Analyzing Bias Mitigation in Automated Decision Systems

Tanya Naheta
NYU Center for Data Science
tsn6109@nyu.edu

Gordon Chan
NYU Center for Data Science
ghc260@nyu.edu

Abstract

This report examines the effectiveness of an Automated Decision System (ADS) designed for toxicity detection in online platforms, focusing on its ability to manage bias and ensure fairness across diverse user demographics.

1 Background

In the digital world that we live in today, technological innovation and the advent of the Internet have made it possible to communicate online with anyone, anywhere, and anytime. At the same time, this has made it easier than ever for any person to anonymously spread toxicity online without consequences, making the detection of toxicity an increasingly important part of moderating online content. However, initial toxicity models built by Conversation AI, a research initiative jointly co-founded by Jigsaw and Google, would incorrectly learn to associate the names of frequently attacked sensitive groups with toxicity, predicting a high likelihood of toxicity for comments containing words such as "gay", even when the comment was not toxic. As noted by Buolamwini and Gebru, such biases can lead to intersectional accuracy disparities in commercial applications [3].

Addressing such biases in toxicity models is particularly important, as imbalances in the training data can reflect pre-existing biases that may eventually lead to potential emergent bias if such biased models are used to identify and filter toxicity. This is significant as toxic or hate speech is almost always directed towards sensitive groups, and it is those belonging to such minority groups who are negatively impacted the most, as not only do such toxic comments perpetuate and reinforce societal prejudices against them, but may even encourage toxic or hateful behavior towards minority groups in the real world. As such, the purpose of the ADS we chose to audit was to not only be able to identify toxicity but to do so while minimizing the bias to mentions of sensitive groups. However, these goals introduce a potential tradeoff between minimizing bias and maximizing the model's overall accuracy. Furthermore, due to the inclusion of identities from various sensitive groups, the tradeoffs become more complex and intricate as balancing the bias and accuracy for certain groups may affect the balance of other groups. Although in general, we would expect that efforts to minimize bias should correspond with a decrease in accuracy, in this particular context removing such unintended biases may in fact improve the model's overall accuracy if the false positive rate can be lowered enough.

Showing no signs of slowing down or going away, toxicity in the online world will only continue to increase and become more widespread as time passes and technology improves. With such toxicity often being targeted at sensitive groups that already have to face prejudice and discrimination in real life, not only is it vital to have models that can detect such toxicity to help with content moderation, but it is perhaps even more important to ensure that such models do not learn unintended biases from the data, to prevent them from exacerbating the problem instead of helping.

Preprint. Under review.

2 Input and output

2.1 Dataset Description and Collection

The dataset utilized by the ADS originates from the Civil Comments platform, consisting of approximately 2 million public comments spanning from 2015 to 2017 across various English-language news sites. After the platform's closure, the dataset was augmented by Jigsaw, who sponsored the annotation of these comments to identify various forms of toxicity and mentions of identity to deepen insights into conversational civility. This dataset was chosen due to its broad coverage and diversity, which are critical for research into online toxicity, aligning with the need for comprehensive datasheets as recommended by Gebru et al. [4].

2.2 Input Features and Data Profiling

Data Types and Description: The primary input feature, `comment_text`, is textual data. Supplementary features include toxicity scores (target) and several identity labels such as male, female, black, etc.

Missing Values and Distribution: There is a significant prevalence of missing data, particularly in identity labels, with around 77.55% data missing across these categories. The toxicity scores are heavily skewed towards non-toxic comments, which could influence model training by underrepresenting toxic data points.

