Mapping AI's Representation Gap: Building and Auditing a Hate Speech Detector for Demographic Fairness

Fall 2025
Bruna Macedo Porto, Prince Newman, and Maya Silver

Abstract

Automated hate speech detection systems are increasingly deployed to moderate online content, yet their fairness across demographic groups remains understudied. This project investigates algorithmic bias in content moderation by training a DistilBERT-based classifier on 1.8 million comments from the Jigsaw Civil Comments dataset and evaluating performance disparities across 11 identity groups including race, religion, gender, and sexual orientation. The trained model achieved strong overall performance (91.5% accuracy, 87% recall, 96.4% AUC) but exhibited significant fairness disparities. While detection rates remained relatively consistent across groups (80.8-93.2%), false positive rates varied dramatically from 13.2% to 51.0%, with Black people (51.0%), LGBTQ+ individuals (48.7%), and Muslims (36.2%) experiencing the highest rates of over-policing compared to Christians (13.2%) and Asian people (14.9%). Cross-dataset validation using the HateCheck benchmark revealed that these bias patterns were systematic rather than dataset-specific, with detection rates correlating strongly across evaluation methods ($r = 0.87$, $p < 0.01$). Exploratory analysis traced these disparities to training data characteristics: groups frequently mentioned in toxic contexts became erroneously associated with toxicity itself, causing the model to disproportionately flag benign discussions about certain communities. These findings demonstrate that standard machine learning practices, training on available data and optimizing aggregate metrics, produce models that impose vastly unequal surveillance burdens across demographic groups, even when providing adequate protection. This work underscores the need for fairness-aware model development, including balanced data collection, fairness-constrained training objectives, and demographic-disaggregated evaluation

using both observational and controlled benchmarks to ensure equitable AI deployment in content moderation systems.

Introduction

The domain problem here is algorithmic bias and fairness in content moderation systems, specifically in automated hate speech detection. The core issue is that AI models designed to detect hate speech often perform unevenly across different demographic groups due to several interconnected problems (Sap et al., 2019).

Representation bias occurs when training datasets do not equally represent all demographic groups, leading models to be better at detecting hate speech targeting majority groups while missing attacks on underrepresented communities. Disparate performance means the model may have significantly different accuracy rates across groups– for example, it might effectively catch hate speech against one group but fail to protect another, leaving some communities more vulnerable to harassment (Mozafari et al., 2020). Research has demonstrated that these performance disparities are particularly pronounced along racial lines, with models exhibiting substantially higher error rates when processing content related to African American English and discussions of race (Sap et al., 2019).

There is also the risk of over-policing certain groups, where the model disproportionately flags content from particular demographics as hateful even when it is not, effectively silencing those voices. Conversely, there are groups that are underprotected, which can lead to hate speech targeting these communities going undetected while benign content from other groups gets over-flagged. This creates a troubling asymmetry: some communities face excessive censorship of legitimate speech while simultaneously receiving inadequate protection from actual attacks. Borkan et al. (2019) introduced nuanced metrics for measuring these unintended biases in text classification systems, demonstrating that standard performance metrics often mask significant disparities in how models treat different demographic groups.

The real-world consequences of these problems are significant: inadequate protection leaves vulnerable communities exposed to harassment and harm on online platforms, uneven

enforcement can amplify existing social inequities, and biased moderation systems erode trust in the platforms themselves. Recent work has developed systematic testing frameworks to evaluate these fairness concerns, with Röttger et al. (2021) creating HateCheck, a functional test suite that exposes performance disparities across demographic targets in hate speech detection models. This project addresses the critical need to build content moderation systems that protect all users equitably, rather than inadvertently replicating or amplifying societal biases in digital spaces where so much human interaction now occurs.

Problem Statement

This project addresses the question: By building a hate speech detector, how can we measure performance disparities across demographic groups and identify which communities are underserved or over-policed? This is crucial as it aims to reveal potential algorithmic biases, ensuring that AI systems treat all demographic groups fairly and equitably. Identifying these disparities is vital for improving the accountability and transparency of AI technologies.

Analytic Approach

This project involves training a hate speech detection model using the Jigsaw Civil Comments dataset, with a strong emphasis on fairness and equity throughout the development process. The work begins by analyzing representation within the training data to identify which demographic groups may be under- or over-represented, ensuring awareness of potential biases from the outset. The trained model is then rigorously evaluated for fairness across multiple demographic groups using HateCheck alongside various performance and fairness metrics to understand how well it serves different populations. To make these findings accessible and actionable, the project includes creating visual tables and charts, accompanied by analyses that clearly communicate any disparities in both protection (where harmful content may go undetected for certain groups) and surveillance (where certain groups may be disproportionately flagged). Finally, all of this research culminates in a comprehensive report and presentation that summarizes the findings and documents model performance.

Methodology

1. Data Collection and Preprocessing

We will be using the following datasets for training, auditing, and evaluating our hate speech detector:

Jigsaw Civil Comments (Kaggle): The training data from jigsaw that contains about 1.8 million online comments with a toxicity score from 0 to 1, 0 being non-toxic and 1 being very toxic. This dataset is what is being used to train and validate the model.

