Centering the Margins: Outlier-Based Identification of Harmed Populations in Toxicity Detection

Vyoma Raman* UC Berkeley vyoma.raman@berkeley.edu

Eve Fleisig* UC Berkeley efleisig@berkeley.edu

Dan Klein UC Berkeley klein@berkeley.edu

Abstract

A standard method for measuring the impacts of AI on marginalized communities is to determine performance discrepancies between specified demographic groups. These approaches aim to address harms toward vulnerable groups, but they obscure harm patterns faced by intersectional subgroups or shared across demographic groups. We instead operationalize "the margins" as data points that are statistical outliers due to having demographic attributes distant from the "norm" and measure harms toward these outliers. We propose a Group-Based Performance Disparity Index (GPDI) that measures the extent to which a subdivision of a dataset into subgroups identifies the subgroups facing increased harms. We apply our approach to detecting disparities in toxicity detection using the Perspective API. We find that text targeting outliers is 28% to 86% more toxic for all types of toxicity examined. We also discover that model performance is consistently worse for demographic outliers, with disparities in error between outliers and non-outliers ranging from 28% to 71% across toxicity types. Our outlier-based analysis has comparable or higher GPDI than traditional subgroup-based analyses, suggesting that outlier analysis enhances identification of subgroups facing greater harms. Finally, we find that minoritized racial and religious groups are most associated with outliers, which suggests that outlier analysis is particularly beneficial for identifying harms against them.

1 Introduction

Society often erects barriers that hinder marginalized individuals from receiving essential social and infrastructural access. Disability studies offers valuable insights into their experiences, illuminating the oppressive and restrictive nature of the construction of "normalcy" within society. This

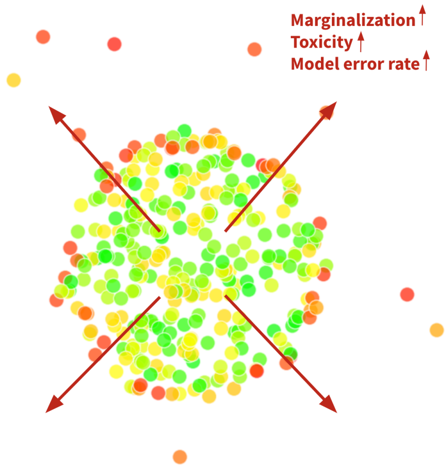

Figure 1: Outliers on the basis of demographic identity face harms resulting from high model error compared to non-outliers. Further analysis of the attributes of demographic outliers can reveal the specific subgroups experiencing harm.

rich body of literature highlights how prevailing social structures marginalize certain groups and further perpetuate their exclusion from mainstream societal participation. Extending the scope of these critiques to the realm of artificial intelligence, we recognize that AI models also encode prevailing notions of normalcy. When models optimize for an aggregate metric, they prioritize patterns and distributions for data points with more common characteristics; that is, the "norm" of the dataset. As such, individuals who fall outside the normative boundaries of the training data are more likely to be subject to model error and consequent harm. The lack of representation and tailored accommodation for these marginalized groups within the training data contributes to biased and exclusionary AI models.

In the evolving landscape of powerful tools like artificial intelligence, it is crucial to critically evaluate their application to avoid reinforcing systemic biases and instead promote equitable outcomes. Al-

*Eve and Vyoma co-created the theoretical framework for this paper, and Vyoma implemented it.

gorithm auditing plays a vital role in assessing the real-world impact of AI, especially in identifying and scrutinizing potential harm to specific "protected" demographic subgroups and their intersections (Mehrabi et al., 2021). However, the current subgroup-based analysis used in algorithm fairness evaluations is fraught with challenges. Two notable concerns emerge: First, identifying the relevant subgroups of concern is not always straightforward (Kallus et al., 2022), as there may be hidden or overlooked patterns in how populations are affected by algorithmic harms. For example, dividing into individual disability subgroups may conceal shared harms experienced by people with different types of disabilities. Second, when considering intersectional subgroups across multiple demographic categories, including race, gender, sexual orientation, disability, religion, and age, the sheer number of potential subgroups becomes overwhelming while the size of each subgroup decreases; thus, particularly severe or unique harms faced by many smaller, intersectional subgroups may be overlooked (Kong, 2022). These limitations make it challenging to conduct thorough and accurate audits.

