A Comparative Analysis of Pre-Trained NLP Models for Toxicity Detection and Unintended Bias

Sreeniyathi Kasireddy
skasireddy@umass.edu

Mantra Burugu
mburugu@umass.edu

Farhana Rahman
farrahman@umass.edu

Abstract

This project is a comparative analysis of pre-trained Natural Language Processing models by evaluating how accurately they detect toxicity as well as whether they show unintended identity bias. Since some models are trained using social media data, they can learn biased patterns and falsely flag harmless comments that are related to identities like gender, religion, race, or sexual orientation.
Using the 2019 Jigsaw Unintended Bias dataset, we compare three models, Detoxify Original, roberta-hate-speech-dynabench-r4-target, and Twitter-RoBERTA Offensive, against human annotated labels already given with the dataset. We analyze the differences in how each model performs with toxicity detection on unseen text and their bias-handling mechanism. This analysis shows how each model operates, and help in choosing the right tool for fair content moderation and point toward specific areas for future improvement to reduce bias.

1 Introduction

Recent studies have shown that pre-trained Natural Language Processing models perform well on identifying and classifying toxic comments. However, while these models achieve high overall accuracy, they often display unintended identity bias by misclassifying comments that mention certain identities (gender, race, or religion) as toxic, even when they are not intended to be. This can occur in cases where online users utilize reclaimed slurs or counter-speech online, and can be unfairly flagged for using terms in a neutral or positive light. This bias can have real-world consequences, such as inadvertently silencing the marginalized communities that the models were built to protect against online hate.
To explore this misclassification, we compare three domain-adapted pre-trained models, Detoxify (Wikipedia), roberta-hate-speech-dynabench-r4-target, and Twitter-RoBERTa Offensive (Twitter), using the 2019 Jigsaw Unintended Bias dataset which contains identity markers as features.
The scope of our proposed work includes:

1. Evaluating the three pre-trained models on the Jigsaw Unintended Bias dataset.
2. Measuring overall performance (accuracy, F1) and fairness using subgroup bias metrics.
3. Analyzing and visualizing which models perform most fairly across different identity categories.

Through this analysis, we hope to understand not just which model is most accurate, but which is most fair, and how domain pretraining can potentially impact unintended bias in online toxicity detection.

2 Related Work

Our first piece of relevant prior work is *A Comparative Study of Using Pre-trained Language Models for Toxic Comment Classification* (Zhoa et Al, 2021). This paper compares three pre-trained models, BERT, RoBERTa, XLM, for toxic comment classification (TCC) across several datasets and found that BERT/RoBERTa generally outperform XLM (Zhao et. al, 2021). We will be further exploring two of these three models, roberta-hate-speech-dynabench-r4-target and Twitter-RoBERTA, in our own research.
However, the comparison conducted by Zhao et al. only targets overall TCC accuracy and does not analyze unintended identity bias. In order to address this, we will be comparing our pre-trained models on the 2019 Jigsaw Uninteded Bias Dataset, which contains labels for different identity markers, such as gender, religion, race and others to quantify fairness across these identity subgroups.

Our second piece of relevant prior work is HATECHECK: Functional Tests for Hate Speech Detection Models (2021). This paper complements our work by 29 small, targeted tests for hate-speech models (reclaimed slurs, negated hate, and counter-speech) which we can utilize to properly compare our models effectiveness. Instead of simply looking at the overall F1 score, these tests can help identify where our pre-trained models falsely flag neutral or positive identity mentions as toxicity (Röttger et. al, 2021).

We will use the HATECHECK research study to gain a better understanding on how to run group-specific checks and quantify unintended identity bias. We will then use our findings from this study to create an evaluation plan that ensures accuracy in our data and metrics.

3 Dataset

We used the Jigsaw Unintended Bias dataset, which contains comments from the Civil Comments platform, that collected public comments on English-language news sites between 2015 and 2017. These comments are labeled for toxicity along with identity related labels such as gender, race, religion and sexual orientation in the dataset. It is publicly available on Kaggle and can be downloaded as a CSV file by creating an account on Kaggle. We compared each model's predictions with the human-annotated true labels in the dataset to assess how accurately they detect toxicity and whether they show bias toward identity-related comments.

