ADS Audit: Assessing Performance and Fairness in Comment Toxicity Classification

Veronica Zhao, Wang Xiang

May 10, 2024

1 Background

The purpose of the Automated Decision System (ADS) is to effectively detect and classify toxic comments in online environments. The primary goals of this ADS are that it ensures high accuracy in toxicity detection and minimize unintended model bias. Thus, in this report, we aim to further analyze the ADS to maintain high accuracy performance while reducing biases related to identity attributes such as gender, race, sexual orientation, and disability. These are among the much broader aims within a competition taken by Jigsaw/the Conversation AI [1], reflecting the theme of responsible development and incorporation of the tools of data science in society and market, by setting an aim to develop a tool to compete with existing societal biases.

However, achieving the goals presents challenges, especially the potential trade-off between accuracy (measured by the Area Under the Curve/AUC in this competition) and fairness. High accuracy may be reached at the cost of fairness because the model would have ranked predictive power over equal treatment of all identity groups. This ADS seeks to explore and reconcile these competing objectives.

The following audit aims to evaluate the effectiveness of the ADS in managing these trade-offs through a critique of its effectiveness at bias reduction, validation process, and its efficiency in evaluating the toxicity assessment among demographic groups. The insights gained from this evaluation will contribute to the ongoing conversation about the intersection of AI, ethics, and social impact, particularly in the field of text analytics.

2 Input and Output

The datasets in our analysis include public comments sourced from the Civil Comments platform, curated by Jigsaw for the competition's use. The training set is constituted by 1804874 unique textual comments with associated toxicity scores (annotated by various human annotators), identities of users, and metadata. The testing set is constituted by 97320 textual comments.

The output is the target column, a float score ranging from 0.0 to 1.0, with a value larger or equal than 0.5 being positive (toxic) label for accuracy evaluation.

The data type and the number of missing values of each feature can be observed from Figure 1 below. The ADS has only used *comment text* as the feature and *target* as the output, which are object and float respectively. Other features are provided by the competition for further research. All identity features are the probability score (float) of the user belonging to a certain identity and are all extracted by human annotators, so most of the comments have no identity data.

```
Data columns (total 46 columns):
 #   Column                                 Dtype
---  ------                                 -----
 0   id                                     int64
 1   target                                 float64
 2   comment_text                           object
 3   severe_toxicity                        float64
 4   obscene                                float64
 5   identity_attack                        float64
 6   insult                                 float64
 7   threat                                 float64
 8   asian                                  float64
 9   atheist                                float64
 10  bisexual                               float64
 11  black                                  float64
 12  buddhist                               float64
 13  christian                              float64
 14  female                                 float64
 15  heterosexual                           float64
 16  hindu                                  float64
 17  homosexual_gay_or_lesbian              float64
 18  intellectual_or_learning_disability    float64
 19  jewish                                 float64
 20  latino                                 float64
 21  male                                   float64
 22  muslim                                 float64
 23  other_disability                       float64
 24  other_gender                           float64
 25  other_race_or_ethnicity                float64
 26  other_religion                         float64
 27  other_sexual_orientation               float64
 28  physical_disability                    float64
 29  psychiatric_or_mental_illness          float64
 30  transgender                            float64
 31  white                                  float64
 32  created_date                           object
 33  publication_id                         int64
 34  parent_id                              float64
 35  article_id                             int64
 36  rating                                 object
 37  funny                                  int64
 38  wow                                    int64
 39  sad                                    int64
 40  likes                                  int64
 41  disagree                               int64
 42  sexual_explicit                        float64
 43  identity_annotator_count               int64
 44  toxicity_annotator_count               int64
 45  word_count                             int64
dtypes: float64(32), int64(11), object(3)
memory usage: 633.4+ MB
```

```
Missing Values:
id                                          0
target                                      0
comment_text                                3
severe_toxicity                             0
obscene                                     0
identity_attack                             0
insult                                      0
threat                                      0
asian                                 1399744
atheist                               1399744
bisexual                              1399744
black                                 1399744
buddhist                              1399744
christian                             1399744
female                                1399744
heterosexual                          1399744
hindu                                 1399744
homosexual_gay_or_lesbian             1399744
intellectual_or_learning_disability   1399744
jewish                                1399744
latino                                1399744
male                                  1399744
muslim                                1399744
other_disability                      1399744
other_gender                          1399744
other_race_or_ethnicity               1399744
other_religion                        1399744
other_sexual_orientation              1399744
physical_disability                   1399744
psychiatric_or_mental_illness         1399744
transgender                           1399744
white                                 1399744
created_date                                0
publication_id                              0
parent_id                              778646
article_id                                  0
rating                                      0
funny                                       0
wow                                         0
sad                                         0
likes                                       0
disagree                                    0
sexual_explicit                             0
identity_annotator_count                    0
toxicity_annotator_count                    0
dtype: int64
```