Inspired by disability studies's argument that those who fall outside the norm experience greater adversity, we propose using outlier detection to statistically determine who is assumed to be "normal" and analyze algorithmic harms with respect to this boundary in the domain of algorithmic content moderation. By introducing a Group-Based Performance Disparity Index (GPDI) that quantifies the extent to which a disaggregated analysis uncovers harm, we demonstrate that disaggregating harm by outlier group membership reveals high model error disparities compared to other schemes for breaking down the data. Additionally, we identify discrepancies in toxicity detection model performance between outliers and non-outliers and analyze these groups to determine the demographic makeup of those experiencing the most harm. This approach allows for the identification of subgroups and intersections thereof that are particularly vulnerable to harm in the model's behavior.

This paper presents three primary contributions to advance our understanding of algorithmic harm towards marginalized groups in toxicity detection:

1. We propose and implement a method for identifying marginalized groups at risk of AI-induced harm using outlier detection techniques and a unique model harm identification framework.

2. We introduce the Group-Based Performance Disparity Index (GPDI), a metric for quantifying disparities in AI model performance. GPDI offers a concrete method to pinpoint groups experiencing the most harm, and we use it to validate our focus on outlier groups.

3. We examine model performance disparities between outliers and their complements to highlight the detection of severe toxicity and identity attacks as areas of concern and identify demographic subgroups that are particularly susceptible to harm from disparities in toxicity predictions.

Our work underscores the importance of incorporating critical theory into auditing practices in artificial intelligence models, specifically in toxicity detection, to ensure they do not exacerbate societal biases or inadvertently harm marginalized communities. The methodologies and tools we present serve as resources for those seeking to create more equitable and inclusive AI systems.

2 Prior Work

We ground our work on prior research on algorithmic content moderation and its potential harms, the measurement and evaluation of these harms, and the relationship between outlier detection and the social construction of the "other."

2.1 Harms in Content Moderation

Advances in machine learning and natural language processing (NLP) have enhanced the efficiency and scalability of moderating user content on online platforms (Gorwa et al., 2020; Gillespie, 2018), but this progress has been shadowed by the emergence of critical challenges and harms. Significant among these are allocative harms, which occur when content moderation decisions disproportionately amplify or suppress content by or about specific groups. These can also veer into representational harms, which involve the systematic silencing or misrepresentation of marginalized groups (Crawford, 2017). For example, toxicity detection algorithms have disproportionately flagged content from minority communities (Hutchinson et al., 2020) and also failed to adequately address hate speech targeted at these groups (Binns et al., 2017).

In toxicity detection settings, classifiers are prone to label dialects like African American English as abusive more often, instigating a discriminatory effect in content moderation (Davidson et al., 2019; Sap et al., 2019). A different challenge lies in detecting implicit hate speech, which frequently manifests through coded or indirect language (ElSherief et al., 2021; Waseem et al., 2017).

The academic community has developed various approaches to comprehend and address these issues in the face of these multifaceted challenges. Notably, Sap et al. (2020) proposed Social Bias Frames, a conceptual formalism that attempts to understand and capture the subtle ways social biases and stereotypes are projected in language. Furthermore, to standardize the testing of robustness and fairness in models, researchers have designed benchmark datasets, such as the HateXplain dataset, to facilitate the reduction of unintended bias (Mathew et al., 2021).

2.2 Measuring Algorithmic Harms

The measurement of harm caused by content moderation AI has been approached from varied theoretical perspectives, including discussing the moral implications of algorithmic decision-making, ensuring legal compliance with anti-discrimination laws, and computationally quantifying fairness with audits that systematically test the model's response to different inputs (Mittelstadt et al., 2016; Barocas and Selbst, 2016; Sandvig et al., 2014).

Researchers have proposed fairness metrics that capture different aspects of algorithmic performance, such as demographic parity, equalized odds, and equal opportunity (Dwork et al., 2012; Hardt et al., 2016), but these have been criticized for failing to account for underlying differences between groups (Corbett-Davies and Goel, 2018). These metrics often emphasize disaggregating analyses into protected subgroups to highlight potential disparities and biases that may not be evident in aggregate performance measures (Barocas et al., 2021; Buolamwini and Gebru, 2018). This focus on disaggregated analysis has enabled a more nuanced understanding of algorithmic harms and informed the development of fairer and more inclusive AI systems.

2.3 Qualitative and Quantitative Representations of the "Other"

Disability studies provides a valuable perspective on understanding the social construction of the "other." Scholars in this field have argued that disability is a social phenomenon as well as a biological category (Shakespeare, 2006). They contend that all bodies and minds are part of a spectrum of natural human diversity and that the distinction between disabled and nondisabled arbitrarily divides "normal" and "abnormal" embodiment and behavior. In this way, society partitions natural variation and diversity into categories (Davis, 2014). This approach explicitly grapples with the construction of social categories distinguishing an in-group ("us") from an out-group ("them"), often favoring the dominant group and marginalizing the "other" (Said, 1988). These principles apply to various marginalized groups, including racial and ethnic minorities and gender and sexual minorities, in addition to the disabled (Goffman, 1963; Butler, 1990).