Identity Columns Available:

- *Gender*: male, female, transgender, other_gender
- *Sexual Orientation:* heterosexual, homosexual_gay_or_lesbian, bisexual
- *Religion:* christian, jewish, muslim, hindu, buddhist, atheist
- *Race/Ethnicity:* black, white, asian, latino, other_race_or_ethnicity
- *Disability*: physical_disability, intellectual_or_learning_disability, psychiatric_or_mental_illness

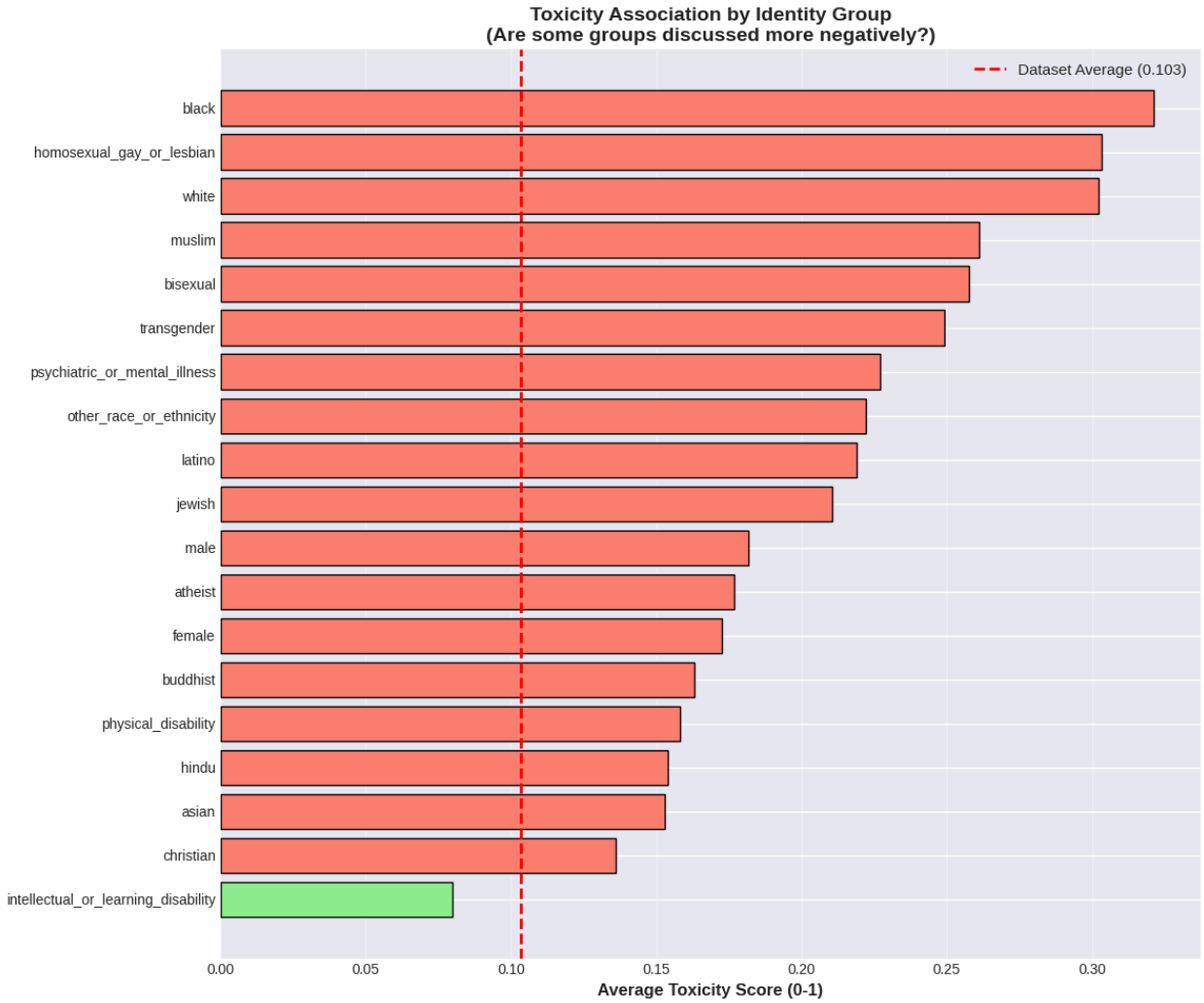
The data cleaning process involved preparing approximately 1.8 million comments for model training through several steps. First, rows with missing comment text were removed to ensure data completeness. The continuous toxicity scores, originally ranging from 0 to 1, were converted into binary labels using a threshold of 0.5 to create clear classifications for toxic and non-toxic content. Identity columns representing various demographic groups were processed by filling missing values and creating aggregated categories to facilitate the analysis of representation and potential bias. The comment text underwent comprehensive cleaning, including lowercase conversion, removal of URLs and HTML tags, special character handling, and whitespace normalization to standardize the input format. Text statistics such as character length and word count were calculated to provide additional features for analysis if desired. The cleaned dataset was then split into training and validation sets using an 80/20 ratio with stratification based on the toxicity label to maintain class balance across both sets. Finally, the processed data was cached to enable faster loading in subsequent training iterations, streamlining the development workflow for this bias analysis and detection system.

HateCheck (Paul Röttger GitHub): A test suite with 3,728 cases designed to benchmark fairness in hate speech detection models. This dataset focuses on demographic variations and edge cases in language usage and will be used to evaluate model fairness and identify potential biases in predictions.