3.1 Data Pre-processing

We ran preliminary experiments on random samples of 10,000 instances, which revealed that performance metrics were unstable for identity groups with very few examples. To address this issue, we narrowed our scope to focus specifically on race and gender-related identities, and constructed two separate balanced datasets for these categories.

For each balanced dataset, we first selected the relevant identity indicator columns (race-related or gender-related). We then filtered the data to keep only comments in which a given identity was explicitly mentioned (i.e., entries with a value greater than 0). To balance the representation across identity groups, we computed the number of rows in each group and identified the smallest group size. Each identity subset was then randomly sampled to this minimum size, ensuring equal representation.

Finally, the sampled subsets were concatenated and shuffled to remove any ordering effects and bias. The balanced datasets contain ∼10,000 gender-related data and ∼34,000 race-related data.

3.2 Sample Size By Subgroup Identity

These graphs show the sample sizes of each subgroup of gender identities and race identities.

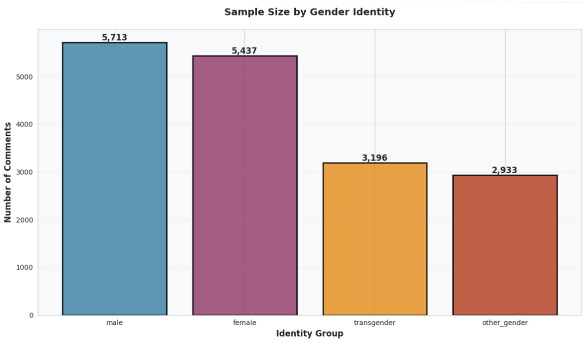


Figure 1: Sample Size by Gender Identity

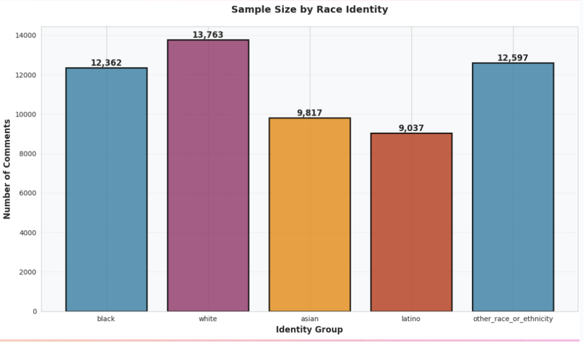


Figure 2: Sample Size by Race Identity

4 Methods

We evaluated three publicly available transformer-based NLP toxicity detection models on the 2019 Jigsaw Unintended Bias dataset to examine variation in performance across different identities. Our original model set included Detoxify Original, HateBERT, and Twitter-RoBERTa Offensive. However our preliminary experiments revealed that, HateBERT assigned nearly all comments to the non-toxic class, resulting in precision, recall, and F1 scores of 0.0. This indicates that the model is not well-aligned with toxicity detection on this dataset, we replaced it with Facebook's RoBERTa-based hate speech model (roberta-hate-speech-dynabench-r4-target), introduced by Vidgen et al. (2021). None of the evaluated models were trained on the Jigsaw data, so this gives us a proper generalization and bias assessment.

The final set of evaluated models is as follows:

- **Detoxify Original**: A transformer-based model trained on the 2018 Jigsaw Toxic Comment Classification dataset. It provides a baseline for how a Wikipedia-trained toxicity classifier behaves on identity-related language.
- **roberta-hate-speech-dynabench-r4-target**: A RoBERTa-based classifier trained on the Dynabench dataset, which consists of dynamically collected hate speech examples.
- **Twitter-RoBERTa Offensive**: A RoBERTa-based model fine-tuned on abusive and disrespectful tweets from Twitter, optimized for social media style text.

All models were run using HuggingFace's pipeline interface. For each comment, we obtained the model's predicted probability of toxicity and assigned a binary label using a threshold of 0.5. Model predictions were compared against the ground-truth target toxicity labels in the Jigsaw dataset. We evaluated each model on two balanced subsets (gender focused and race focused). All these experiments can be found hyperlinked here: NLP Repository

5 Evaluation Metrics

To assess each model's performance, we computed accuracy, precision, recall, and F1 score for each identity label using the binary toxicity labels. We computed the false positive rate (FPR) for comments containing identity terms grouped by gender and race to analyze fairness and potential identity-related bias. FPR is particularly important in toxicity detection because it captures cases where non-toxic comments are incorrectly flagged as toxic, disproportionately affecting marginalized identities. We did not include a traditional machine learning baseline because our focus is on comparing the behavior of pre-trained transformer models.