Figure 1: Data Types and Missing Values

Word Clouds of comments in different levels of toxicity score can be observed in Figure 2. From Figure 3, we can notice that the word count of each comment ranges from 0 to 200 and skewed to the left, indicating most comments being a paragraph.

Figure 2: Word Clouds of All, Low Toxicity, High Toxicity Comments

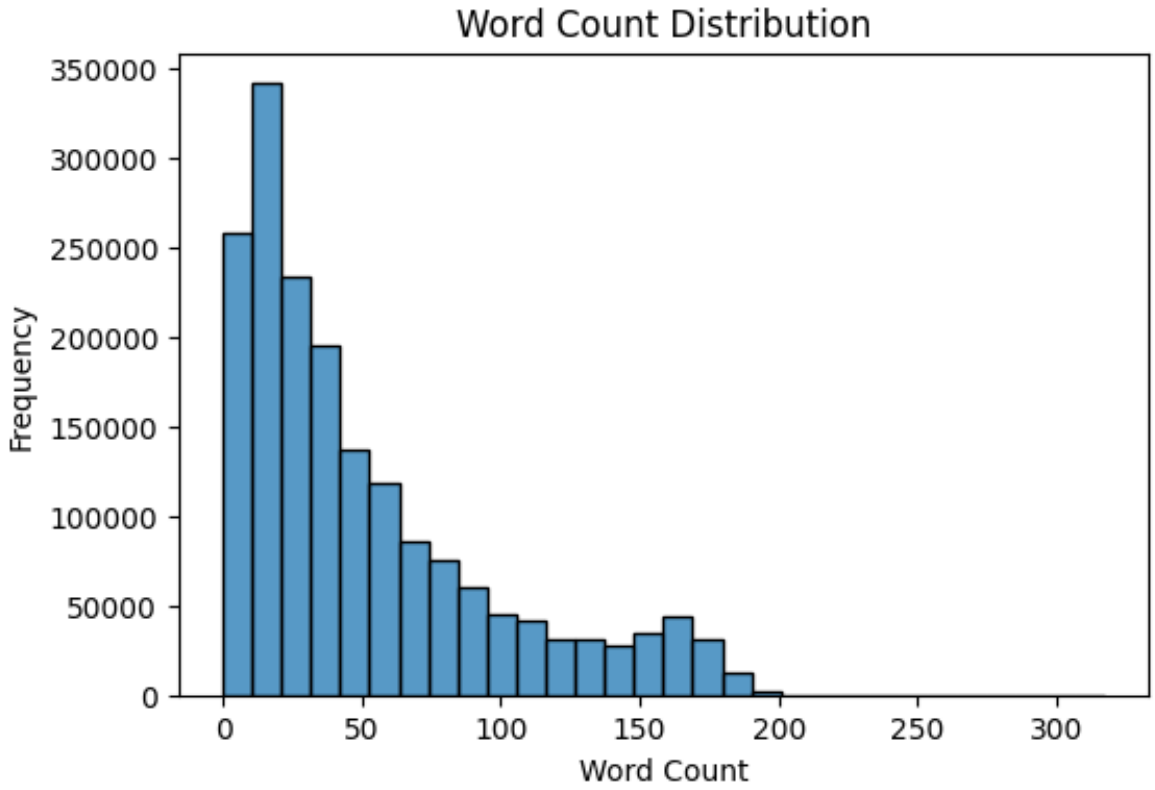

Figure 3: Distribution of Word Counts

From Figure 4, we can examine a high correlation between insult and target, and a moderate correlation between other toxicity labels and target.