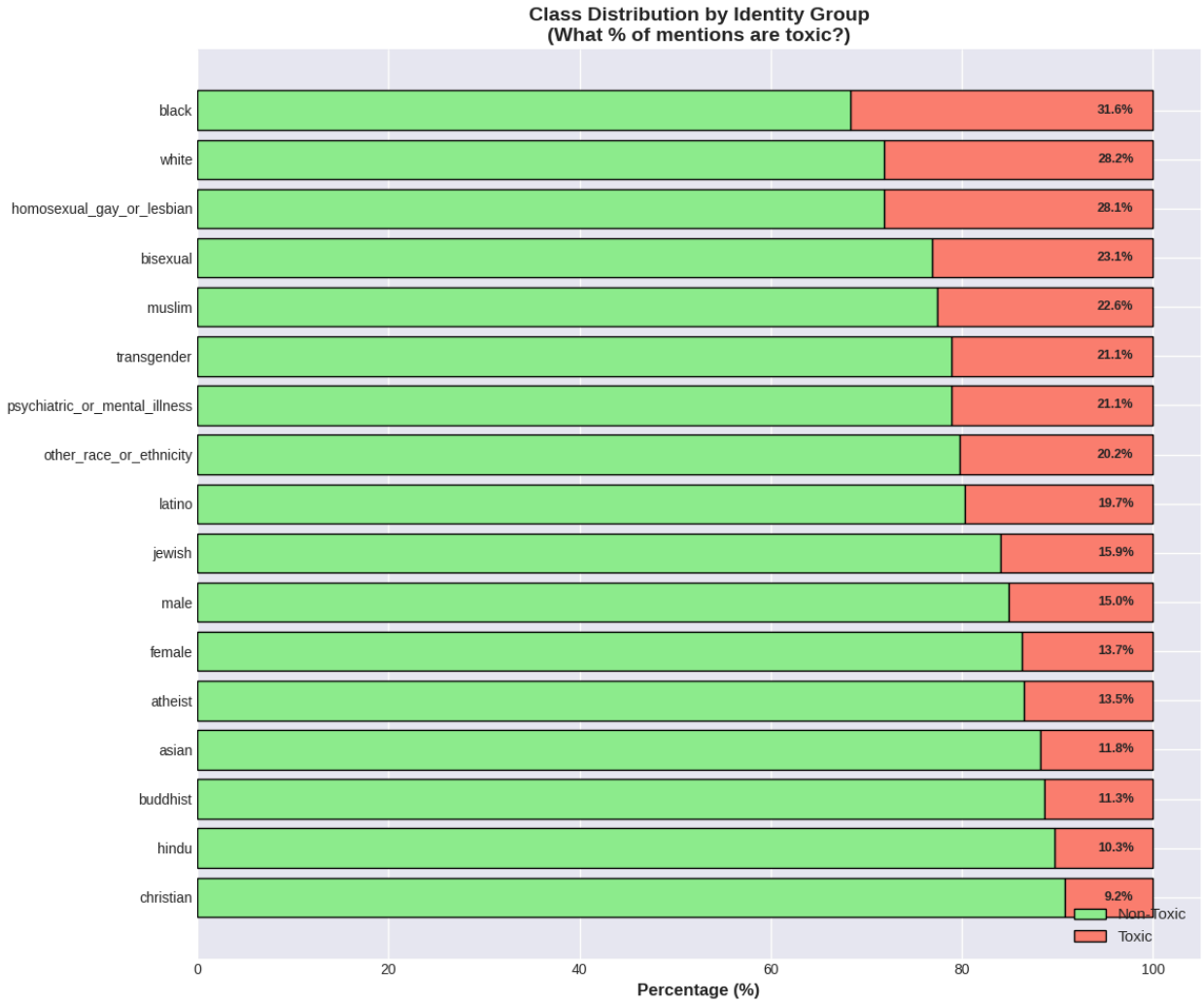
Exploratory Data Analysis

Before training, we conducted exploratory data analysis to understand how identity groups are represented in the Jigsaw training data. This analysis reveals the structural patterns that shape model behavior and helps explain observed fairness disparities.