6 Results and Analysis

6.1 Detoxify Original Results

Identity	Count	Precision	Recall	F1	Accuracy	FPR
asian	9,817	0.6986	0.3233	0.4420	0.8951	0.0206
black	12,362	0.8028	0.3096	0.4469	0.7847	0.0297
latino	9,037	0.7568	0.3096	0.4395	0.8586	0.0217
white	13,763	0.7684	0.2917	0.4229	0.7862	0.0323
other_race_or_ethnicity	12,597	0.7160	0.2923	0.4151	0.8658	0.0226

Table 1: Identity metrics by race

Identity	Count	Precision	Recall	F1	Accuracy	FPR
male	5,713	0.6640	0.3603	0.4672	0.8650	0.0358
female	5,437	0.6494	0.3590	0.4624	0.8606	0.0389
transgender	3,196	0.6621	0.3217	0.4330	0.8411	0.0382
other_gender	2,933	0.7000	0.3313	0.4497	0.8657	0.0282

Table 2: Identity metrics by gender

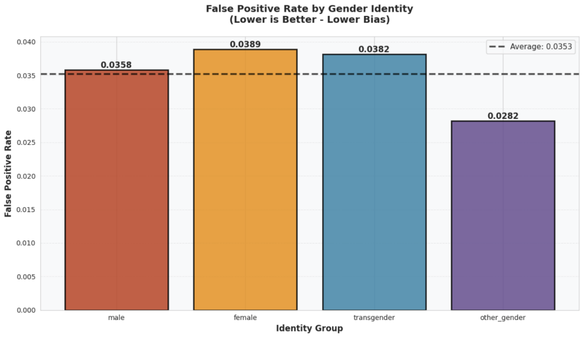


Figure 3: FPR of Gender Identity for Detoxify

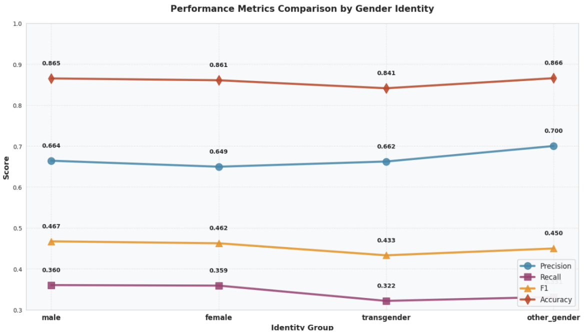


Figure 4: Performance Metrics by Gender Identity for Detoxify

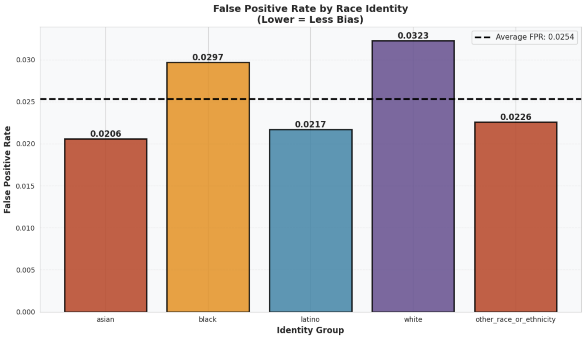
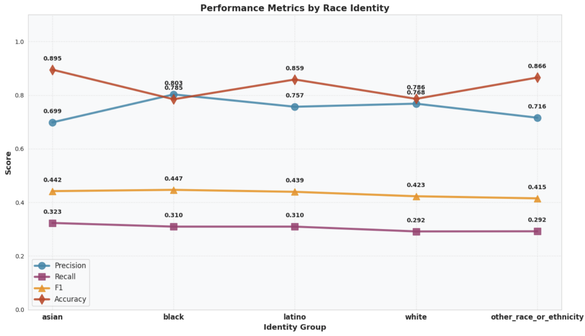


Figure 5: FPR of Race Identity for Detoxify



r

Figure 6: Performance Metrics by Race Identity for Detoxify

6.2 Detoxify Original Analysis

Across racial identity groups, Detoxify exhibits consistently high precision (0.70–0.80) but low recall (0.29–0.32), which reveals that the model is more cautious in predicting toxicity when racial identities are mentioned. In other words, the low FPR values across all race identities (2–3 percent) tells us that when the model predicts a comment as toxic, it is often correct, but it fails to identify a large fraction of truly toxic comments.

