process.

To address the weaknesses of the Human-in-the-Loop (HITL) approach, governments or higher education institutions can establish a specialized department focusing solely on reviewing AI-flagged texts, where the employees should be trained experts in linguistic diversity and AI ethics. By centralizing the review process, this department can relieve educators of the manual reviews, ensuring educators to devote more time in teaching. Additionally, it can serve as a feedback loop to improve AI systems over time. By using the insights from manual reviews to enhance AI accuracy and fairness, HITL approach can becomes more scalable and efficient. What's more, Turnitin's method can further improve the efficiency of this process. Instead of classifying the texts as AI-generated or human-written, the reports from Turnitin highlight the specific sentences or paragraphs in the writings that are likely to be AI-generated. Under this circumstance, human reviewers are able to merely focus on checking these highlighted parts rather than reviewing whole articles (Turnitin, 2024). Therefore, staffs' reviewing efficiency will be significantly improved.

Adjusting classification criteria

Turnitin further enhances its method to mitigate the bias in AI detectors towards non-native speakers by adjusting the classification criteria (Turnitin, 2024). To be more specific, Turnitin only presents the highlights if the overall percentage of AI-generated content in a paper exceeds 20%. If the AI-generated content is below this threshold, no score or highlights are shown to users, since Turnitin treats every score below this threshold to a false-positive value (Turnitin, 2024). Additionally, the threshold of 20% ensures that small instances of flagged text do not lead to unnecessary penalties, which might disproportionately affect NNS whose writing styles differ from those of NS. In a word, Turnitin's pattern allows for more nuanced review, while minimizing unnecessary impact for smaller AI-generated sections.

While this method can effectively reduce the bias, it also has weaknesses. Setting a 20% threshold may lead to misuse. More specifically, if a researcher uses AI to generate 18% of the article, the overall score for that article will still be unsurfaced since it is lower than the threshold. To further improve this strategy, Turnitin can implement a dynamic threshold based on the length and type of the document. For instance, longer research papers could have a lower threshold, such as 10%, to ensure that even smaller portions of AI-generated content are counted in to the overall length. This will help the models to prevent misuse where a researcher avoid penalty for using AI to generate a large part of text but under the 20% limit.

Conclusion

As discussed in this review, mitigating bias in AI detection tools towards non-native English speakers presents both technical and educational challenges. Current AI detectors disproportionately flag the work of non-native speakers as AI-generated, pushing these individuals to modify their writing styles in ways that can hinder their academic and professional growth. While technical solutions like adjusting thresholds, removing sensitive features and balancing data offer ways to address these biases, they come with limitations. These methods need to be enhanced with more inclusive training datasets and continuous oversight to create more balanced and fair detection systems.

While some bias-mitigation methods have already been explored, some other methods with high potential have not yet been applied in this specific field. Among all potential solutions, the Relaxed Equalized Odds Framework and Fairness-Constrained Optimization have a huge potential. However, finding the optimal parameters to balance accuracy and fairness is complex and time intensive. What's more, implementing dynamic thresholds to address bias in AI detectors for NNS possess considerable prospects, but it is also time-consuming and challenging, since it requires continuous recalibration based on diverse language patterns. In contrast, methods like collecting a more inclusive dataset that classifies non-native speakers' writing into three proficiency levels and implementing Adversarial Debiasing are more practical and effective, as they are straightforward and likely to yield quicker results, since they do not need the developers to find the optimized parameters for them. Therefore, future studies may benefit from prioritizing the creation of inclusive datasets and applying adversarial debiasing, as these approaches are likely to offer feedback and solutions sooner to address this urgent issue.

Furthermore, two methods proposed by educational institutions show promise and are recommended for implementation. Establishing a specialized department to train professionals focusing on reviewing Turnitin's AI reports would ensure fairer and more consistent assessments. Additionally, setting various detection thresholds for the articles with different lengths would allow AI systems to further avoid potential cheating. Since both approaches are feasible and beneficial, start working on them can mitigate the bias in AI detection systems effectively. Overall, solving the bias in AI detectors will require a combination of technological advancements and educational reforms to ensure fair and inclusive practices.

Reference

Esposito, C., Landrum, G. A., Schneider, N., Stiefl, N., & Riniker, S. (2021). GHOST: Adjusting the Decision Threshold to Handle Imbalanced Data in Machine Learning. *Journal of Chemical Information and Modeling*, *61*(6), 2623–2640. https://doi.org/10.1021/acs.jcim.1c00160

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *CoRR*, *abs/1610.02413*. http://arxiv.org/abs/1610.02413

Jiang, Y., Hao, J., Fauss, M., & Li, C. (2024b). Towards Fair Detection of AIGenerated Essays in LargeScale Writing Assessments. In A. M. Olney, I. Chounta, Z. Liu, O. C. Santos, & Bittencourt, Ig Ibert (Eds.), *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky* (pp. 317–324). Springer Nature Switzerland.

Pagano, T. P., Loureiro, R. B., Lisboa, F. V. N., Peixoto, R. M., Guimarães, G. A. S., Cruz, G. O. R., Araujo, M. M., Santos, L. L., Cruz, M. A. S., Oliveira, E. L. S., Winkler, I., & Nascimento, E. G. S. (2023). Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods. *Big Data and Cognitive Computing*, *7*(1), 15. https://doi.org/10.3390/bdcc7010015

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J. M., & Weinberger, K. Q. (2017). On Fairness and Calibration. *Neural Information Processing Systems*, *30*, 5680–5689.

T. Kamishima, S. Akaho, & Sakuma, J. (2011). Fairnessaware Learning through Regularization Approach. 2011 IEEE 11th International Conference on Data Mining Workshops, 643–650. https://doi.org/10.1109/ICDMW.2011.83

Turnitin. (2024). *AI Writing Detection | AI Tools | Turnitin*. Www.turnitin.com. https://guides.turnitin.com/hc/en-us/articles/28294949544717-AI-writing-detection-model

US Government. (2023, May 24). *U.S. Department of Education Shares Insights and Recommendations for Artificial Intelligence*. U.S. Department of Education. https://www.ed.gov/about/news/press-release/us-department-of-education-shares-insights-and-recommendations-for-artificial-intelligence

Wei, Q., & Dunbrack, R. L. (2013). The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics. *PLoS ONE*, *8*(7), e67863. https://doi.org/10.1371/journal.pone.0067863

Xu, Y., Shang, L., Ye, J., Qian, Q., Li, Y.-F., Sun, B., Li, H., & Jin, R. (2021). *Dash: Semi-supervised learning with dynamic thresholding* (M. Meila & T. Zhang, Eds.; Vol. 139, pp. 11525–11536). PMLR. https://proceedings.mlr.press/v139/xu21e.html

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. *CoRR*, *abs/1801.07593*. http://arxiv.org/abs/1801.07593

Zhu, H., Dai, E., Liu, H., & Wang, S. (2023). Learning fair models without sensitive attributes: A generative approach. *Neurocomputing*, *561*, 126841. https://doi.org/10.1016/j.neucom.2023.126841