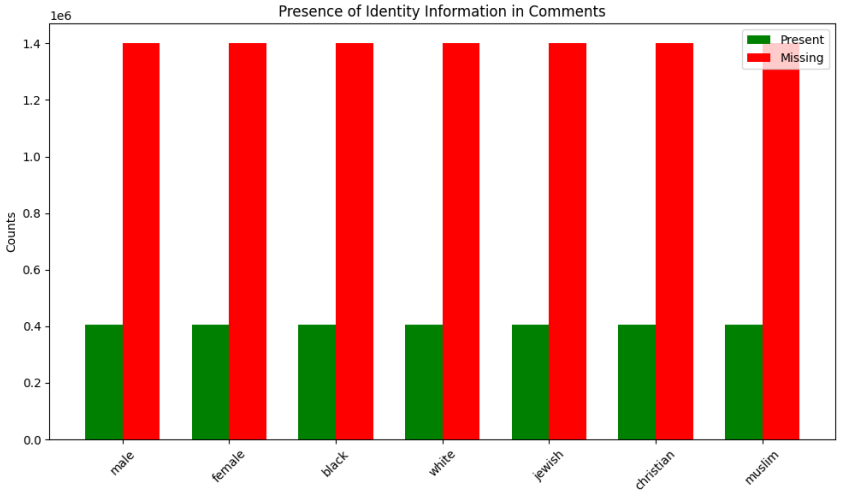

Figure 1: Presence of Identity Information in Comments

Correlation Analysis: Strong correlations were noted between target and subtypes like insult and obscene, suggesting these features are closely linked to overall toxicity perceptions. The correlation matrix also highlighted potential biases where certain identities might correlate with higher perceived toxicity.

The correlation matrix indicates strong associations between target and toxicity subtypes like insult and obscene, suggesting these features are critical indicators of overall toxicity. The matrix also shows correlations between identity mentions and toxicity, which could inform bias mitigation strategies in the ADS.

2.3 System Output and Interpretation

The output of the ADS is a continuous probability score ranging from 0 to 1, representing the likelihood of a comment being perceived as toxic. The ADS employs an ensemble of neural network models, including LSTMs and BERT-based models, to capture a wide range of linguistic features and nuances. The system's validation involves a robust set of metrics, including accuracy, precision, recall, and F1-score, ensuring it performs reliably across all demographics without bias [7].This score

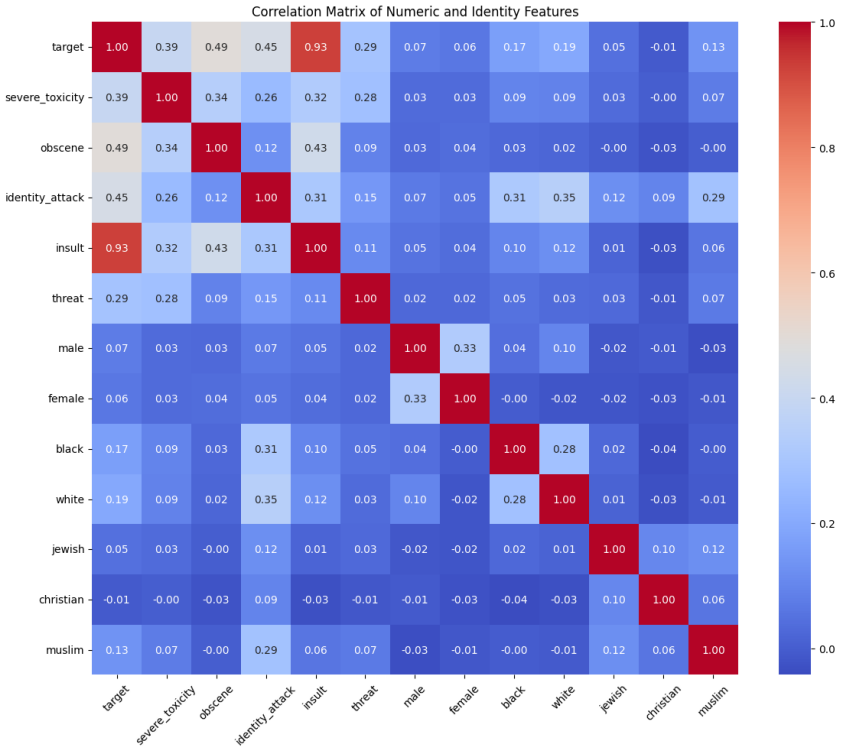

Figure 2: Correlation Matrix of Numeric and Identity Features

facilitates nuanced moderation strategies by providing a gradient measure of toxicity rather than a binary classification, allowing for tailored responses to different levels of toxicity.