The gender-based results are similar to the race-based trends but with slightly higher recall and F1 scores overall, suggesting that Detoxify performs marginally better on gender-related content. False positive rates are slightly higher for gender identities than for race identities (2.8–3.9 percent), with female and transgender groups exhibiting the highest FPRs. This suggests that non-toxic comments mentioning gender may be more likely to be incorrectly flagged as toxic than those mentioning race.

Across both race and gender analyses, Detoxify Original demonstrates an importance on precision by prioritizing avoiding false positives, resulting in low FPRs, but at the expense of recall. This tradeoff is especially evident in identity-related content, where the model often fails to flag toxic comments that reference minority groups.

6.3 roberta-hate-speech-dynabench-r4-target Results

Identity	Count	Precision	Recall	F1	Accuracy	FPR
black	12,362	0.3901	0.3592	0.3740	0.6623	0.2193
white	13,763	0.3500	0.2993	0.3227	0.6626	0.2040
asian	9,817	0.1883	0.4398	0.2637	0.6842	0.2797
latino	9,037	0.2683	0.3096	0.2875	0.7252	0.1841
other_race_or_ethnicity	12,597	0.2341	0.3648	0.2852	0.7020	0.2324

Table 3: Identity metrics by race

Identity	Count	Precision	Recall	F1	Accuracy	FPR
male	5,713	0.2937	0.2953	0.2945	0.7677	0.1395
female	5,437	0.2940	0.3238	0.3082	0.7572	0.1559
transgender	3,196	0.3095	0.4229	0.3574	0.7131	0.2194
other_gender	2,933	0.3183	0.3621	0.3388	0.7658	0.1541