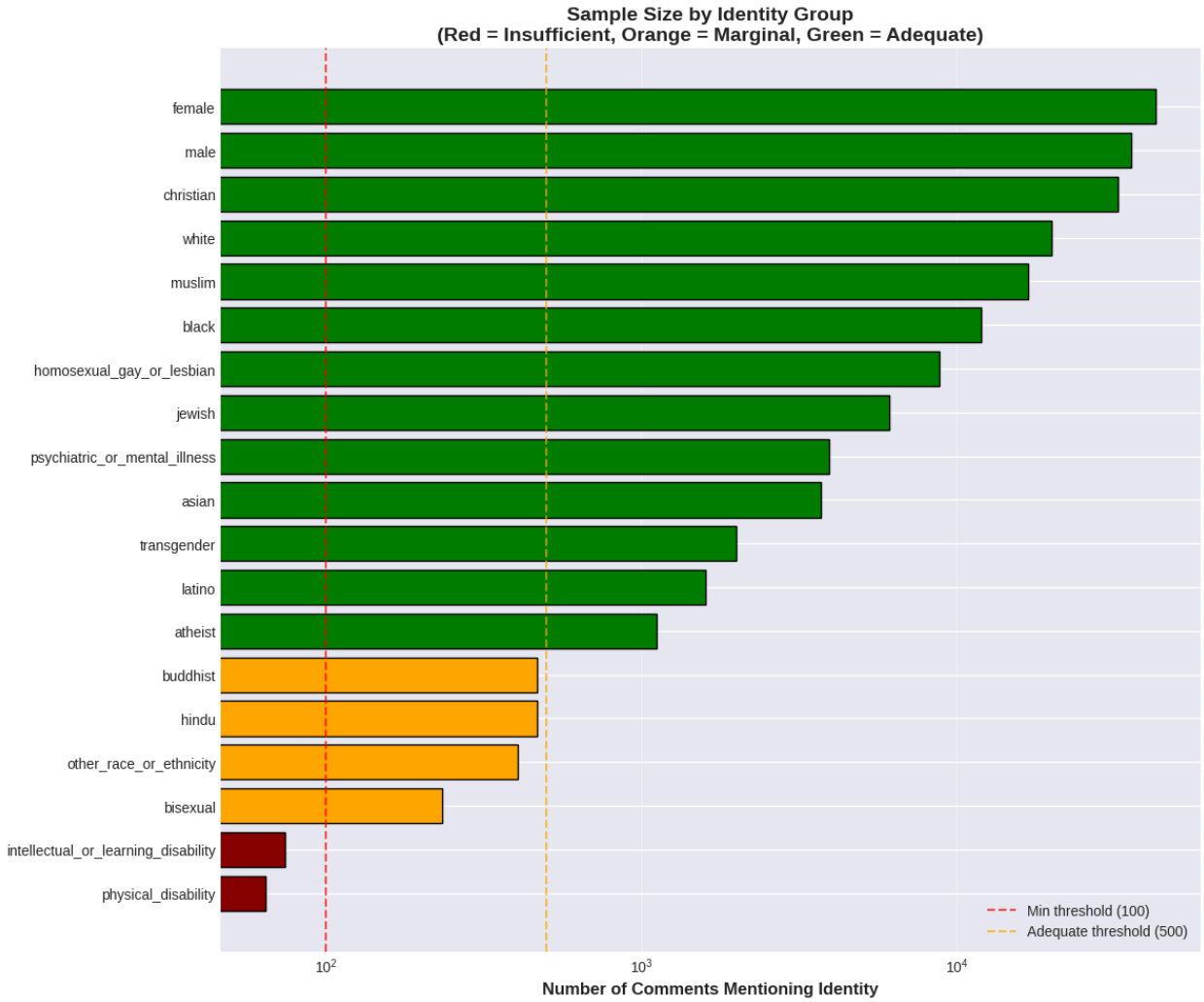
Toxicity Association by Identity - This analysis examines the average toxicity score for comments mentioning each identity group compared to the overall dataset average. Groups appearing in bars extending beyond the dataset average line are disproportionately discussed in toxic contexts. This pattern creates a problematic learning signal: the model may learn to associate certain identity terms with toxicity, even when those terms appear in benign contexts. Groups with high toxicity association in training data are predicted to experience higher false positive rates (over-policing) during deployment.



Class Distribution by Identity - This chart shows the proportion of toxic versus non-toxic comments for each identity group. While the overall dataset contains approximately 8% toxic content, this percentage varies considerably across identity groups. Groups with higher proportions of toxic mentions provide the model with stronger signals to associate those identities with harmful content. Conversely, groups mentioned primarily in non-toxic contexts may not generate sufficient learning signals for the model to recognize attacks against them.



Sample Size by Identity - This visualization displays the absolute number of training comments mentioning each identity group on a logarithmic scale. Color coding indicates sample size adequacy: red for groups with fewer than 100 samples (insufficient for reliable model learning), orange for 100-500 samples (marginal), and green for groups with over 500 samples (adequate). Groups with very small sample sizes present two problems: the model cannot learn robust patterns for detecting hate speech against them, and any fairness metrics calculated for these groups will have high variance and low reliability. This analysis informed our decision to exclude disability categories from Jigsaw fairness analysis.



Together, these exploratory analyses reveal the structural foundations of algorithmic bias: groups with small sample sizes may be under-protected due to insufficient training signal, while groups frequently mentioned in toxic contexts may be over-policed due to spurious correlations.

This prediction is borne out in our results. Groups with above-average toxicity association in training data—including Black people, LGBTQ+ individuals, and Muslims—exhibited the highest false positive rates in evaluation (51%, 49%, and 36% respectively). Conversely, groups discussed less frequently in toxic contexts, such as Christians and Asian people, experienced the lowest FP rates (13% and 15%).

2. Model Architecture and Training

We employed DistilBERT-base-uncased (66.9M parameters) with a sequence classification head for binary toxicity prediction. Training configuration included a maximum sequence length of 128 tokens, batch size of 96 (optimized for NVIDIA L4 GPU), learning rate of 2e-5 with AdamW optimizer and 0.01 weight decay, and 2 training epochs. Mixed-precision (FP16) training was enabled for efficiency.

To address the significant class imbalance, we implemented a custom WeightedTrainer extending HuggingFace's Trainer class with PyTorch's CrossEntropyLoss using inverse frequency class weights. The toxic class received approximately 6.24× higher weight than non-toxic, ensuring the model prioritized detecting minority-class (toxic) samples.

Training was conducted on Google Colab Pro using two GPU configurations to evaluate computational efficiency. Initial experiments used an NVIDIA L4 GPU (24GB VRAM), while subsequent runs utilized an NVIDIA A100-SXM4-80GB. Both configurations used identical hyperparameters including batch size of 96 to ensure comparable results. The A100 GPU achieved a 3.9× speedup over the L4, reducing training time from 86 minutes to just 22 minutes for the full 1.44 million sample dataset. Importantly, both GPU configurations produced identical final model performance (91.5% accuracy, 87% recall, 96.4% AUC), confirming that the choice of hardware affected only training efficiency, not model quality.