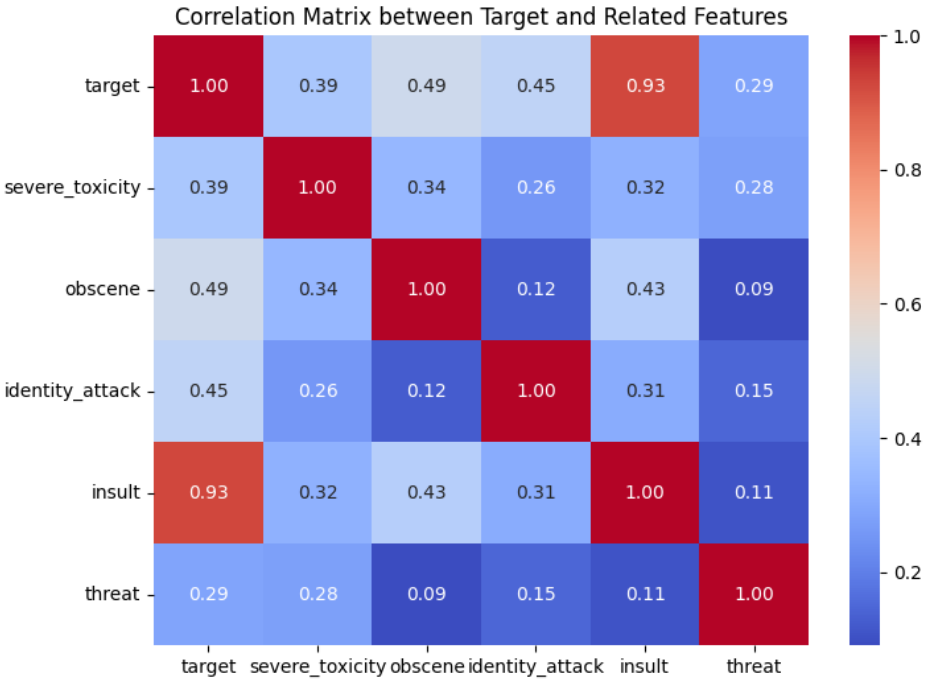

Figure 4: Correlations of Toxicity Score and Labels

Figure 5 and Figure 6 illustrates the distributions of the target variable and groups of demographic data. It can be observed that the dataset is mostly labelled as negative (less than 0.5), which means it is an imbalanced dataset and may cause problem in the prediction process.