Table 4: Identity metrics by gender

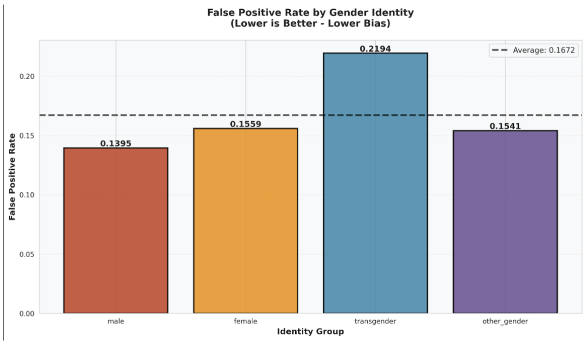


Figure 7: FPR of Gender Identity for Dynabench

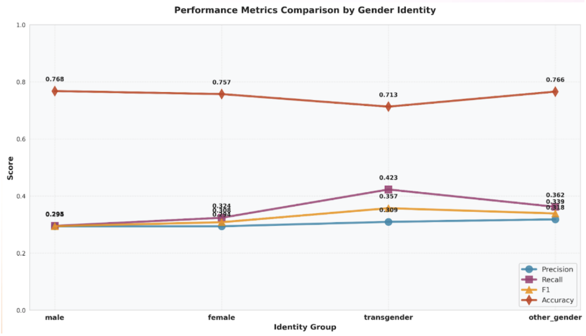


Figure 8: Perfomance Metrics of Gender Identity for Dynabench

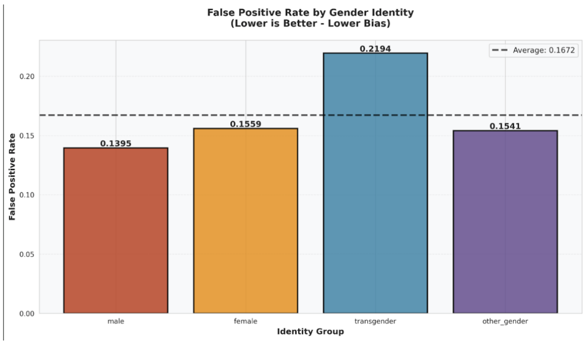


Figure 9: FPR of Race Identity for Dynabench

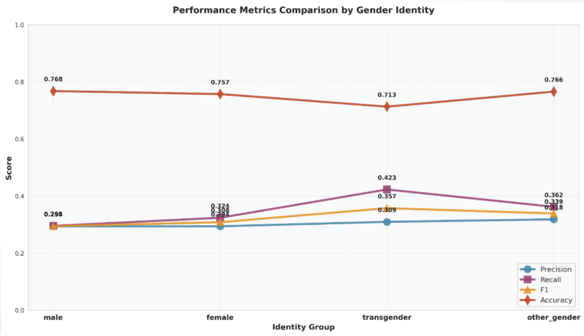


Figure 10: Perfomance Metrics of Race Identity for Dynabench

6.4 roberta-hate-speech-dynabench-r4-target Analysis

Across racial identities, the Dynabench R4 model exhibits low precision and moderate recall, resulting in modest F1 scores across all groups. Precision ranges from 0.19 to 0.39, indicating that a large fraction of comments that mention race are predicted as toxic when they actually are not non-toxic.

Despite balanced sample sizes across race groups, false positive rates vary considerably, suggesting that these disparities aren't driven by sampling bias, but by the model's sensitivity to context clues. The model over-predicts toxicity for comments referencing racial identities, particularly Asian and Black identities.

Gender-based evaluation reveals a similar pattern of high recall paired with low precision, with uneven error rates across gender identities. The transgender identity demonstrates the highest recall (0.4229) and FPR (0.2194), indicating that comments mentioning transgender identities are significantly more likely to be incorrectly flagged as toxic than comments including male or female identities.

The Dynabench R4 model prioritizes recall over precision, aggressively flagging potentially harmful content. While this increases sensitivity to hate speech, it also leads to over-flagging comments with identity mentions, especially for marginalized groups. High false positive rates across both race and gender identities indicate that the model often correlates identity references with toxicity.

6.5 Twitter-RoBERTA Results

Identity	Count	Precision	Recall	F1	Accuracy	FPR
black	12,362	0.3252	0.9150	0.4798	0.4428	0.7416
white	13,763	0.3101	0.9315	0.4654	0.4253	0.7604
asian	9,817	0.1735	0.9065	0.2912	0.4328	0.6371
latino	9,037	0.2399	0.9135	0.3800	0.4663	0.6312
other_race_or_ethnicity	12,597	0.2076	0.9250	0.3391	0.4125	0.6873

Table 5: Identity metrics by race

Identity	Count	Precision	Recall	F1	Accuracy	FPR
male	5,713	0.2201	0.9094	0.3545	0.4562	0.6329
female	5,437	0.2223	0.9053	0.3569	0.4552	0.6350
transgender	3,196	0.2433	0.8988	0.3829	0.4534	0.6502
other_gender	2,933	0.2204	0.8909	0.3533	0.4596	0.6261

