ADS Audit: Assessing Performance and Fairness in Comment Toxicity Classification

DS-UA 202 Project by Wang Xiang, Veronica Zhao

05/08/24

Background Information

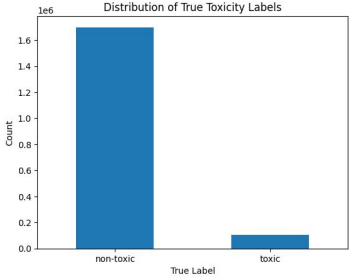
- Jigsaw Unintended Bias in Toxicity Classification Challenge
- Purpose of the ADS: ensures high accuracy in detecting toxic comments with minimizing unintended model bias
- Data Source: public comments collected from Civil Comments platform
- Challenge: Accuracy vs Fairness (Trade-off)
- Our Goal
    - Analyze the challenge and the ADS's performance in both accuracy and fairness
    - Proposing viable methods of maintaining high accuracy with more bias across demographic groups

NYU

Exploratory Data Analysis

- Input: User's textual comments
- Output: Toxicity score (0~1), label=toxic if score>=0.5
- True label: toxicity score annotated by human annotators
- Sensitive features: sex, race, religion, sexual orientation, disability identified from comments by human annotators

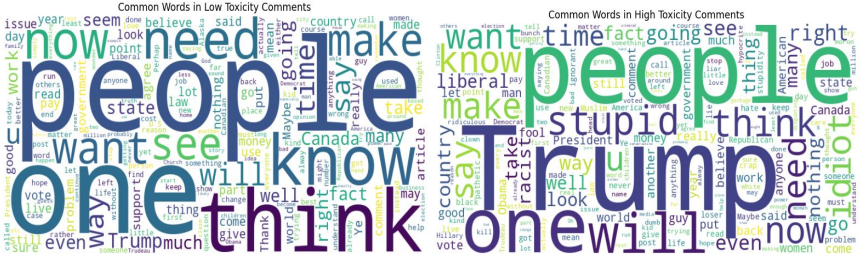
Observation 1: Imbalanced dataset



Observation 2: Fairness can only be measured in a subset of dataset

- # of rows in test set: 194640
- # of rows with labels in test set: 42870

Observation 3: Word Cloud



NYU

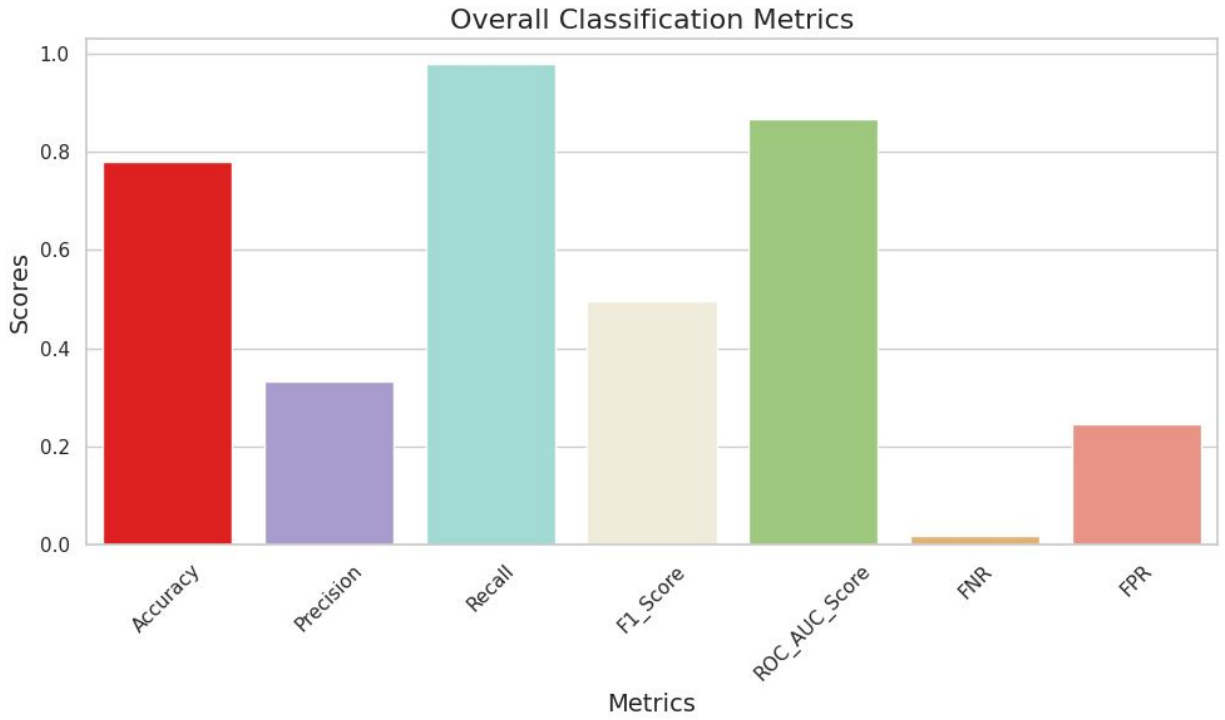
Automated Decision System

- Approach: 8th ranked participant on challenge leaderboard
  - Text Preprocessing
  - Ensemble method combining predictions from BERT-base-uncased, BERT-large-uncased, GPT2, XLNET models (weighted average)
- Performance:
  - AUC: 0.9466

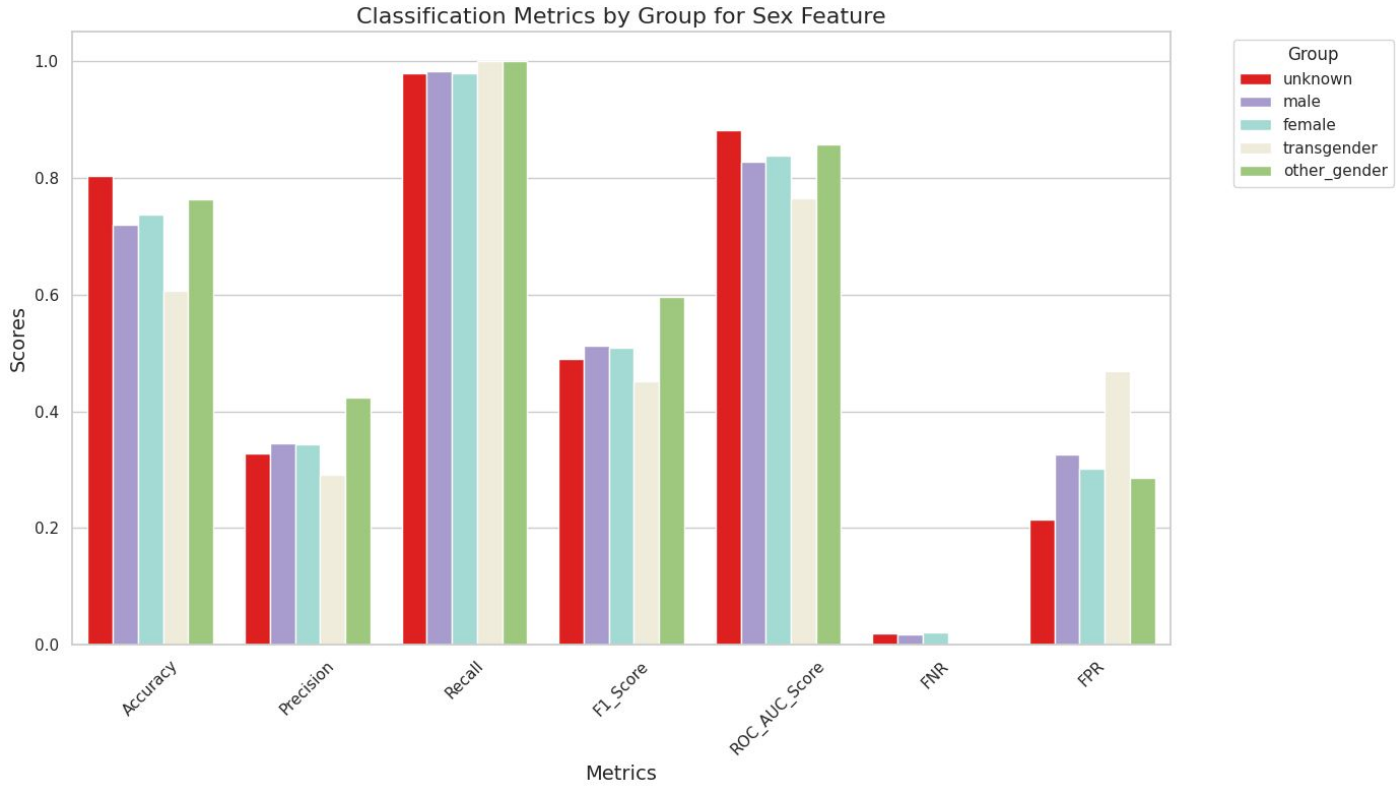
Preprocessing and Implementation

- Data Cleaning
  - Remove rows without sensitive labels
  - Drop irrelevant columns
  - Concatenate 2 expanded test sets
- Inverse one-hot encoding
  - Convert several dummy sensitive labels into a single categorical feature
- Implementation of the ADS on the test set
- Performance and fairness evaluation