3 Implementation and validation

The ADS employs an ensemble of neural network models, including LSTMs and BERT-based models, to capture a wide range of linguistic features and nuances. The system's validation involves a robust set of metrics, including accuracy, precision, recall, and F1-score, to ensure it performs reliably across all demographics without bias.

Our approach to fairness is informed by Barocas et al.'s discussion on fairness and abstraction in sociotechnical systems [2], and we assess model fairness using metrics such as equality of opportunity as defined by Hardt et al. [5].

3.1 Data Cleaning and Pre-processing

Pre-processing involves handling missing data, especially in identity-related columns, potentially using techniques such as imputation or exclusion of incomplete records. Text data were pre-processed through standard NLP techniques including tokenization, normalization (e.g., converting to lowercase), and removal of extraneous characters to prepare the data for model ingestion. Further steps include normalizing Unicode characters to standardize text data, handling obscenities by converting starred words to a uniform representation, and eliminating non-Unicode characters which includes accents and other diacritical marks that could affect text processing.

There are also updates made to the language-specific characters in the data. For instance, all Hebrew characters are replaced with a common Hebrew letter, which reduces the diversity of characters and simplifies the model's processing requirements. Similar approaches are applied to Arabic, Chinese, and Japanese characters. Additionally, emojis are replaced with their textual aliases plus a special token indicating the presence of an emoji, which helps the models to recognize emotional and contextual nuances conveyed by these symbols.

3.2 High-Level System Implementation

The ADS is implemented as an ensemble of various models, strategically combining the strengths of 12 LSTM-based models, 17 BERT-based models, and 2 GPT-2-based models. This ensemble method was chosen to improve model robustness and accuracy by leveraging diverse architectural strengths and reducing the risk of bias inherent in any single model approach.

LSTM Models: These models utilize a bidirectional architecture with two LSTM layers and incorporate character-level embeddings. A unique feature of these models is the pooling operation from both LSTM layers to enrich the feature set before it reaches the dense classifying layers. This diversity is further expanded by using different sets of word embeddings and initializing models with different random states.

BERT Models: All BERT models, including variations like cased, uncased, and fine-tuned uncased, are used to understand contextual nuances in the text. These models benefit from an optimized padding strategy that reduces inference time significantly by adjusting to the longest comment in each batch.

GPT2 Models: These models are adapted with a CNN-based classifier head instead of the typical linear classifier head, which is intended to capture more complex patterns in the data.

3.3 Validation and Performance Metrics

The validation of the ADS was carried out using metrics such as Overall AUC and Bias AUCs across different identity groups to ensure the model performs well across diverse scenarios. The Overall AUC provides a measure of the model's accuracy, while Bias AUCs help assess fairness by determining how well the model identifies toxicity across comments mentioning various identities. This dual metric approach ensures that the ADS is not only accurate but also fair, reducing unintended bias.

4 Outcomes

4.1 Performance Analysis Across Subpopulations

The Automated Decision System (ADS) designed for toxicity detection was assessed for accuracy using Area Under the Receiver Operating Characteristic Curve (AUC), precision, recall, and F1 score. These metrics were chosen based on their relevance to classification problems where both the presence and absence of bias can significantly impact the model's utility.

AUC: This metric is utilized as a primary indicator of the model's capability to differentiate between the classes (toxic and non-toxic) under various threshold settings. The model achieved a high AUC value that indicates very good model performance, with the overall AUC being 0.975. However, subgroup-specific AUCs showed some variation (0.956 for Sex, 0.939 for Religion, and 0.912 for Race), indicating a decrease in discriminatory power for these subgroups compared to the overall results.