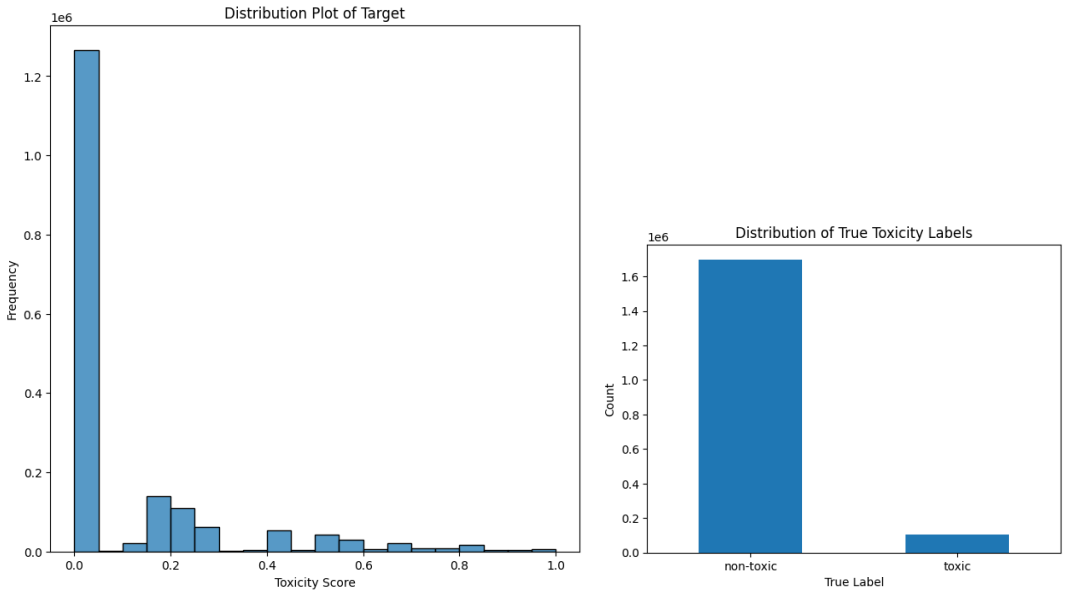

Figure 5: Distribution of Target (Toxicity Scores/Labels)