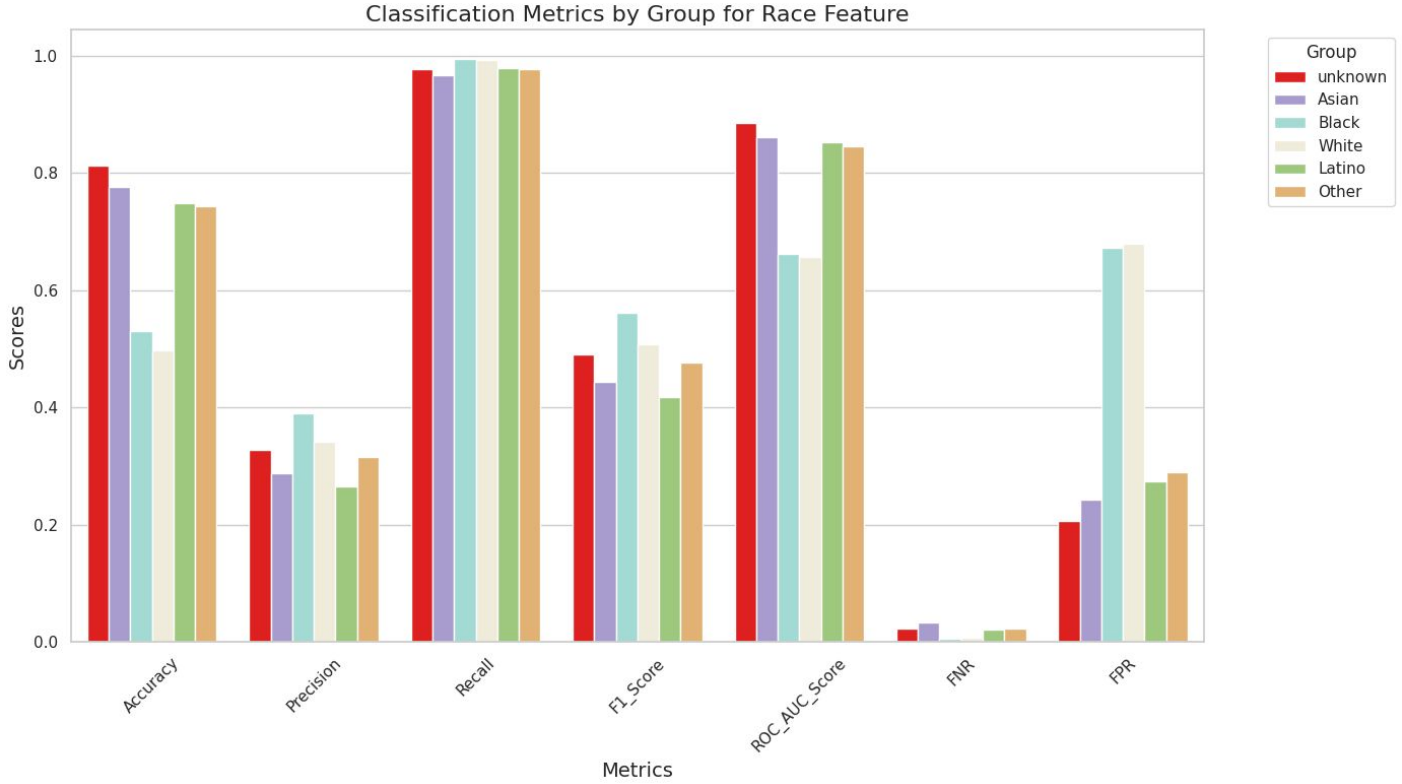
Overall Accuracy Metrics



Accuracy Metrics by Group

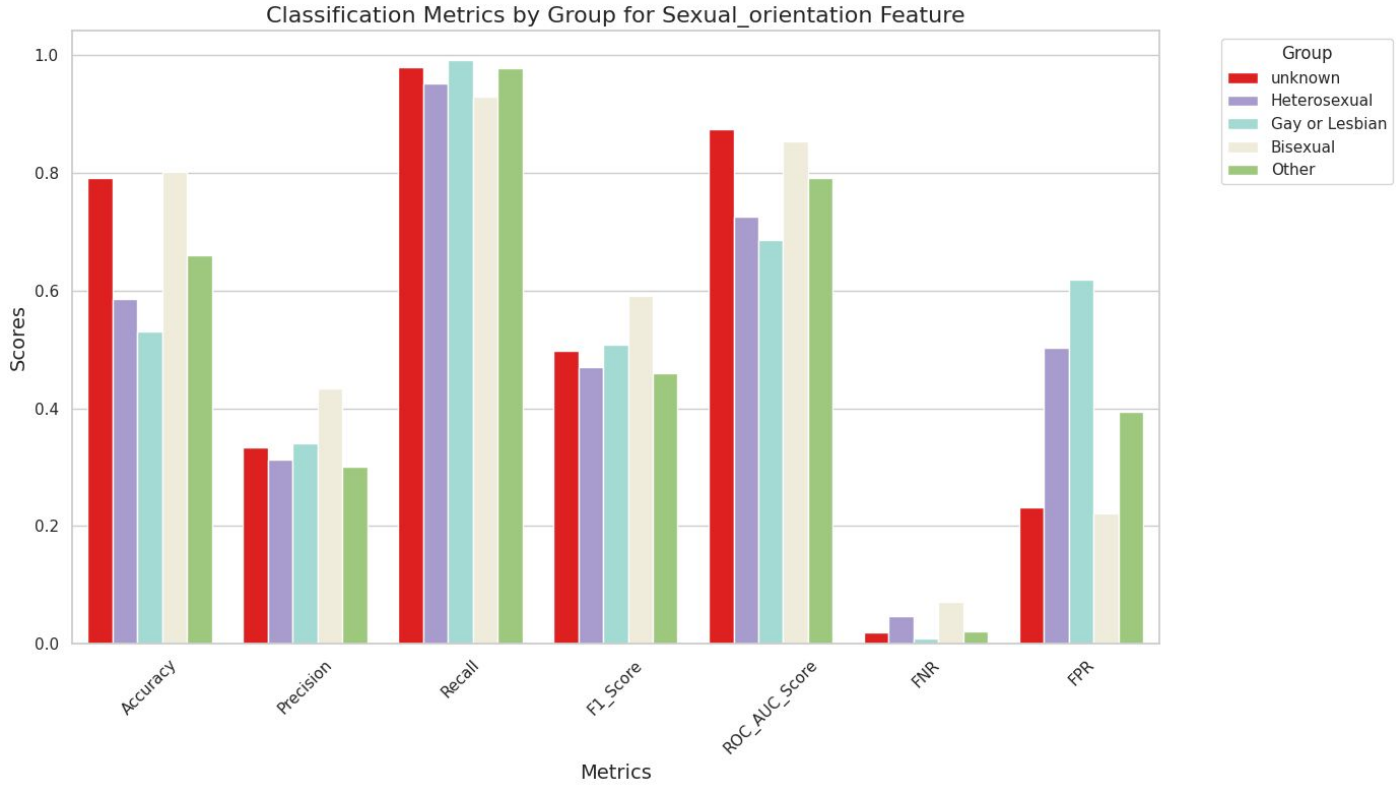


Accuracy Metrics by Group