Table 6: Identity metrics by gender

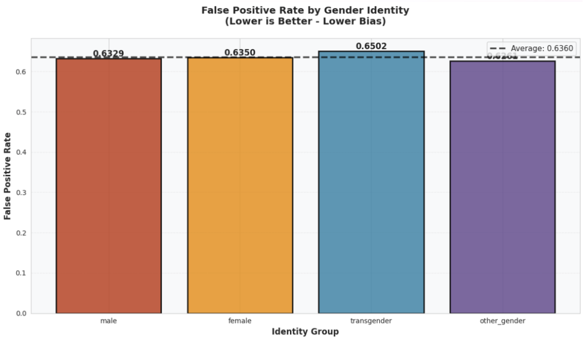


Figure 11: FPR of Gender Identity for Twitter-RoBERTA

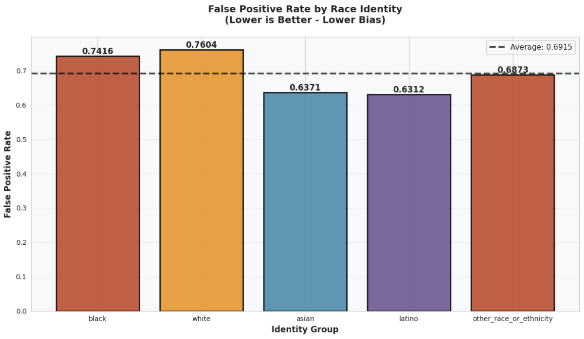


Figure 12: FPR of Race Identity for Twitter-RoBERTA

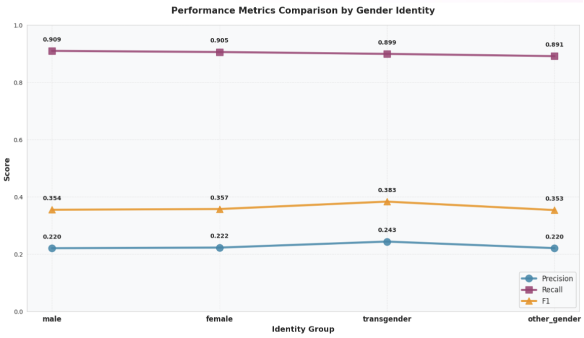


Figure 13: Perfomance Metrics of Gender Identity for Twitter-RoBERTA

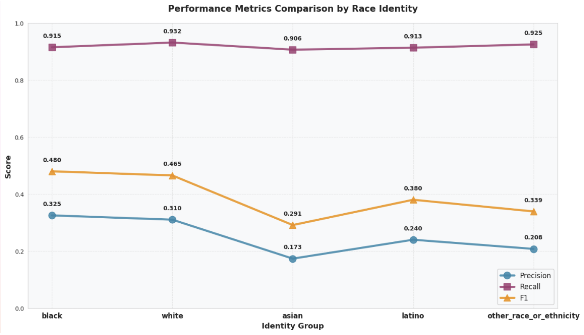


Figure 14: Perfomance Metrics of Race Identity for Twitter-RoBERTA

6.6 Twitter-RoBERTA Analysis

Twitter-RoBERTa exhibits extremely high recall across all identity groups (0.89–0.93) but very low precision (0.17–0.33). This indicates over-flagging, where the model aggressively labels comments that mention identity as toxic, resulting in a large number of false positives. While this strategy minimizes missed toxic content, it substantially harms reliability and fairness for identity-mentioning text.

Across racial identity groups, recall is consistently high, exceeding 0.90 for Black, White, Asian, Latino, and Other Race categories. However, precision remains low, particularly for Asian (0.1735) and Other Race or Ethnicity (0.2076), meaning that the majority of comments predicted as toxic for these groups are actually non-toxic. However, False positive rates are high across all racial identities, which demonstrates that Twitter-RoBERTa strongly associates racial identity mentions with toxicity.

A similar pattern is demonstrated across gender identities. Recall remains very high for all groups while precision is uniformly low (0.22–0.24), indicating that most toxic predictions are incorrect. As a result, F1 scores remain modest but false positive rates are high across all subgroups.

Compared to more conservative models, Twitter-RoBERTa trades precision and fairness for recall. While it rarely misses toxic content, this comes at the cost of bias, where identity mentions are treated as toxic regardless of context.

7 Comparative Analysis

7.1 Dataset Composition and Sample Sizes

We evaluated all three models on the same balanced datasets to ensure the comparison is fair. The gender dataset contained 10,892 comments distributed across four identity groups: male (5,713), female (5,437), transgender (3,196), and other gender (2,933). The race dataset comprised 13,469 comments with more balanced representation: white (13,763), black (12,362), Asian (12,597), Latino (12,597), and other race/ethnicity (12,597). This consistent data composition allowed for direct model comparison while controlling for dataset effects.

7.2 False Positive Rate Analysis

From figures above, we can see that the false positive rates showed clear differences in model behavior across different identity groups. Detoxify showed the lowest FPR overall, with a narrow range of 2-3% across both gender and race datasets. In contrast, Dynabench R4 showed moderate FPRs ranging from 13-21% for gender identity (with transgender at the highest) and 18-27% for race identity (peaking at 27% for Asian comments). Twitter-RoBERTa had the highest bias, with FPRs ranging from 63-65% for gender and 63-76% for race, which are much higher than both competing models. These differences suggest that Twitter-RoBERTa is often confused and incorrectly labels harmless comments that include identity-related terms as toxic.