Metric	**Value**
Overall AUC	0.975293
Sex - Generalized Mean of Bias AUCs	0.956823
Religion - Generalized Mean of Bias AUCs	0.939337
Race - Generalized Mean of Bias AUCs	0.912179
Final Model Score	0.945908

Table 1: Final AUC Scores

The AUC values indicate a high overall capability with variances across subgroups, suggesting areas for model improvement. Discrepancies in subgroup AUCs could indicate potential biases in how the model processes data from different demographics.

These figures highlight the model's challenges in balancing precision and recall, critical in contexts where both false positives and negatives bear significant consequences.

Group	Metric	Value
Overall	AUC	0.975293
Overall	Precision	0.0815522
Overall	Recall	0.999871
Overall	F1	0.150804
Sex	AUC	0.944224
Sex	Precision	0.174168
Sex	Recall	0.998489
Sex	F1	0.296506
Religion	AUC	0.914865
Religion	Precision	0.175145
Religion	Recall	1
Religion	F1	0.295908
Race	AUC	0.866566
Race	Precision	0.325757
Race	Recall	1
Race	F1	0.491031

Table 2: Accuracy Metrics

Precision and Recall: These metrics are crucial in contexts where the cost of false positives and false negatives are substantial. For toxicity detection, a high recall should be considered vital, as the failure to identify toxic comments can have severe repercussions. Nonetheless, that needs to be balanced with the system's overall precision, which was measured at 0.08 and indicates a high rate of false positives. This means that the model output could lead to excessive moderation and potentially stifle free expression.

F1 Score: This metric balances the precision-recall trade-off and is particularly useful in skewed datasets or where both types of classification errors are costly. The relatively low F1 scores across all subgroups signify that the model struggles to balance precision and recall effectively.

The accuracy assessment using these metrics has revealed that while the ADS performs well overall, its effectiveness varies slightly across different subgroups. In order to resolve this issue, some potential enhancements could include diversifying the training data or implementing more nuanced loss functions that weigh subgroup performance differently. This analysis also highlights the importance of maintaining a balance between detecting toxicity and preserving the nuances of human communication.

4.2 Fairness Analysis

The fairness of the ADS is analyzed using a suite of metrics that evaluate the impact of the system on different subgroups, thereby ensuring that it does not perpetuate any biases. Demographic Parity and Equal Opportunity metrics are especially critical for assessing the fairness of our system, following frameworks established by Kusner et al. for counterfactual fairness [6] and the ongoing importance of monitoring machine bias as discussed by Angwin et al. [1].

Demographic Parity (DP): This metric measures whether decision outcomes are independent of the sensitive attribute. There were negative DP values for Sex and Religion indicating that these groups are less likely to be predicted as toxic compared to their proportions in the population. This suggests an under-detection bias, where toxic behaviors in these subgroups might be underreported by the ADS.

Equal Opportunity (EO): This is focused on equality in the true positive rates. EO assesses if all subgroups have an equal chance of being correctly identified when the actual condition is positive. The minimal differences close to zero across all groups imply that the ADS is equally likely to correctly identify true positives among all subgroups.

Equalized Odds: Extends EO by also considering false positive rates. There were disparities in FPRs, especially for the Sex subgroup, pointing to potential biases where this group might be over-classified as toxic.

Disparate Impact (DI): This metric is measured as a ratio, where a value close to 1 indicates fairness, while deviations suggest potential discrimination. For instance, the DI value measured for the Sex subgroup (0.8296) means that a lower likelihood of predicting comments as toxic for this group compared to others, which signifies an underlying bias in model predictions.