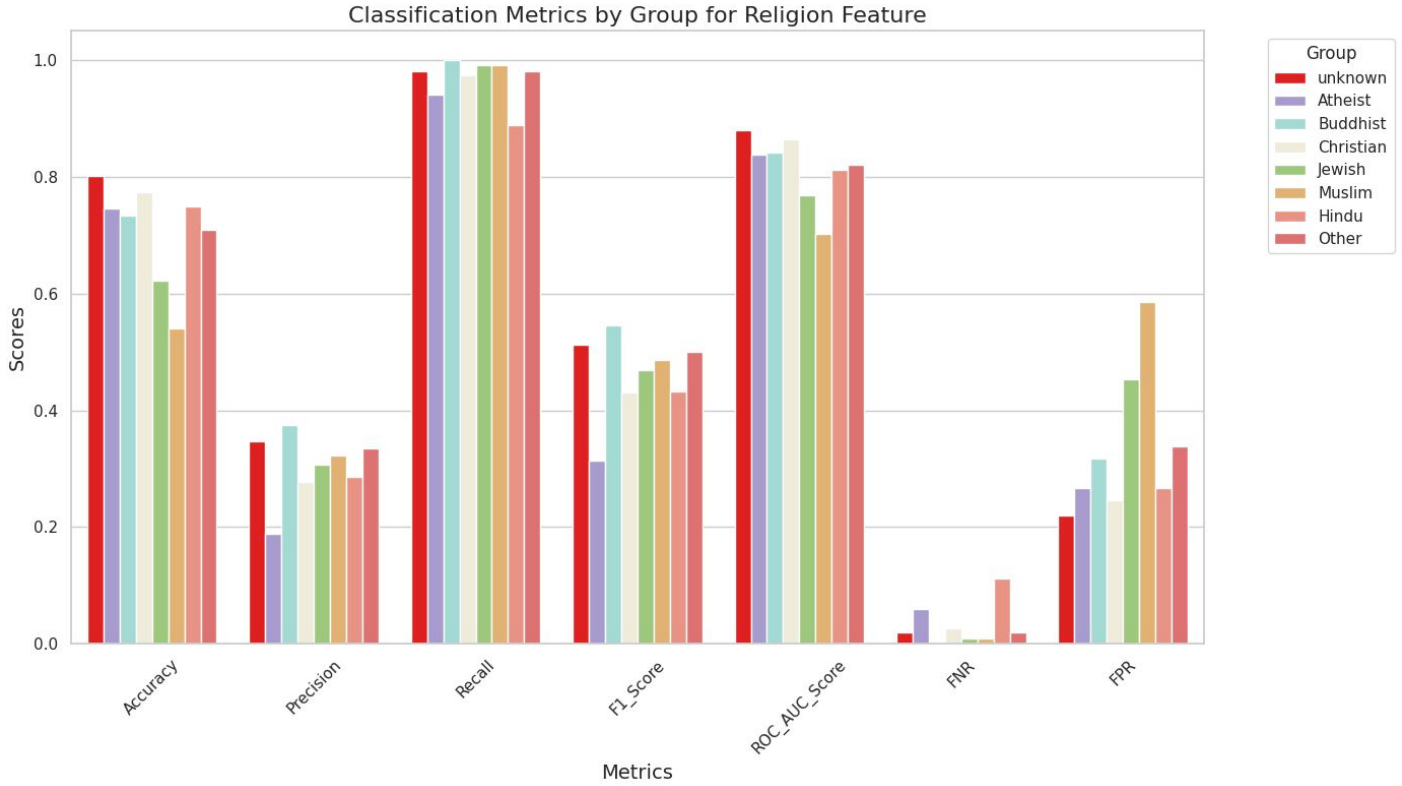


NYU

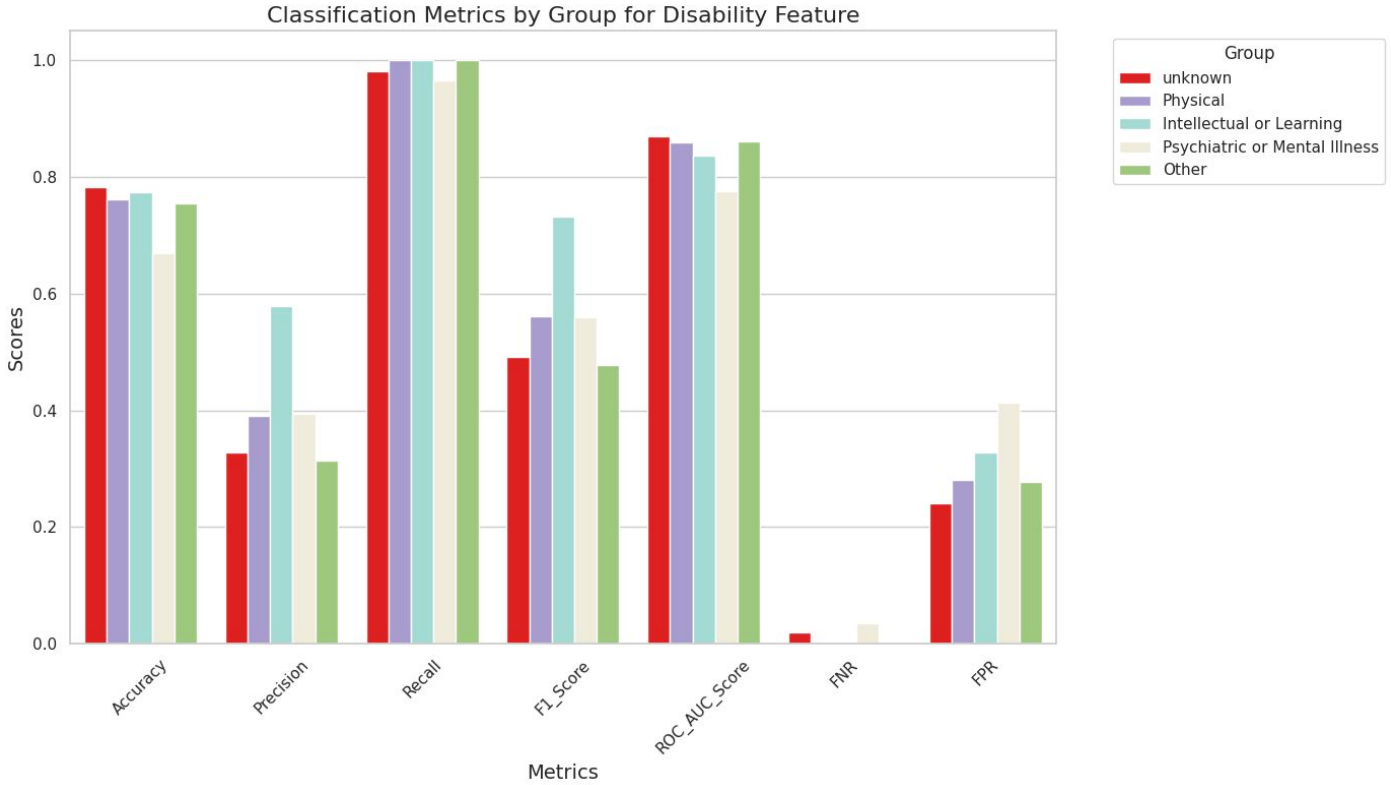
Accuracy Metrics by Group



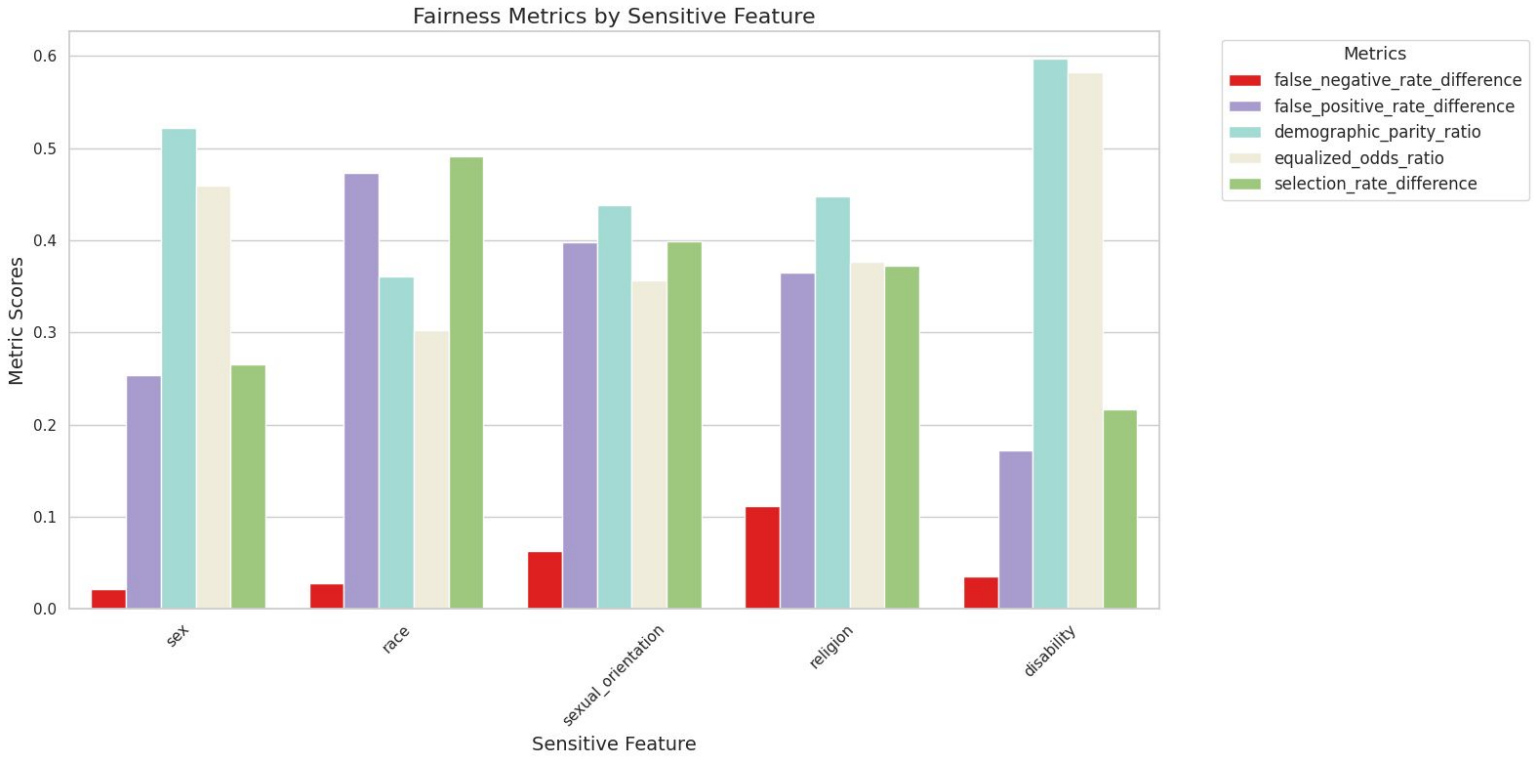