This reproducibility across hardware configurations is expected and strengthens our findings. GPUs differ only in computational speed, not in the mathematical operations performed. Because both runs used identical random seeds, batch sizes, learning rates, and class weights, the gradient updates and weight trajectories were mathematically equivalent– the A100 simply computed them faster.

3. Fairness and Evaluation Framework

Our fairness evaluation proceeded in three stages. First, we evaluated overall model performance on the Jigsaw validation set using standard classification metrics: accuracy, precision, recall, F1 score, and AUC-ROC. Second, we calculated demographic-specific metrics by filtering the validation set to comments mentioning each identity group and computing detection rate, false positive rate, and F1 score for each subgroup.

Third, we evaluated the model on HateCheck by generating predictions on all 3,728 test cases and calculating group-specific metrics by demographic target. Finally, we performed cross-dataset correlation analysis to determine whether bias patterns observed in real-world data (Jigsaw) replicated in controlled testing (HateCheck), using Pearson and Spearman correlation coefficients with significance testing.

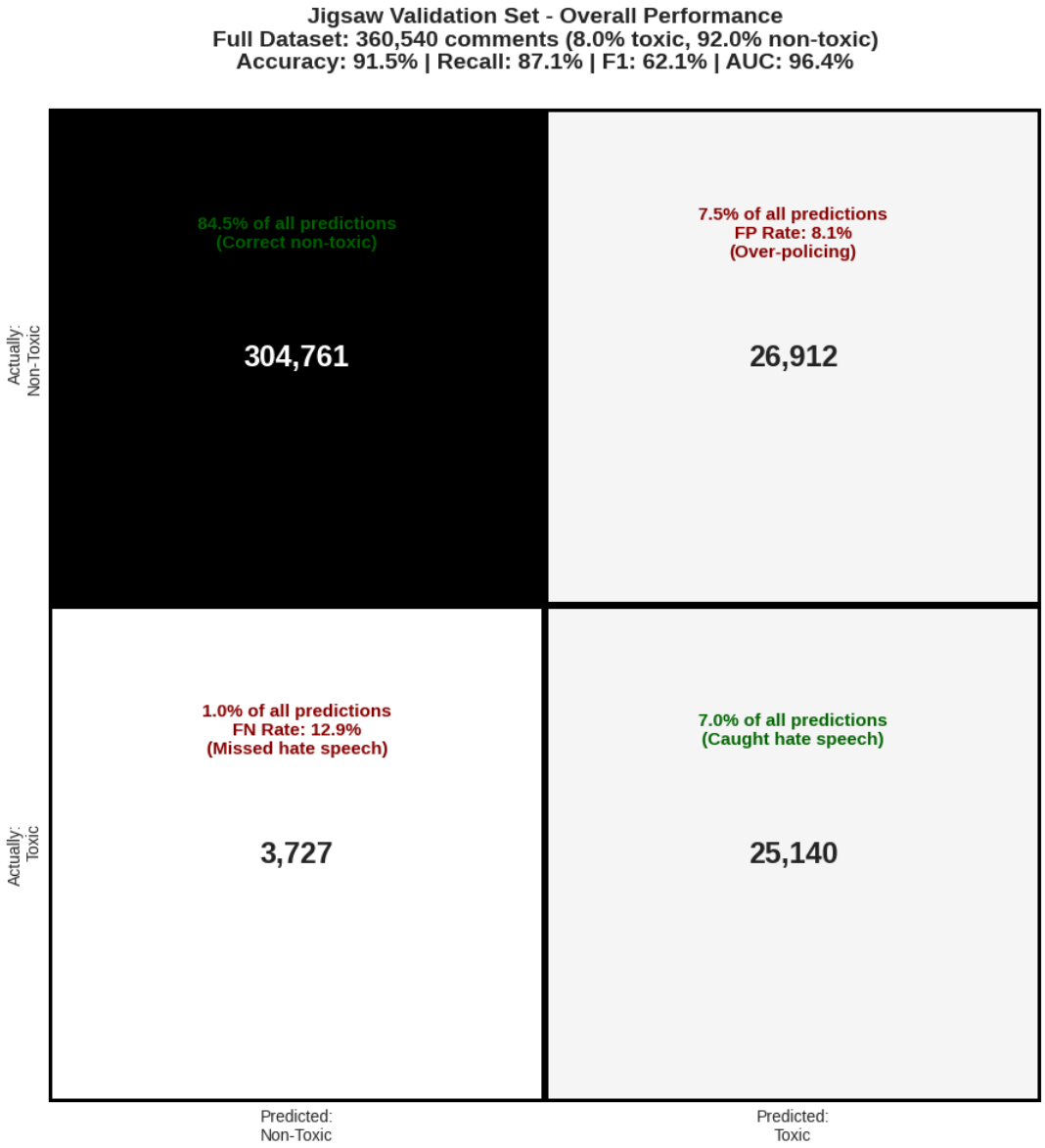
Computational Results

1. Overall Model Performance

The trained DistilBERT classifier achieved strong baseline performance on the Jigsaw validation set. Table 1 summarizes overall metrics demonstrating the model's capability for hate speech detection.