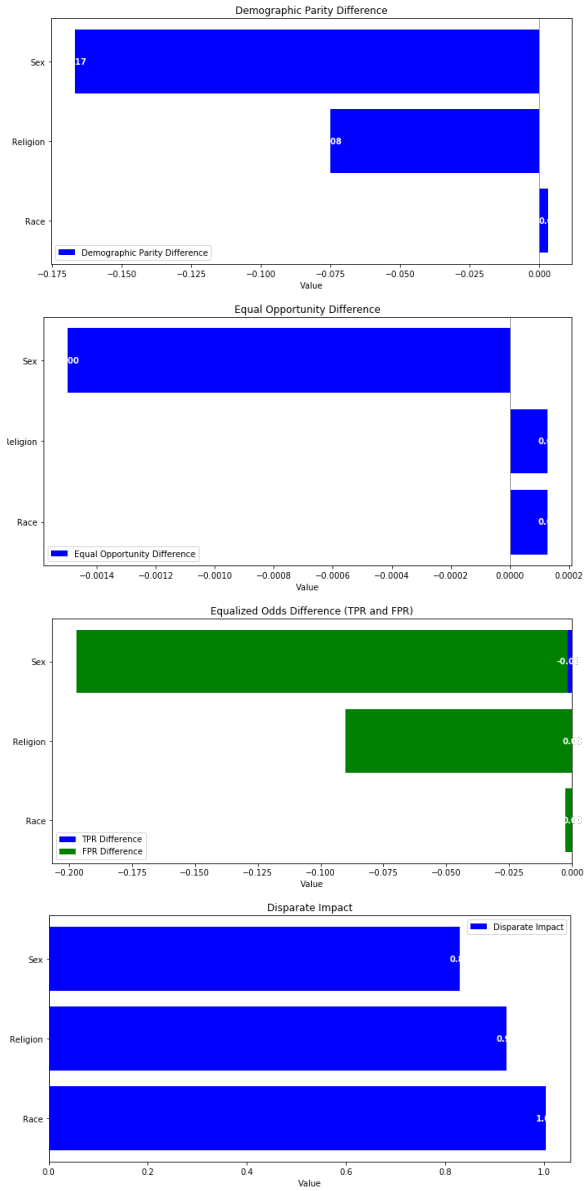

Figure 3: Plot of different fairness metrics across subgroups

The graphical representations of Demographic Parity and Equal Opportunity highlight disparities that suggest an under-detection of toxicity within certain subgroups, requiring further model tuning to address these biases. The nuances of Disparate Impact and the balance of True Positive and False Positive Rates is essential for understanding how the ADS performs in distinguishing toxic from non-toxic comments across different identities.

The identified disparities above necessitate interventions to achieve equity across all demographic groups. This analysis is crucial for the improvement of the ADS, ensuring it adheres to ethical standards and remains a reliable tool in toxicity detection across diverse populations.

4.3 Additional Performance Analysis Methods

Further methods to enhance the ADS robustness include stress testing, robustness checks, and performance evaluations on edge cases. Monitoring for concept drift ensures the ADS adapts to evolving language use and societal norms, maintaining relevance and effectiveness.