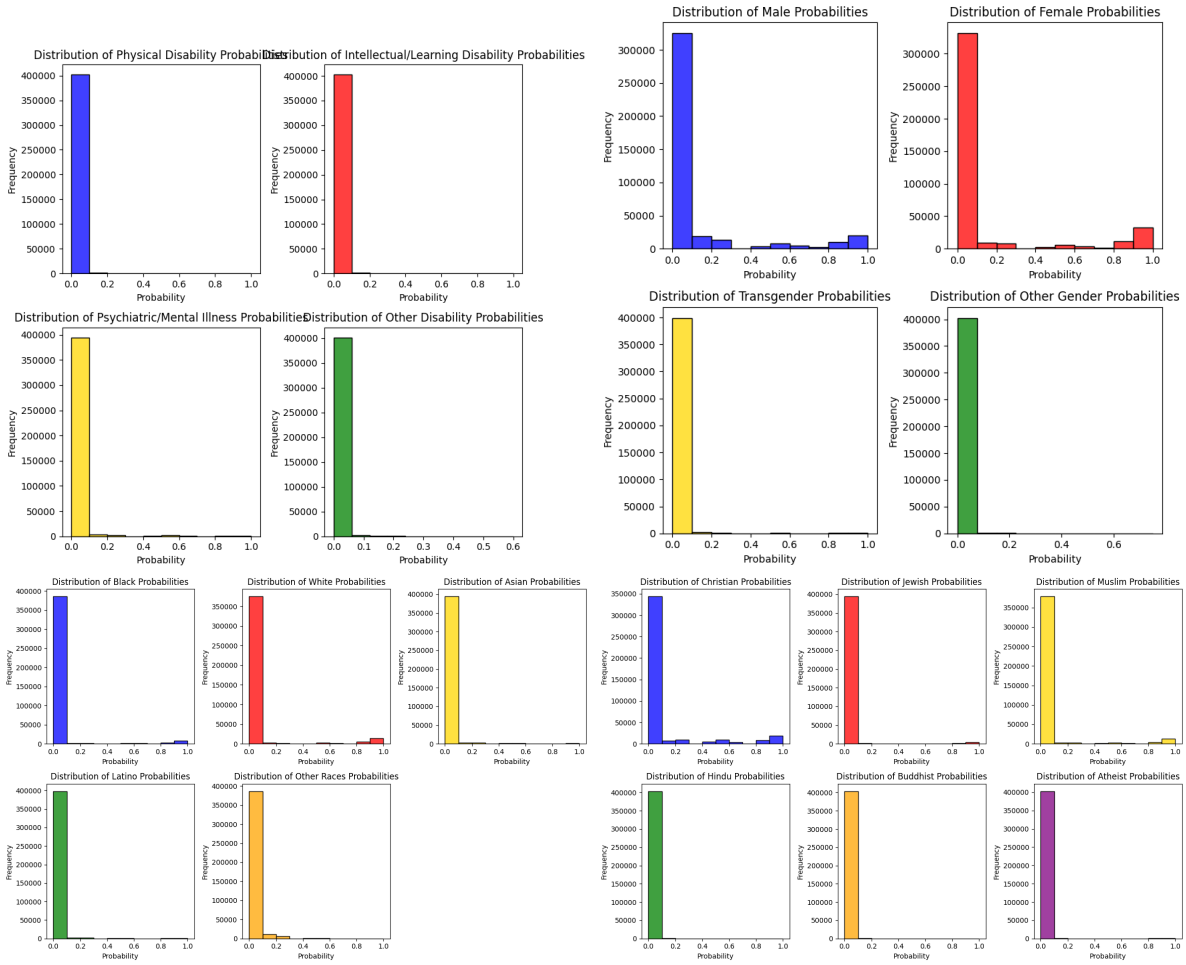

Figure 6: Distribution of Identity Information

3 Implementation and Validation

We plan to adopt the approach by the 8th-ranked participant on the Kaggle leaderboard, who utilized an ensemble method combining predictions from multiple BERT models, along with an XLNet model and a GPT model, to assess toxicity. The model's accuracy has already been evaluated on the test set in the challenge with an Area Under Curve (AUC) of 0.9466. However, in order to evaluate accuracy on sensitive groups, we need to use the privately and publicly expanded test sets with demographic labels. Thus, we concatenate the 2 test sets as they are not overlapped.

For data cleaning, all rows without sensitive labels are removed from training set and test set. Then irrelevant columns such as *created date* and *article id* are removed. Lastly, all subtypes of demographic groups are inversely one-hot encoded that each column includes the categorical value of each group (E.g. In *race* column, 'asian'=1, 'white'=2, 'black'=3).

Afterwards, the chosen ADS is implemented on the new test set for performance evaluation. It firstly preprocesses the comments by isolating symbols, abbrevations, and using Treebank Tokenizer, a standard tokenizer that utilizes regular expressions to tokenize texts. Then it loads the 4 pre-trained language models and takes a weighted average of the predictions as the result.

As the AUC of the model is 0.9466, indicates that the model has a 0.9466 chance of correctly distinguishing between toxicity and non-toxicity comments in the test set, demonstrating strong predictive accuracy and effective separation between the classifications.

However, we need to further examine the accuracy and fairness measures among different demographic groups, and if it is performing equally well for both positive and negative classes.

For further accuracy and fairness evaluation, we evaluated Accuracy, Precision, Recall, F1 score, ROC AUC score, FNR (false negative rate), FPR (false positive rate) as the accuracy measure; we evaluated false negative difference, false positive difference, demographic parity ratio, equalized odd ratio, and selection rate difference, on the prediction values both overall and within groups using the Fairlearn package.