Alongside these social constructs, the concept of "normalcy" emerged, heavily influenced by the rise of statistical methodologies (Davis, 1995). The advent of statistics paved the way for tools that reified the distinction between normal and abnormal. This was seen in early applications like IQ tests (Fass, 1980) and phrenology (Twine, 2002), which attempted to classify individuals based on certain traits or characteristics. Modern statistics has further advanced these capabilities through the notion of outliers, observations in a dataset that deviate significantly from the norm. Common outlier detection methods include techniques involving standard deviation, the interquartile range, and clustering algorithms like Local Outlier Factor (Breunig et al., 2000). Outlier detection involves quantitative methods to identify deviations from the norm, which has the potential to further solidify the dichotomy between the "norm" and the "other".

By applying outlier detection to demographic information, it is possible to identify minoritized demographic subgroups that represent marginalized people. Groups that are further from the "norm" face greater societal harms, and those that are less represented in data may similarly face greater allocative harms due to poor performance of a model on data points it had less exposure to during training.

3 Methods

Current quantitative methods for determining the impact of content moderation systems on vulnerable populations rely on identifying demographic subgroups represented in the dataset and examining the model behavior toward these groups. However, this approach has two significant barriers: First, if researchers are unaware of a particular group that might be affected, they cannot specify it for measuring potential harms using established methods such as fairness metrics. This can lead to unaddressed biases and unintended consequences in AI systems. Second, intersectional harms can be particularly acute and are crucial to measure. This has typically involved a disaggregated analysis of individuals along different demographic categories like race and gender (Buolamwini and Gebru, 2018). However, as more and more demographic categories are considered, the number of subgroups increases exponentially and their size decreases exponentially, making it challenging to identify and address specific problem areas.

In both situations, researchers risk missing significant harms that marginalized groups may experience. We propose a model harm identification framework to address these limitations and determine which groups and subgroups experience poorer model performance. This approach aims to facilitate a more comprehensive understanding of algorithmic harms and inform the development of fairer and more inclusive AI systems.

3.1 Theoretical Framework

Disability studies provides valuable insights into the experiences of marginalized individuals, highlighting that those on the fringes of society are more likely to face adversity due to social structures that are not designed to accommodate them (Goodley, 2014). This concept can be extended to AI models: individuals who are on the margins of a training dataset are more likely to experience model harms, as the model has not been specifically trained to anticipate their needs.

While researchers might not always have access to the training data, it is possible to conduct analyses on other datasets representative of the original data to identify marginalized individuals. Our approach uses outlier detection techniques to select data points on the fringes of the model's exposure, thereby quantitatively defining the "other." By examining the impact of model predictions on these data points, researchers can assess how the model affects this marginalized group. Further, inspecting the demographic makeup of these points can reveal which specific populations are adversely affected by the AI system.

This methodology not only helps to uncover new subgroups of people who may experience harm from AI, but it also addresses the potential invisibility of these groups within the dataset. Insights from this analysis can inform future measurement and mitigation methods, ultimately contributing to developing more equitable AI systems sensitive to diverse needs (Section 5).

3.2 Model and Data

In this study, we examined the impact of AI-driven content moderation on various demographic groups by conducting our analyses on the Perspective API, a toxicity detection tool developed by Jigsaw (Zhang et al., 2018). The Perspective API is used by multiple publishers and platforms, including *The New York Times* and OpenWeb, to assess user-generated content for toxicity. The API predicts several toxicity-related attributes, including toxicity, severe toxicity, identity attack, insult, obscenity, and threat, each associated with specific definitions provided by Jigsaw (Appendix A).

We selected the Jigsaw Unintended Bias dataset due to its detailed demographic annotations and the alignment of its toxicity annotations with our model of interest, the Perspective API. The dataset contains comment text with labels for 24 demographic subgroups across five categories and six types of toxicity. Demographics and toxicity labels are obtained by averaging the opinions of 4-10 annotators, resulting in labels ranging from 0 to 1. We employed the Perspective API to obtain model predictions for each toxicity category and used a 50% threshold to create binary labels for identities, toxicities, disagreements, and model decisions. To manage computational resource constraints, we applied stratified sampling to select a subset of the data, ensuring a proportional representation of all demographic subgroups. The final dataset comprised 20,589 rows and 131 columns.