4.3.1 System Stress Testing and Robustness

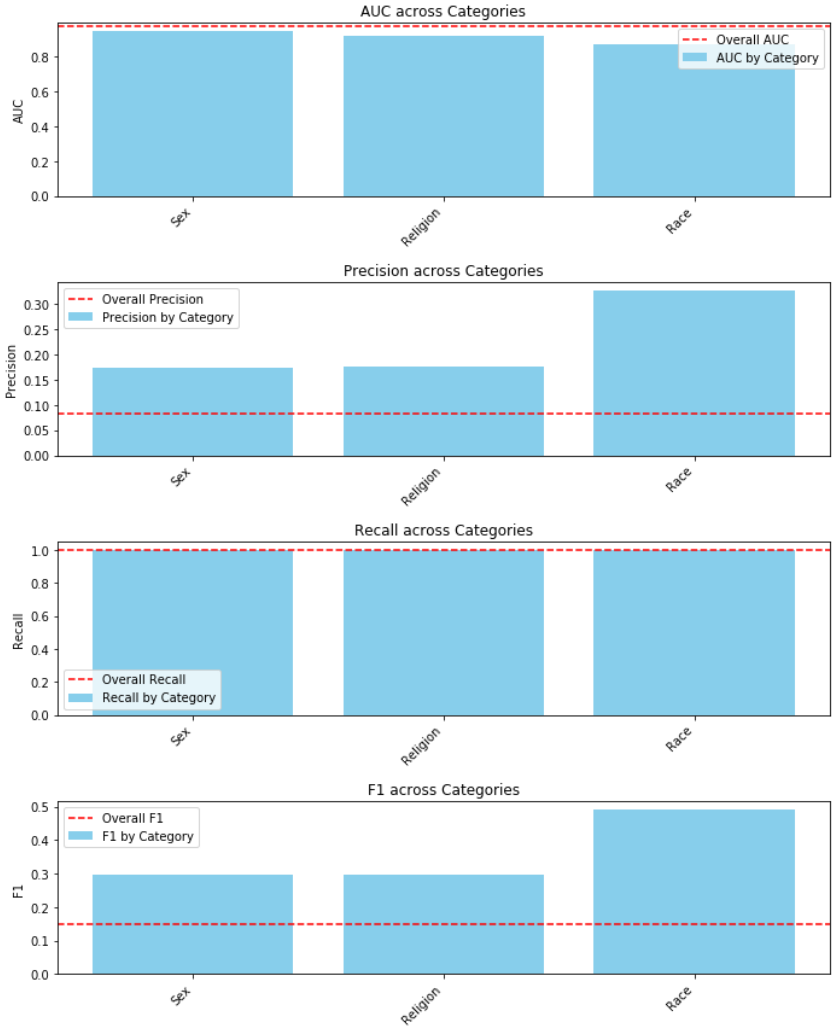

Figure 4: Performance metrics under varied conditions illustrating system resilience and adaptability

This testing ensures that the ADS can handle real-world complexities, maintaining high performance and fairness under dynamic conditions. To enhance the robustness of our system, we consider the impacts described by O'Neil in "Weapons of Math Destruction", highlighting the need for careful, ethical use of machine learning technologies [8].

Stress Testing: This method would evaluate how the ADS performs under extreme conditions, such as sudden influxes of data or inputs featuring complex linguistic nuances. It would be really useful in understanding the limits of the system's resilience and its capacity to handle difficult content without degradation in performance. In order to implement this, we could feed the system with complex data streams to observe how well it maintains functionality.

Robustness Checks: This method would involve introducing slight variations in input data to see if small changes can disproportionately impact the outcomes. For instance, minor alterations in the

wording of inputs or the inclusion of noise would test the ADS's ability to deliver consistent results, highlighting potential vulnerabilities or overfitting issues.

Performance on Edge Cases: This would involve testing the model on subtle or complex cases such as sarcasm or culturally specific idioms. These instances are often the most challenging parts of toxicity detection and can reveal significant insights into the model's nuanced understanding and any inherent biases.

Monitoring for Concept Drift: This method would implement continuous monitoring to detect shifts in the data landscape over time. As societal norms and language evolve, the ADS must adapt to maintain its fairness and accuracy. Monitoring for concept drift would ensure that the system remains effective and relevant by adjusting to new patterns and contexts in the data it processes.

These methods are recommended based on the necessity to ensure that the ADS not only performs well under ordinary circumstances but also maintains high performance and fairness levels in dynamic, unpredictable real-world scenarios. By incorporating these additional strategies, the ADS can be better prepared to handle the complexities and variability of real-world applications, ensuring its effectiveness and reliability.