Accuracy Metrics by Group



Accuracy Metrics by Group



Fairness Metrics by Sensitive Feature



Introduction to Population Stability Index (PSI)

https://www.aporia.com/learn/data-science/practical-introduction-to-population-stability-index-psi/#

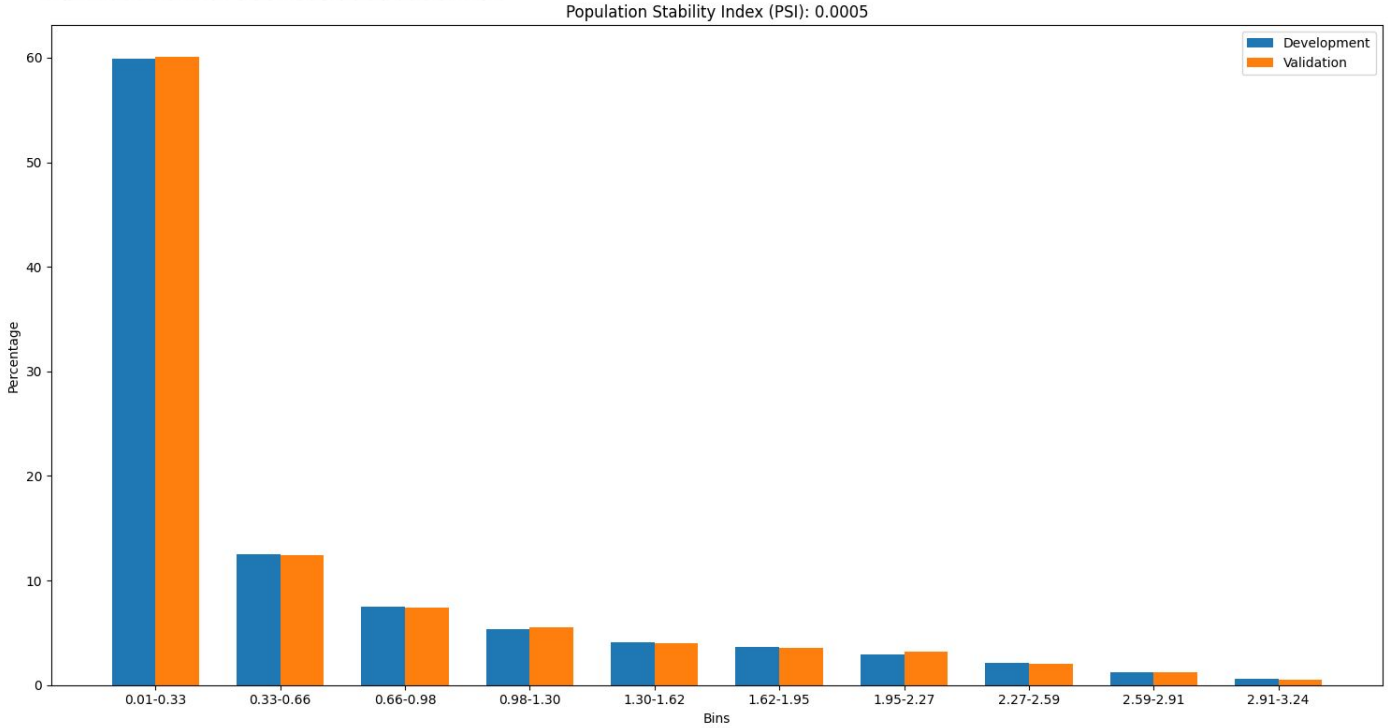
- **Definition:**
  PSI is a measure used to assess the stability of a model by comparing the distribution of a variable between a development dataset and a validation dataset.