3.3 Outlier Detection and Analysis

We used the Local Outlier Factor (LOF) method for outlier detection (Breunig et al., 2000).[1] LOF's emphasis on data point density serves as a quantitative representation of othering and belonging. We implemented LOF on multiple vector sets to

[1] Unless otherwise stated, all data manipulation and analysis were performed using the Pandas, Scikit-Learn, and Gensim libraries (McKinney, 2010; Pedregosa et al., 2011; Rehurek and Sojka, 2011).

examine different types of outliers: text-based, disagreement-based, and demographic-based. Text-based outliers are determined using Doc2Vec embeddings of the text of each comment. In this context, outliers are comments with unusual words, phrases, topics, syntax, or other textual patterns. Disagreement-based outliers are determined from binary vectors indicating annotator disagreement on demographic and toxicity scores. These indicate comments for which annotators clashed in unusual ways. Demographic-based outliers are computed using vectors of demographic labels annotated on the dataset, and they represent comments that discuss an unusual demographic group or set of groups.

We expected model performance for all these types of outliers to be poorer than for non-outliers: text-based outliers may contain unusual linguistic harm patterns that are not recognized by the model or reclaim offensive language in positive ways; disagreement-based outliers may simply be more ambiguous in their toxicity than non-outliers; and demographics-based outlier comments may include mentions of groups or include demeaning content that the model has less exposure to.

For each outlier type, we examined significant differences in average toxicity between outliers and non-outliers to identify any disparities in the experiences of these groups. To determine if this pattern is pervasive throughout the dataset, we repeated the analysis on each demographic subgroup. This approach allowed us to gain a deeper understanding of the impact of toxicity both generally and on various demographic groups within the context of outliers and non-outliers.

In Section 4.4, we also analyzed associations between general and intersectional demographic characteristics. We inspected correlation plots to determine relationships between particular demographic characteristics and outlier types to identify any connections between specific demographic attributes and the classification of data points as outliers or non-outliers. Furthermore, we used these plots to examine relationships between different demographic characteristics within a particular outlier type. Finally, we inspected the proportion of members of demographic subgroups considered outliers and non-outliers to determine the distribution of outliers within each demographic subgroup. These analyses helped us to determine individual identities and intersections of demographic characteristics and their associations with outlier classification, which can help to further assess the potential impact of AI models on these populations.

3.4 Model Performance Disparity Scoring

Comparing model performance between outliers and non-outliers allowed us to assess the impact of the model on various groups, considering both overestimation and underestimation of toxicity. We stratified the results by demographic subgroups to determine the pervasiveness of different issues.

When determining what harms particular subgroups face, a critical step is breaking the dataset down into subgroups to identify the ones facing increased harms. To experimentally test our hypothesis that outlier-focused analysis spotlights the subpopulation of marginalized individuals facing greater model harms, we introduce the Group-Based Performance Disparity Index (GPDI), a metric for determining whether a subdivision of a dataset into subgroups identifies the subgroups facing increased harms.

For a demographic attribute i and a full set of data points D, the set g_i is defined as the members of D where the i^{th} attribute is true:

$$g_i = \{x \in D : x_i == True\}$$

$size(g_i)$ represents the cardinality of this set. When splitting the dataset by group membership, we define the GPDI of a single group g_i to be the average percent difference in the model's mean squared error between g_i and its complement:

$$\text{GPDI}_{g_i} = \sum_{t \in T} \frac{\max(\text{MSE}(g_i, t) - \text{MSE}(\neg g_i, t), 0)}{\text{MSE}(\neg g_i, t)}$$

We are interested in areas where the model performs worse for the in-group than the out-group, so we assign the alternative to a score of 0.

The GPDI score for a single group g_i ranges from 0 to 1 and can be interpreted as follows: A score close to 1 signifies a high error rate for members of g_i, indicating significantly more harm caused by the model due to worse predictions. For instance, considering 'Female' as a specific demographic attribute i, the $GPDI_{female}$ score elucidates the model's error difference between data points for women in the dataset (in-group) and the rest of the dataset (out-group). A $GPDI_{g_i}$ score of zero, however, indicates that the model's performance is equal or higher for the specified group compared to the rest of the dataset.

Outlier Type	# Neighbors	Silhouette Score
Text	1,000	0.725
Demographics	4,500	0.266
Disagreement	4,500	0.265

Table 1: Outlier detection hyperparameter and silhouette score for text, demographics, and disagreement outliers. Silhouette score is a cluster goodness metric ranging from -1 to 1; a high score indicates that objects are more similar to their own clusters and less similar to other clusters.