4 Outcomes

- The high accuracy of 0.78 indicates that a large proportion of all predictions (toxic and non-toxic) are correctly classified. It's a straightforward metric to evaluate the overall performance of the model but can be misleading since the class distribution is imbalanced.
- A precision of 0.333 and a recall of 0.98 are crucial for evaluating the effectiveness of identifying toxic comments. The lower precision with higher recall suggests the model identifies a large part of potential toxic comments but at the cost of including more false positives.
- The F1-Score is 0.497 as a moderate balance of precision and accuracy, which implies there's room for improvement in identifying toxicity.
- The high ROC-AUC score of 0.868 suggests the model does relatively well at distinguishing between toxic and non-toxic classes across all thresholds.
- FNR and FPR are crucial for understanding model biases. A low FNR of 0.02 is desirable that toxic content is hard to slip through, while a higher FPR of 0.245 indicates it is more often for non-toxic comments to be misclassified as toxic, possibly limiting free speech.

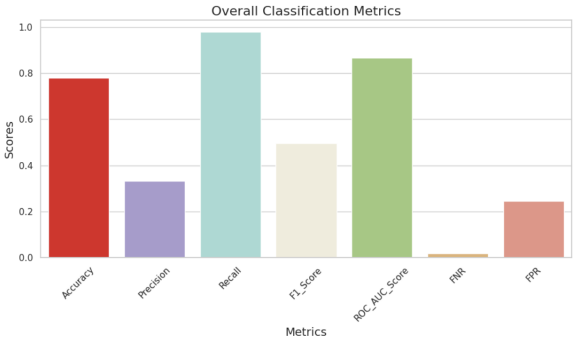

Figure 7: Overall Accuracy Metrics

For accuracy over subpopulation:

- In sex groups, performance metrics like precision and recall does not vary notably between male, female, and other genders. However, the accuracy measures and FPR of transgender group is distinct from other groups.
- Significant disparities are evident among race groups, especially in accuracy, precision, ROC AUC, and FPR among black and white groups, which suggest potential biases in detecting toxicity which could misrepresent these racial groups either by over-flagging or under-flagging comments.
- The model performs unevenly across different religious groups among all metrics, especially for accuracy, precision, and FPR. The FPR is highest for Muslim group, which potentially indicates a bias in recognizing toxicity in comments related to religions.