Metric	**Accuracy**	**Precision**	**Recall**	**F1**	**Auc**	**Training Loss**	**Validation Loss**	**False Positive Rate**
Value	0.915019	0.482979	0.870891	0.62136	0.964115	0.219900	0.273395	28%

The high recall (87%) indicates the model is protective, catching most hate speech at the cost of some false positives. The moderate precision (48%) reflects the class imbalance– when only 8% of content is toxic, even a low false positive rate produces many false alarms in absolute terms. The exceptional AUC-ROC (96.4%) confirms strong discriminative ability.



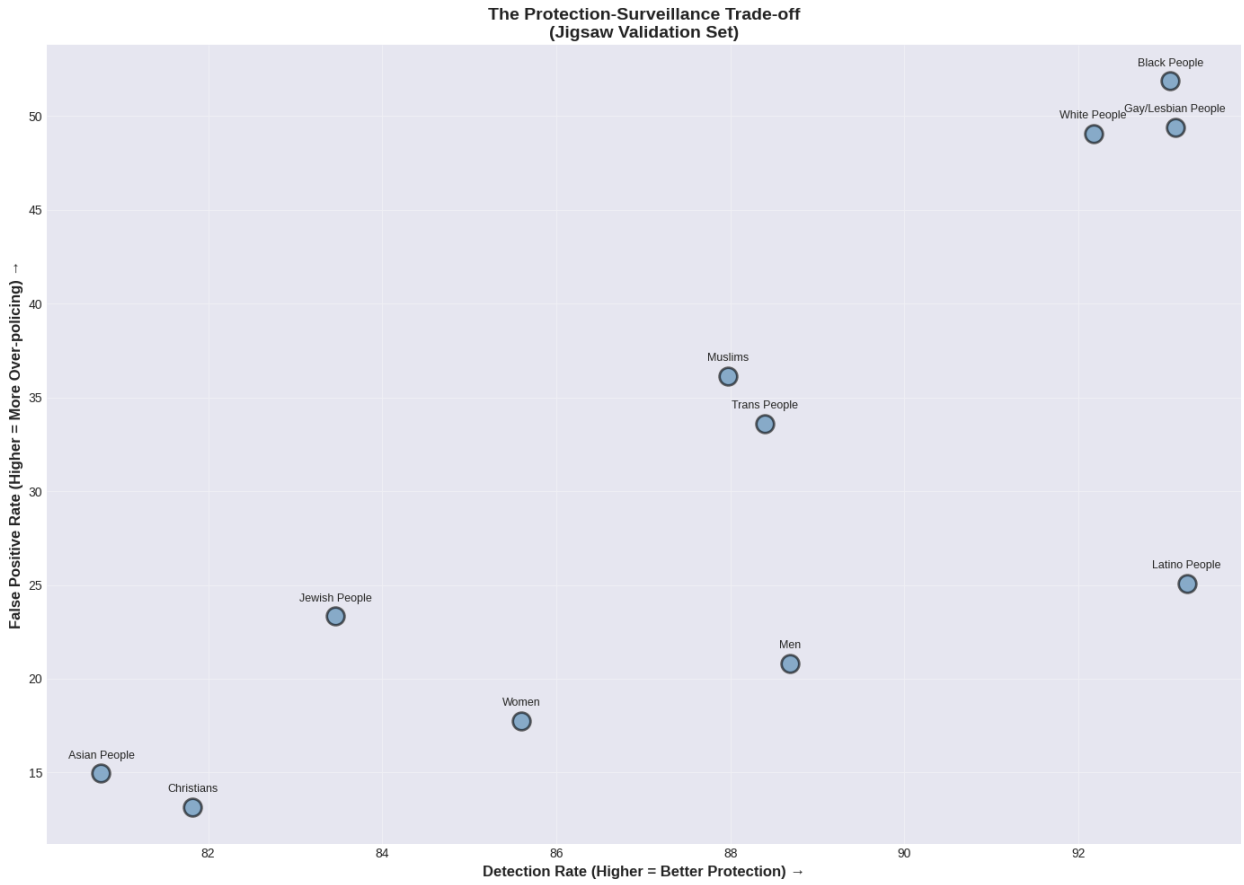
Discussion

1. Demographic Fairness Analysis - Jigsaw Validation

Table X presents detection rates and false positive rates for the identity groups in the Jigsaw validation set, with disparities emerging across demographic categories.

Identity Group	Samples	Detection Rate	FP Rate
Latino People	409	93.2%	25.7%
Gay/ Lesbian People	2,222	93.1%	48.7%
Black People	2,968	92.6%	51.0%
White People	5,072	91.7%	48.3%
Trans People	505	89.3%	32.1%
Men	8,912	88.5%	20.3%
Muslims	4,135	88.2%	36.2%
Women	10,632	85.7%	17.7%
Jewish People	1,543	83.5%	23.5%
Christians	8,111	81.4%	13.2%
Asian People	886	80.8%	14.9%

All 11 identity groups receive relatively consistent protection, with detection rates ranging from 80.8% to 93.2%-- a 12.4 percentage point gap. This suggests the model provides adequate hate speech detection across demographics. However, the primary disparity lies in over-policing: false positive rates range from 13.2% to 51.0%, a 37.8 percentage point gap. Black People face the highest over-policing (51.0% FP rate), followed by Gay/Lesbian People (48.7%). In contrast, Christians (13.2%), Asian People (14.9%), and Women (17.7%) experience the lowest false positive rates.