This metric is combined across the set of groups G to compute the GPDI across the set:

$$\text{GPDI}_G = \frac{\sum_{g_i \in G} \text{GPDI}_{g_i} \cdot size(g_i)}{\sum_{g_i \in G} size(g_i)}$$

GPDI is a metric for identifying whether splitting a dataset into a set of groups G helps to identify the groups with the worst model performance disparities between members and non-members. We are interested in sets of groups with the highest GPDI since this implies that the split G identifies the groups for which the model performs much worse on average for in-group members. Due to the weighted averaging technique, GPDI is best suited for sets of groups with low levels of overlap between groups. We also note that GPDI can be arbitrarily maximized by applying it to groups that have already been determined to be harmed. However, its utility lies in identifying groups whose vulnerabilities have not already been revealed. Thus, GPDI helps to select a split into subgroups that identifies the subgroups facing the greatest disparities.

By comparing the GPDIs of various group breakdowns, we can assess the value of considering outlier status in the analysis of model performance.

4 Results

We used outlier detection to identify outliers on the basis of text, demographics, and annotator disagreement. Approximately 1,000 samples were labeled as outliers in each method since we set the contamination parameter in the Local Outlier Factor (LOF) algorithm to 0.05. We varied the n_neighbors parameter by outlier type. Table 1 lists the LOF hyperparameter and silhouette score for each outlier type. Although text-based outliers demonstrated the best clustering, all the scores indicated at least moderately good clustering quality.

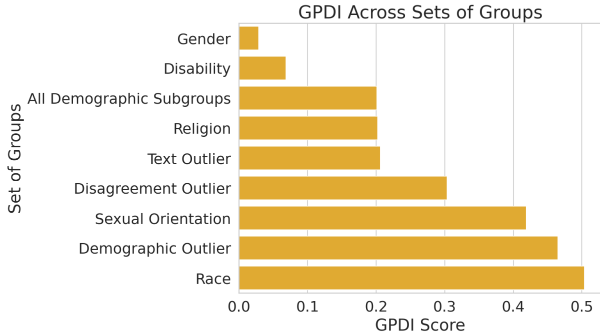

Figure 2: A comparative analysis reveals that race and demographic outliers have the highest GPDI of nine different group breakdowns, suggesting that these analyses are most effective at uncovering which groups face the greatest disparities.

We compared outliers and non-outliers by employing statistical testing to determine the significance of metric differences. We used the non-parametric Mann–Whitney U test for comparing group proportions of binary decisions and the Chi-Square Test of Homogeneity for comparing group average scores, both complemented with a Bonferroni correction.

4.1 Group-Based Performance Disparity Index

Before analyzing outlier-based model performance, we sought to understand how much disparity in model performance is revealed by different types of group breakdowns of the dataset. To do this, we considered the Group-Based Performance Disparity Index (GPDI) scores, defined in Section 3.4, across different types of disaggregation.

We calculated the cumulative GPDI across all 24 demographic groups to be 0.204. Figure 2 illustrates the GPDIs computed by breaking the dataset down by different demographic categories and outlier types. Notably, four out of the eight chosen splits (disagreement outlier, sexual orientation, demographic outlier, and race) offer a higher informational yield than when all demographic categories are considered, illustrating the importance of strategic data breakdowns in uncovering potential model harms. We found that disaggregating by race and demographic outlier status led to the highest GPDI scores of 0.504 and 0.465, respectively, thereby reinforcing the sustained relevance of demographic outliers. Although the GPDI score for race is slightly higher, focusing on demographic outliers helps to identify nuances in harm patterns

Toxicity Type	Demographic Outliers	Text Outliers	Agreement Outliers
Identity Attack	7	6	8
Toxicity	6	5	10
Insult	6	3	11
Severe Toxicity	4	3	1
Obscenity	4	2	0
Threat	1	0	0

Table 2: Number of identity categories with a statistically significant difference in scores for a particular toxicity type for each definition of outliers. Identity attack, general toxicity, and insult have the most pervasive significant differences between outliers and their complements.

Toxicity Type	Overall MSE	Outlier MSE	Non-Outlier MSE	MSE % Decrease for Outliers
Toxicity	0.032	0.041	0.032	30.2%
Severe Toxicity	0.002	0.003	0.002	67.6%
Obscenity	0.009	0.012	0.008	36.9%
Identity Attack	0.030	0.049	0.029	70.8%
Insult	0.022	0.028	0.022	28.1%
Threat	0.006	0.008	0.005	45.5%

Table 3: Model performance overall and divided by outlier status across all types of toxicity. Identity attack and severe toxicity have the greatest significant percent differences between outliers and their complements.

across demographics rather than race alone.

4.2 Toxicity Analysis

Our dataset reveals a pervasive trend where general toxicity (12.2% frequency), identity attack (4.89%), and insult (5.62%) emerge as the most common forms of toxic speech. This is consistent across demographic subgroups.