- There's a clear disparity in how the model performs across different sexual orientations, with gay or lesbian and heterosexual groups showing lower performance metrics, particularly in recall and F1-score. While bisexual group is more accurately classified.
- There are more variation in precision, F1-scores, and FPR in disability groups, indicating inconsistencies in how well it identifies and confirms cases within specific groups like intellectual or learning which has notably lower performance.

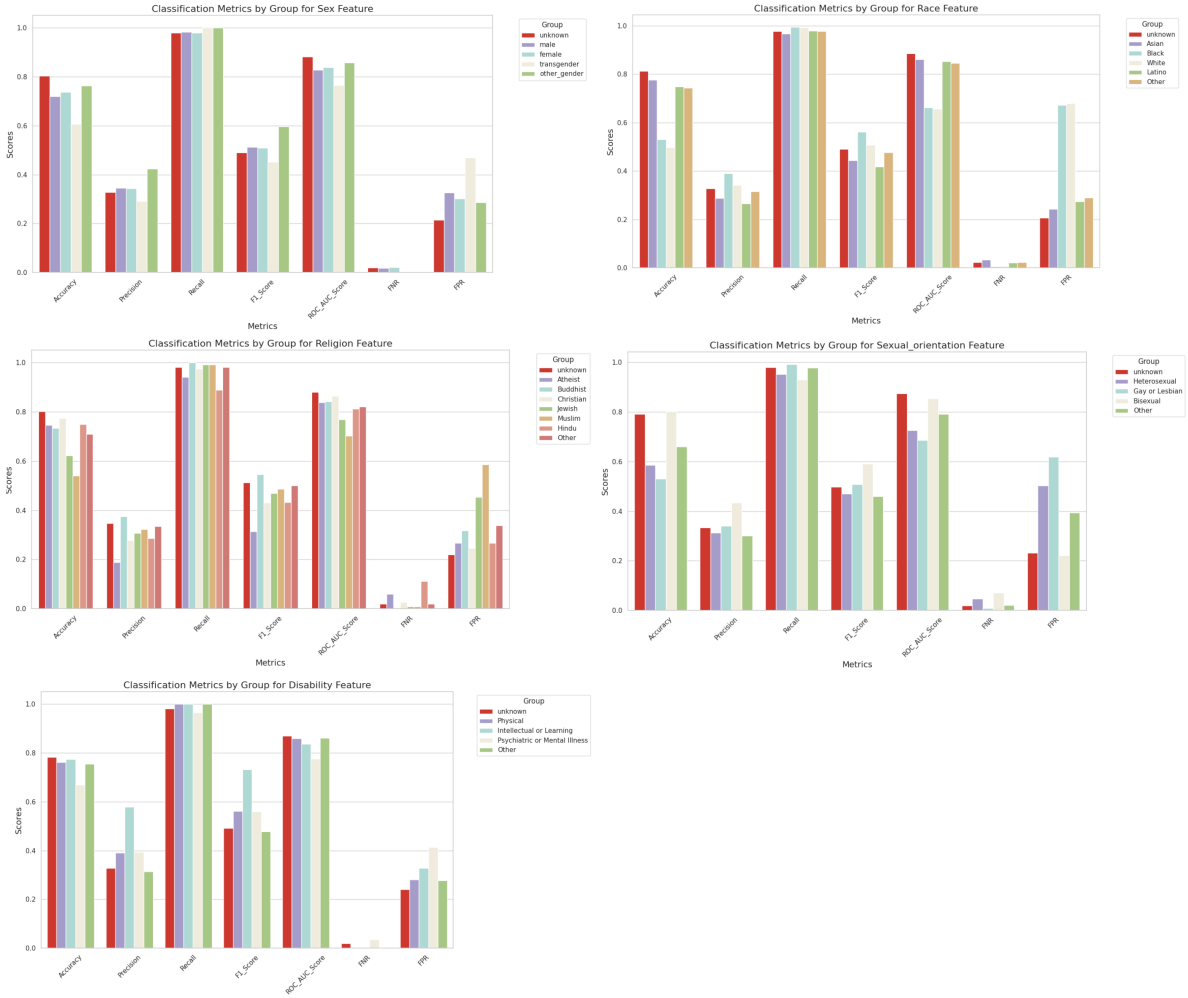

Figure 8: Accuracy in Sensitive Groups

According to the fairness metrics plot, very low FNR difference is achieved among all sensitive features, while FPR difference remains highest and demographic parity ratio remains lowest for different racial groups, which indicates that the model is consistent in minimizing missed detection of toxicity across all groups but struggles with over-predicting toxicity and unequal prediction probabilities among different races, potentially leading to racial bias in the classifications. The equalized ratio is highest for disability groups while lowest for racial groups, and this trend also applies to selection rate. To conclude, the model performs fairer for disability groups but more bias against racial groups.

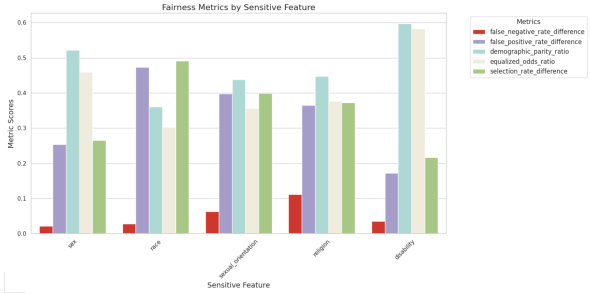

Figure 9: Fairness Measure

To analyze the performance of the ADS and gauge its stability, we utilized the Population Stability Index (PSI), alongside an evaluation of the model's accuracy on marginal cases. The PSI, a measure derived from information theory, effectively quantifies the stability of a model by detecting shifts in data distributions between the model's development and validation datasets. In our analysis, a low PSI value (0.0005) indicates minimal distributional changes, affirming that the model performs consistently over time with no significant drift that might necessitate recalibration. This consistency suggests the model is robust against the dynamic variations often seen in real-world scenarios.

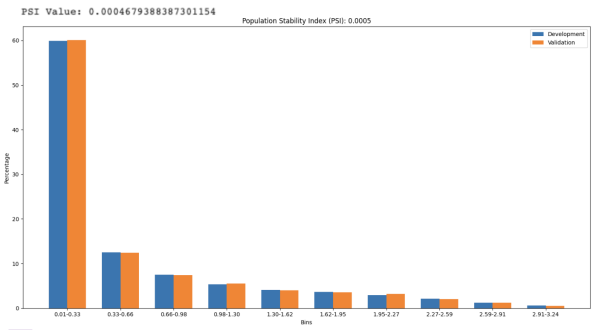

Figure 10: Stability Measure

On the other hand, the evaluation of the model's performance on marginal predictions, where it achieved an accuracy of approximately 54.13% (not much better than random guess), revealed some challenges. We defined marginal cases as cases with predicted toxicity score between 0.4 to 0.6. Despite a high precision and recall for the majority class, there's a noticeable discrepancy in predicting the minority class effectively. This performance gap highlights that the model, when faced with marginal examples, may disproportionately or directly predict as negative (nontoxic comments). This tendency not only raises concerns about the model's ability to handle edge cases but also indicates a serious risk of misclassification in situations where accurate identification of positive cases is crucial, potentially leading to detrimental outcomes in practical applications.