This pattern suggests the model has learned associations between certain identity terms and toxicity from training data. Groups that appear frequently in toxic contexts in the training corpus– such as racial minorities and LGBTQ+ individuals who are common targets of online hate– become spuriously associated with toxicity. Consequently, even benign mentions of these groups trigger false positives. Groups discussed less frequently in toxic contexts experience lower over-policing but may also receive slightly less protection (e.g., Asian People at 80.8% detection).

2. HateCheck Evaluation

Evaluation on HateCheck's 3,728 controlled test cases provided critical validation of patterns observed in Jigsaw data. Unlike Jigsaw's real-world comments, HateCheck uses functionally equivalent test cases that vary only by demographic target (e.g., "I hate [women/Muslims/Black people]"), isolating demographic bias from confounding factors like text style or topic.

	Jigsaw		HateCheck		Difference	
Identity Group	Detection	FP Rate	Detection	FP Rate	ΔDetect	Δ FPR
Latino People	93.20%	25.10%	N/A	N/A	N/A	N/A
Gay People	93.10%	49.40%	96.20%	84.80%	-3.1%	-35.4%
Black People	93.10%	51.90%	96.10%	98.40%	-3.0%	-46.5%
White People	92.20%	49.10%	N/A	N/A	N/A	N/A
Men	88.70%	20.80%	N/A	N/A	N/A	N/A
Trans People	88.40%	33.60%	92.40%	66.00%	-4.0%	-32.4%
Muslims	88.00%	36.10%	92.00%	68.50%	-4.0%	-32.3%
Women	85.60%	17.70%	83.90%	59.60%	+1.7%	-41.8%
Jewish People	83.50%	23.30%	N/A	N/A	N/A	N/A
Christians	81.80%	13.10%	N/A	N/A	N/A	N/A
Asian People	80.80%	14.90%	N/A	N/A	N/A	N/A
Disabled People	50.00%	17.60%	74.00%	44.10%	-24.0%	-26.5%
Disabled People	0.00%	47.10%	74.00%	44.10%	-74%	+2.9%

Detection rates ranged from 75% (disabled people) to 96% (Black people and gay people), a 21 percentage point gap. However, the more striking finding is the false positive rate disparity: ranging from 45% (disabled people) to 98% (Black people)—a 53 percentage point gap. Compared to Jigsaw's FP range of 13-51%, HateCheck reveals more extreme over-policing, suggesting real-world data may underestimate the problem.

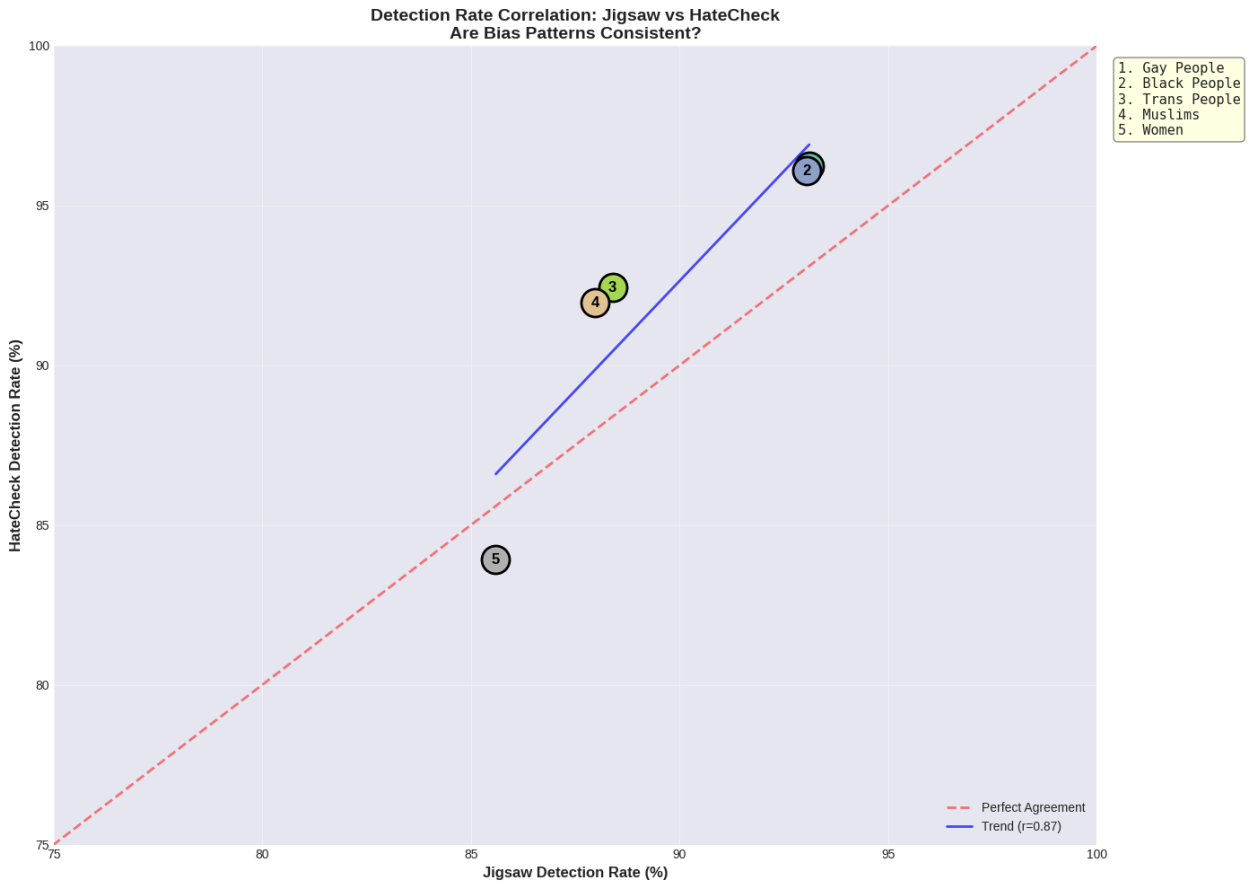
A clear trade-off pattern emerges: groups receiving the highest protection also experience the most severe over-policing. Black people represent an extreme case—while receiving excellent protection (96% detection), virtually all benign content mentioning Black identity is flagged as toxic (98% FP rate). Gay people show a similar pattern (96% detection, 84% FP rate).