We uncovered noticeable differences in how outliers and non-outliers experience various forms of toxicity. Figure 3 illustrates how identity attack (86.5%, 56.4%), severe toxicity (68.7% and 41%), and general toxicity (41.1% and 23.3%) are significantly more severe for demographic and text outliers.

In contrast, we observed that toxicity, identity attack, and insult disparities are negative between disagreement outliers and non-outliers, indicating less toxicity in identity outliers. In examining potential reasons for this trend, we found that disagreement outliers have higher agreement than non-outliers (28-45% more) for these toxicity types compared to other types (8-18% more). This suggests that disagreement outliers contain comments that are widely viewed as harmless.

Moreover, when verifying our results, we found that this pattern repeats at the granularity of demographic subgroups. For each demographic subgroup and each toxicity type, we computed whether the difference in toxicity scores between outliers and non-outliers is statistically significant (Table 2). Consistent with the trend across demographics, the most subgroups experience significant disparities between outliers and non-outliers for identity attack, general toxicity, severe toxicity, and insulting harms. Identity attacks are experienced most intensely (86.5%) and pervasively (significant differences in 29% of subgroups) by demographic outliers as well as by text outliers. This suggests that demographic outliers may be more insightful for understanding the harms toward different subgroups of people in the dataset, supporting our conclusions from the GPDI analysis.

4.3 Model Performance Analysis

We focus on demographic outliers for the remaining results because their consistently high GPDI scores and their comparatively intense experience of identity attacks make them particularly potent for understanding harm from toxicity detection models. The GPDI score of 0.465 for demographic outliers underscores the high degree to which the identification of demographic outliers can expose model harm. Additionally, their heightened degree of identity attacks makes them a particularly valuable group to study with respect to their unusual identity characteristics.

In our analysis, we examined the model performance across different toxicity types (Table 3). The MSE for different toxicity detection scores ranges from 0.002 to 0.032. The steep error for multiple toxicity types shown here is concerning because, depending on the thresholds used for content filtration, it could lead to either the preservation of toxic content or the erasure of benign discourse. With respect to outlier status, we examined the percent difference in MSE between outliers and non-outliers. This ranges from 28% to 71%, with statistically significant differences in MSE across all types of toxicity. As such, the differences in MSE between outliers and non-outliers for all types of toxicity were positive and significant.

Since the percent differences in MSE are positive, the model, on average, has more error in the outlier group than in the non-outlier group. This

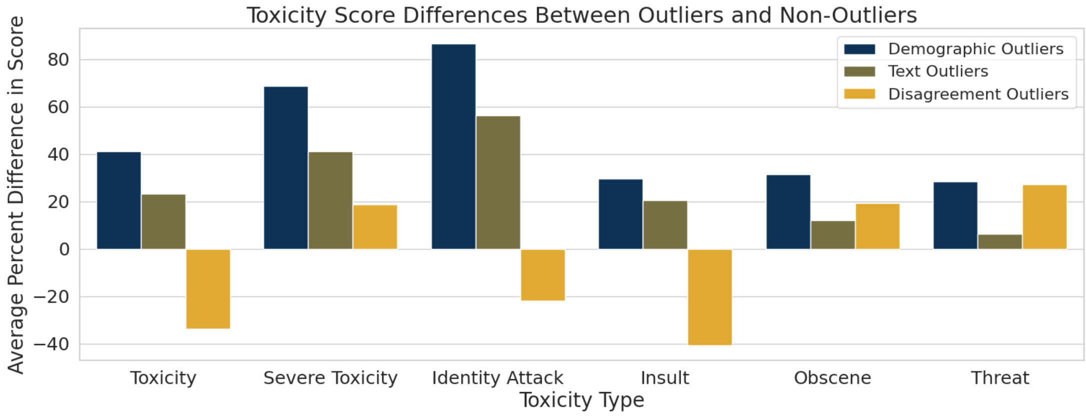

Figure 3: Average differences in toxicity score between three different types of outliers and their complements. Identity attack and severe toxicity have the greatest significant differences.

suggests that the model is not performing as well for these subgroups of individuals and implicates it in harming outlier demographic subgroups. In examining potential sources of this error, we found that 84.8% of all model predictions under-predicted the actual toxicity score of the text. This proportion ranged from 70% to 95.1% across all toxicity types, with 81.1% for identity attack and 90.1% for severe toxicity. The performance discrepancies in the outlier analysis thus suggest that the model does not adequately protect members of less common demographic subgroups from attacking and severely toxic language.