5 Summary

The data used was carefully selected to train the system on a wide array of toxic behaviors and identity mentions. However, the variability in performance across different subpopulations suggests that while the data was generally appropriate, there is room for improvement in ensuring it encompasses a broader spectrum of interactions that more fully represent the minority subgroups. This would help in reducing bias and enhancing the model's sensitivity and specificity across all the segments.

5.1 Robustness, accuracy, fairness

The implementation of the ADS has shown robustness and high accuracy in general scenarios, as evidenced by high AUC scores. The choice of accuracy measures like AUC, precision, recall, and F1 score was selected to provide a comprehensive evaluation of the model's performance in classifying toxic and non-toxic comments. These metrics collectively offer insights into the ADS's ability to correctly identify toxic content while minimizing misclassifications.

Despite high overall accuracy, the ADS has demonstrated relatively low precision. This imbalance implies that while the ADS is effective at detecting potential toxicity (high recall), it also produces a significant number of false positives (low precision). In real-world applications, this could lead to excessive content moderation, where non-toxic comments are mistakenly flagged as toxic. Such over-moderation could stifle free speech or lead to user dissatisfaction if their content was incorrectly moderated, making it crucial for further refinement of the model to better balance recall and precision.

Fairness measures including Demographic Parity, Equal Opportunity, Equalized Odds, and Disparate Impact were used in evaluating whether the ADS operates without bias. These revealed important insights about the system's performance across different subgroups. Notably, negative values in Demographic Parity for the Sex and Religion subgroups suggest an under-detection bias, which indicates that these groups are less likely to be identified as toxic compared to their actual presence in the data. Moreover, disparities in false positive rates, particularly for the Sex subgroup, highlight a potential over-classification bias which could lead to excessive moderation.

The importance of these accuracy and fairness measures becomes even more pronounced when considering deployment in sensitive environments like social media, forums, or customer interaction platforms. Ensuring that the ADS is both accurate and fair is crucial for stakeholders who rely on this technology to maintain safe and inclusive online environments.

Stakeholders such as social media platform maintainers, users, civil rights groups, and regulatory bodies may find these measures particularly relevant. For social media platform maintainers and users, ensuring accuracy and fairness is very important in order to maintain trust and a positive user experience. For civil rights groups and regulatory bodies, these measures verify compliance with legal and ethical standards set in society.

In summary, while the ADS's implementation is robust and largely effective, the noted discrepancies in precision and fairness across subgroups highlights the need for adjustments and improvements. These would not only enhance the system's reliability but also bolster its acceptability and effectiveness in diverse, real-world applications.

5.2 Leveraging the ADS in the public sector

Deploying this ADS in the public sector or industry would require careful consideration. While the system has shown promising results, the noted disparities in performance across different subpopulations and the occurrence of false positives suggest that further tuning is needed before full deployment. Stakeholders would need to consider the potential social implications of incorrect toxicity classifications, which could affect user interactions and trust. Therefore, while the system holds potential, a cautious approach would be recommended, possibly starting with a test in controlled environments where the system's outputs can be monitored and moderated by human overseers.

5.3 Recommended ADS Improvements

Diversification of Training Data: Incorporating a wider variety of data sources, including more nuanced and context-rich examples, as well as more data points that represent the minority subgroups would help in training the models to better understand the numerous nuances.

Enhanced Monitoring for Bias: Implementing continuous monitoring mechanisms to assess and adjust for biases that may develop as language and social norms evolve.

Model Tuning: Refining model parameters and exploring more sophisticated techniques that could improve precision and lower unnecessary content moderation is critical, especially in scenarios where this model could run without any human oversight.

These enhancements would not only improve the performance and fairness of the ADS but also bolster its reliability and acceptance among all stakeholders.

References

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. ProPublica. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, 2016.

[2] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 59–68, 2019.

[3] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

[4] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[5] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[6] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

[7] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229, 2019.

[8] Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown, New York, 2016.