Considering both the PSI analysis and the performance on marginal cases, while the model exhibits

Performance on marginal cases:
Accuracy: 0.5412844036697247

	precision	recall	f1-score	support
0	0.97	0.54	0.69	3055
1	0.04	0.59	0.08	106
accuracy			0.54	3161
macro avg	0.51	0.57	0.39	3161
weighted avg	0.94	0.54	0.67	3161

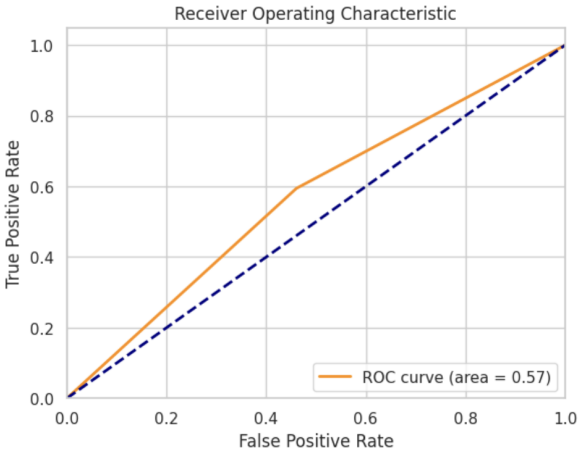

Figure 11: Important Examples

stability, the issue with marginal predictions highlights the need for ongoing monitoring and possibly refining the model to handle edge cases better. This holistic approach to model evaluation, focusing on both stability and detailed performance metrics, ensures a balanced view that addresses both accuracy and fairness, vital for stakeholders relying on the model's outputs for strategic and operational decisions.

5 Summary

The quality of data is appropriate for this ADS as it incorporates a huge number of textual comments and human annotated labels. However, it is very imbalanced for the toxicity labels, thus we can propose further improvement by:

- Oversampling/Upsampling: increase the number of instances in the positive class by duplicating them before training each base model, making class distribution more balanced and allowing each model to better learn the characteristics of the minority class.
- Adding Class Weights: setting higher class weights for the positive class to effectively increase the penalty for misclassifying these instances.

The implementation of the model has been scrutinized for robustness, accuracy, and fairness across multiple sensitive features including sex, race, sexual orientation, religion, and disability. Each feature was assessed using a comprehensive set of metrics like accuracy, precision, recall, F1-score, ROC_AUC score, and false-negative and false-positive rates. Additionally, fairness metrics such as the false negative rate difference, false positive rate difference, demographic parity ratio, equalized odds ratio, and selection rate difference were employed to evaluate the model's fairness. The analysis reveals significant variations across groups, suggesting potential biases that could impact stakeholders differently. For instance, policymakers and regulators in healthcare or financial sectors might find these metrics particularly crucial to ensure equitable service delivery. While the model shows promising accuracy and ROC_AUC scores, the fairness assessments indicate a need for further calibration to minimize bias and enhance reliability across all groups. Furthermore, the PSI is implemented for stability analysis and the marginal cases are evaluated with confusion matrix. Users who post comments that are borderline or ambiguous in terms of toxicity (scores between 0.4 and 0.6) are most likely to be harmed by this model's performance. Due to low precision and F1-score for the positive class, these users may experience their content being unjustly flagged or removed, potentially stifling free expression and causing frustration from perceived unfair treatment.

It is more appropriate to apply this ADS in the entertainment industry, but it still need to be improved to be deployed comfortably since even the overall accuracy is relatively high, the FPR also remains around 0.2, which indicates the possibility of misclassifying non-toxic comments to toxic, therefore

limiting users' freedom of speech on the platform.

For potential improvement, it is possible to use other tokenizers that are more effective to this context, such as the Tweet Tokenizer in NLTK, which may be better at processing text in this social media context. Moreover, the ADS uses a weighted average of prediction scores from each language model without elaboration. The ADS may be improved after evaluating each model's performance and use a more considered weight in the calculation. Considering other approaches, TF-IDF vectorizor and other classification machine learning models may also be viable in this situation and worth trying.

6 Project Contribution

Wang Xiang's contributions encompassed researching the topic, conducting exploratory data analysis (EDA), implementing the algorithmic decision system (ADS), cleaning the data, evaluating accuracy and fairness, and generating the report and presentation slides. Meanwhile, Veronica Zhao engaged in topic research, identified a suitable ADS, performed EDA, implemented the ADS, evaluated accuracy and fairness, assessed stability and marginal performance, and refined the content of the report and slides.

References

[1] inversion Jeffrey Sorensen Lucas Dixon Lucy Vasserman nithum cjadams, Daniel Borkan. Jigsaw unintended bias in toxicity classification, 2019.