7.3 Performance Metrics and Model Behavior

Twitter-RoBERTa showed a pattern of aggressive toxicity detection as we just saw from the high FPR rates. Across both the datasets, recall was very high (89–90%), while precision and F1 scores were substantially lower (around 35%). This trade-off indicates that the model prioritizes catching all toxic comments, even at the cost of numerous false positives. The low precision of the model means that it frequently misclassifies innocent identity comments as toxic. Although the race dataset showed deeper drops in precision and F1 score, which means there is a stronger racial bias, overall the model was consistent across both the datasets.

Dynabench R4 performed at an intermediate level. For the gender dataset, accuracy reached approximately 76%, but precision, recall, and F1 scores remained low (30–40%). The race dataset showed further degradation, with accuracy dropping to 66%. The Asian identity comments triggered higher recall, indicating that the model attempted to catch more toxic instances, but since it also had low precision, many innocent comments were incorrectly flagged. The resulting F1 scores (≈26% for race) reflect this inconsistent performance.

Detoxify Original gave the most balanced and reliable results. For the gender dataset, accuracy remained consistently high across all identity groups (≈86%), with precision significantly outperforming both models at 60–70%. Recall was comparatively lower (≈35%), which suggests that the model used a more conservative approach and only flagged high-confidence toxic predictions. This conservative strategy resulted in fewer false positives while maintaining reasonable true positive detection, therefore getting F1 scores around 45–47%. Similar patterns emerged in the race dataset, where the accuracy ranged from 80–89%, precision was highest (69–78%), and recall remained low (≈30%), producing F1 scores of 42–44%.

7.4 Subgroup Disparities and Data Imbalance

We found that in all the models smaller and underrepresented identity groups consistently performed worse than larger groups. Specifically, transgender, Asian, and other race/ethnicity categories showed lower accuracy and higher false positive rates. This pattern directly correlates with dataset representation where groups with fewer training examples showed greater performance degradation. Also, since these minority groups often appear in more linguistically varied and adversarial contexts within the data, this compounds the bias effect. This means that the imbalance in the data for each subgroup is the main reason behind accuracy and fairness problems.

7.5 Race vs. Gender Bias

We also observed that race-related bias was much higher than gender-related bias across all the models. Twitter-RoBERTa's FPR ranged from 0.63-0.76 for race compared to 0.62-0.65 for gender. Dynabench R4 showed its peak FPR at 0.28 for Asian comments, which was the highest subgroup FPR across all datasets and models. Even Detoxify, the most balanced model, showed a gap when evaluating race identity comments compared to gender identity comments, though the absolute FPRs remained low.

7.6 Overall Performance

According to our evaluation, Detoxify is the most reliable toxicity detector with the smallest subgroup gaps and lowest FPR. Dynabench R4 occupied a middle ground with moderate accuracy but concerning subgroup FPR disparities. Twitter-RoBERTa was the most biased, with high FPRs and showed a high possibility of relating certain identity mentions with toxicity. However, as we can see from the above results, no model achieved true fairness and all three demonstrated measurable bias against underrepresented groups.

8 Future Work

To extend this study, future evaluations can incorporate balanced identity data for all subgroups, not just gender and race. It would be valuable to evaluate intersectional subgroups, such as Black woman, Asian transgender, Latino male etc. Bias is often amplified in these cases because models rarely encounter intersectional identity combinations during training. Prior work illustrates this issue: for example, the Alphabet Perspective API has been shown to reach an 87% false positive rate for the phrase "I am a gay Black woman," with error rates increasing as more identity attributes are combined.

9 References

Jigsaw Unintended Bias in Toxicity Classification. 2019. Kaggle Dataset. Available at: https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification

TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification https://arxiv.org/abs/2010.12421v2

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)*. https://doi.org/10.18653/v1/2021.acl-long.4

Zhao, Z., et al. (2021). *A Comparative Study of Using Pre-trained Language Models for Toxic Comment Classification.* In *Companion Proceedings of the Web Conference 2021*, 500–507. https://doi.org/10.1145/3442442.3452313