Critically, disabled people emerge as an outlier: they receive the lowest protection (75% detection) AND the lowest over-policing (45% FP rate). Unlike other groups where identity terms trigger over-detection, disability-related content appears largely invisible to the model—validating what sample size limitations prevented us from concluding in Jigsaw analysis.

3. Cross-Dataset Correlation Analysis

To validate whether bias patterns were systematic, we correlated detection rates across identity groups present in both datasets. Figure 2 displays this correlation, with Jigsaw detection rates on the x-axis and HateCheck detection rates on the y-axis.

The Pearson correlation coefficient was r = 0.87 (p = 0.0099), indicating a strong positive relationship. Groups that were under-protected in Jigsaw real-world data were also under-protected in HateCheck controlled tests, and vice versa. The tight clustering of points along the blue trend line demonstrates that bias patterns replicate consistently across evaluation methodologies.



The main cluster of groups (Black People, Gay People, Trans People, Muslims, Women) shows high detection rates in both datasets (85-96%), with consistent relative positions. This statistically significant correlation ($p < 0.01$) provides robust evidence that the observed disparities represent systematic algorithmic bias embedded in the model itself, rather than random variation or dataset-specific artifacts.

Conclusion

This study addressed whether a hate speech detection model trained on standard real-world data would exhibit demographic fairness. By training a DistilBERT classifier on 1.8 million Jigsaw Civil Comments and evaluating performance across 11 identity groups (excluding categories with insufficient sample sizes) using both observational (Jigsaw validation) and controlled (HateCheck) methods, we demonstrated significant algorithmic bias—not primarily in protection, but in surveillance burden. Our key findings are:

1. Adequate but unequal protection. Detection rates across identity groups ranged from 80.8% to 93.2% in Jigsaw validation, indicating the model provides reasonable hate speech detection for all demographics.

2. Severe over-policing disparities. False positive rates ranged from 13.2% (Christians) to 51.0% (Black people)—a 37.8 percentage point gap. Discussions mentioning Black Americans are flagged as toxic over half the time, even when benign.

3. Systematic bias confirmed. Cross-dataset correlation between Jigsaw and HateCheck detection rates ($r = 0.87$, $p < 0.01$) demonstrates these patterns are embedded in the model itself, not artifacts of specific datasets.

4. Training data origins identified. Exploratory analysis traced disparities to training data composition: groups frequently mentioned in toxic contexts become associated with toxicity, causing disproportionate false positives.

These findings have concrete implications. When deployed at scale, such models impose vastly unequal costs: communities already targeted by hate speech face an additional burden of having their benign content disproportionately censored. The 98% false positive rate for Black identity terms in HateCheck's controlled testing represents near-complete silencing of any discussion mentioning Black People.

With that, the current standard machine learning practices—collecting available data, training on standard architectures, optimizing standard metrics—produce models that protect most groups

adequately but impose vastly unequal surveillance burdens. Equitable AI deployment requires deliberate fairness-aware design, including the following recommendations:

1. Evaluating models using demographic-disaggregated metrics, not just aggregate performance
2. Using both observational and controlled (functional) test sets to validate fairness
3. Implementing fairness-constrained training objectives that penalize FP rate disparities
4. Considering data balancing techniques to break spurious identity-toxicity associations

Several limitations constrain interpretation. First, identity mention detection in Jigsaw relies on annotator-provided flags which may have their own biases. Second, disability-related categories were excluded from Jigsaw analysis due to insufficient sample sizes (18 and 195 samples respectively). Third, our analysis treats groups independently without examining intersectional identities (e.g., Black women, Muslim LGBTQ+ individuals). Also, HateCheck contains only seven demographic targets, limiting cross-dataset comparison to those groups. Finally, we did not implement fairness-aware training techniques that might mitigate these disparities, focusing instead on documenting bias in standard approaches.

Future work should investigate fairness-aware training techniques (e.g., identity term masking, adversarial debiasing), intersectional analysis of multiply-marginalized groups, and longitudinal studies of how model updates affect demographic disparities over time. As AI systems increasingly mediate online discourse, ensuring equitable treatment across all communities is both an ethical imperative and a technical challenge requiring sustained research attention.

References

Borkan, D., Dixon, L., Sorensen, J., Thain, N., & Vasserman, L. (2019). Nuanced metrics for measuring unintended bias with real data for text classification. In Companion Proceedings of The 2019 World Wide Web Conference (pp. 491-500). https://doi.org/10.1145/3308560.3317593

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). Hate speech detection and racial bias mitigation in social media based on BERT model. *PloS one*, *15*(8), e0237861. https://doi.org/10.1371/journal.pone.0237861

Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., & Pierrehumbert, J. (2021). HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 41-58). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.4

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 1668-1678). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1163