4.4 Demographic Analysis

Our analysis found that demographic outliers exhibited significant differences in the proportion of each demographic subgroup that is classified as an outlier (Figure 4). Curiously, none of the following groups had any outliers: miscellaneous genders, miscellaneous sexual orientations, intellectual disabilities, and miscellaneous disabilities. In contrast, more than half of the points with the following demographic attributes were outliers: Buddhist, Hindu, bisexual, and physical disability. Our methodology thus identifies these groups as potential candidates for protected groups with respect to which to monitor model performance.

We also found a significantly higher average number of identities mentioned for outliers compared to non-outliers: 3.71 versus 1.53. This indicates that outliers are significantly more intersectional than non-outliers, which suggests that individuals belonging to multiple marginalized groups may experience compounded harm from the model. Our analysis also revealed that demographic outliers have a clustering of demographic binaries within a particular axis, specifically within race and religion. This suggests that demographic outlier analysis captures harms toward intersectional groups. It also allows us to identify these groups as potential candidates for intersectional disaggregated analysis to better understand the experiences of individuals at the intersections of multiple marginalized identities. By focusing on these intersections in the error measurement and mitigation process, we can better address the model harms faced by these individuals.

5 Conclusion

Our research leverages disability studies's insights to illuminate unintended harms inflicted by AI systems, particularly those operating in toxicity detection. We have shown how societal constructions of "normalcy"—often shaped and reinforced by statistical methodologies—can lead to biases and exclusions in AI models, which in turn amplify societal barriers and inequities. In direct response to this challenge, we have contributed in three significant ways: First, we have proposed and implemented a method for identifying marginalized groups at risk of AI harm by using outlier detection techniques to create and examine three different types of outliers. Second, we introduced the Group-Based Performance Disparity Index (GPDI), a measure that effectively quantifies disparities in AI model performance, offering a practical way to compare sets of groups experiencing model harms. We found that breaking the dataset down by race or by demographic outlier status gave the highest GDPI, indicating that these breakdowns expose the greatest disparities in model error. Finally, we

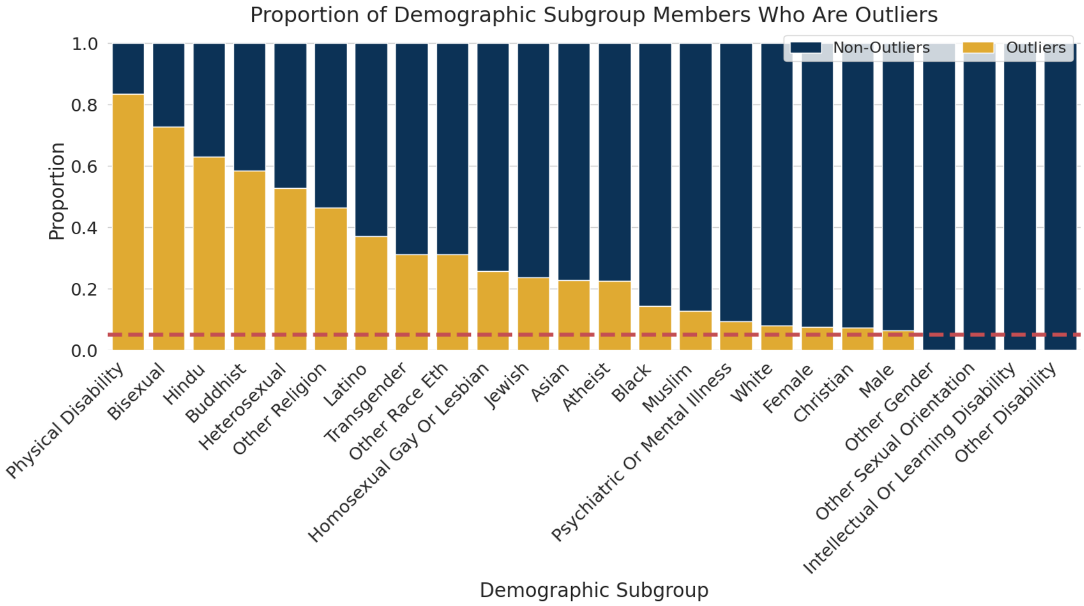

Figure 4: Proportions of each demographic subgroup that are considered outliers or non-outliers. The red line indicates the overall proportion of outliers. Four subgroups are >50% outliers, and four have no outliers.

critically examined model performance disparities across six types of toxicity, finding identity attacks to be particularly acute and pervasive for demographic outliers, alongside general toxicity, severe toxicity, and insults to a lesser degree. We also identified racial and religious groups as members of demographic outliers, as well as people with physical disabilities and bisexual people. These demographic subgroups are thus particularly vulnerable to harm from such disparities.

Our research findings and methodologies pave the way for immediate application and future exploration in AI auditing and fairness mitigation. Operationalizing GPDI in AI auditing practices can aid in identifying, measuring, and eventually mitigating model harms across different groups. A promising direction for future research lies in assessing the use of outlier detection to identify algorithmically harmed groups in scenarios lacking explicit demographic data. This approach could broaden fairness auditing's scope and deepen our understanding of statistical normalcy, social marginalization, and their manifestations in AI systems.

Ethics Statement

The research presented in this paper was conducted using the Jigsaw Unintended Bias in Toxicity Classification dataset, which covers English data only. We selected this dataset for its granular demographic labeling and connection to the Perspective API, which was the focus of this research. However, we note that information on the geographic, linguistic, and demographic background of the commenters and annotators has not been provided, so it is unclear what worldviews the dataset reflects regarding what constitutes harmful content and what types of demographic groups are familiar to the annotators as potential targets of harmful speech. Furthermore, since this study only analyzed the Perspective API, further research is needed to extend our findings beyond this context, such as on other models, tasks, or datasets.

Our introduction of the Group-Based Performance Disparity Index (GPDI) serves as a discovery tool for evaluating and mitigating algorithmic harm within a specific context. Despite its beneficial applications, we recognize the potential misuse of GPDI. Specifically, there is a risk that it could be exploited to cherry-pick groups within a set to justify particular courses of action or to arbitrarily maximize its value (e.g., by duplicating groups to upweight their importance).

Our work proposes using outlier detection as a means of identifying groups potentially harmed by algorithmic bias. This approach has the potential to minimize the need for demographic label inference in future AI fairness evaluations, thus avoiding potential pitfalls and biases associated with such inferences. However, this method needs further exploration and rigorous testing to confirm its efficacy and examine tradeoffs before being used in high-stakes domains.

References

Solon Barocas, Anhong Guo, Ece Kamar, Jacquelyn Krones, Meredith Ringel Morris, Jennifer Wortman Vaughan, W. Duncan Wadsworth, and Hanna Wallach. 2021. Designing disaggregated evaluations of AI systems: Choices, considerations, and tradeoffs. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM.

Solon Barocas and Andrew D. Selbst. 2016. Big data's disparate impact. *SSRN Electronic Journal*.

Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like trainer, like bot? inheritance of bias in algorithmic content moderation.

Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. 2000. LOF. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*. ACM.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR.

Judith Butler. 1990. *Gender Trouble*. Routledge.

Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning.

Kate Crawford. 2017. The trouble with bias. keynote talk at Neural Information Processing Systems (NIPS '17).

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

L. J. Davis. 1995. *Enforcing normalcy: Disability, deafness, and the body*. Verso Books.

L. J. Davis. 2014. *The end of normalcy: Identity in a biocultural era*. The University of Michigan Press.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ACM.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

P. S. Fass. 1980. The iq: A cultural and historical framework. *American Journal of Education*, 88.

T. Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

E. Goffman. 1963. *Stigma: Notes on the management of spoiled identity*. Touchstone.

D. Goodley. 2014. *Dis/ability Studies: Theorising disablism and ableism*. Routledge.

Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data &amp Society*, 7(1):205395171989794.

Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

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

Nathan Kallus, Xiaojie Mao, and Angela Zhou. 2022. Assessing algorithmic fairness with unobserved protected class using data combination. *Management Science*, 68(3):1959–1981.

Youjin Kong. 2022. Are "intersectionally fair" AI algorithms really fair to women of color? a philosophical analysis. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.

Wes McKinney. 2010. Data structures for statistical computing in python. In *Proceedings of the Python in Science Conference*. SciPy.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35.

Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2):205395171667967.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

Edward W Said. 1988. *Orientalism*. Vintage Books, New York, NY.

Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cédric Langbort. 2014. Auditing algorithms : Research methods for detecting discrimination on internet platforms.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Tom Shakespeare. 2006. *Disability rights and wrongs*. Routledge, London, England.

Richard Twine. 2002. Physiognomy, phrenology and the temporality of the body. *Body & Society*, 8(1):67–88.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

A Jigsaw Toxicity Types

Toxicity Type	**Definition**
Toxicity	A rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion.
Severe Toxicity	A very hateful, aggressive, disrespectful comment or otherwise very likely to make a user leave a discussion or give up on sharing their perspective. This attribute is much less sensitive to more mild forms of toxicity, such as comments that include positive uses of curse words.
Identity Attack	Negative or hateful comments targeting someone because of their identity.
Insult	Insulting, inflammatory, or negative comment towards a person or a group of people.
Obscenity	Swear words, curse words, or other obscene or profane language.
Threat	Describes an intention to inflict pain, injury, or violence against an individual or group.