- Importance of PSI: It helps detect shifts in data distribution, indicating potential model drift, which can affect the model's predictive accuracy and reliability.

- Application: Regular use in monitoring models, especially in sectors like finance for credit scoring, to ensure models remain accurate over time.

PSI Analysis and Interpretation

PSI Value: 0.0004679388387301154



PSI Interpretation Guide

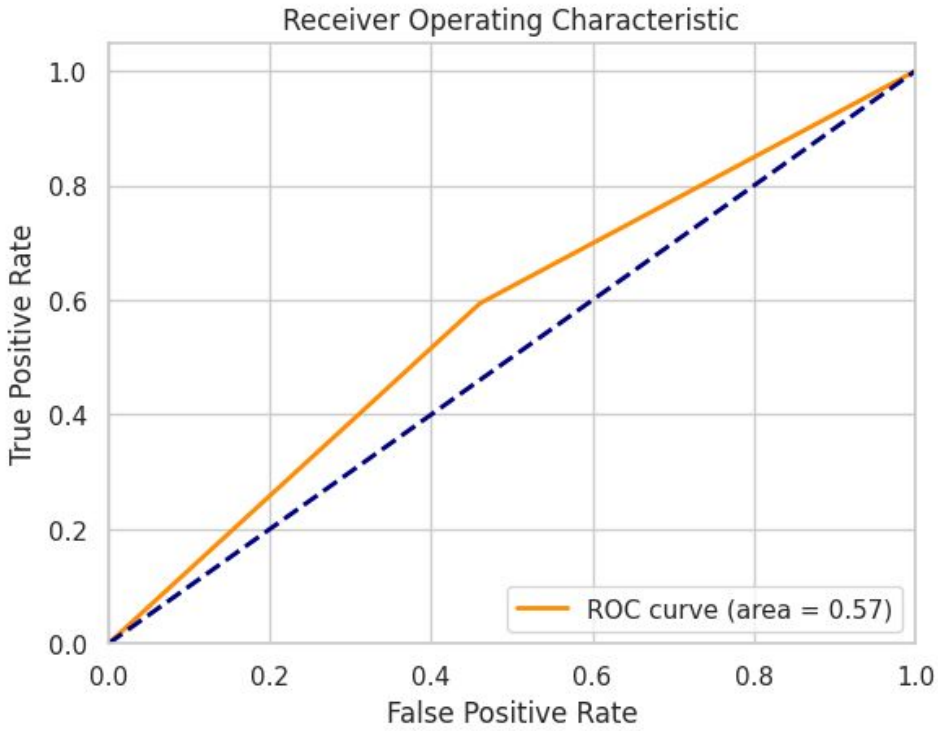
PSI Value	Description
PSI < 0.1	Very low change between the two groups, considered stable.
0.1 <= PSI < 0.25	High change between the two groups, is considered significant.
PSI >= 0.25	High change between the two groups, considered significant.

- ➔ Bisn shows nearly identical percentages, demonstrating high stability with minimal differences.
- ➔ The low PSI value affirms no significant data shifts, suggesting the model's predictions remain reliable.
- ➔ No recalibration needed at present. Continuous monitoring is advised to maintain model accuracy.

Performance on Marginal Predictions

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



NYU

Conclusion

- Stakeholder: Comment platform, users
- ADS should be further improved for deploy in industry
- Accuracy: overall performs well, FPR remains relatively high
- Fairness: should be improved especially for racial groups
- Potential Improvement
  - Imbalanced dataset - oversampling / undersampling
  - Tokenizers for social media platform - TweetTokenizer in NLTK
  - Revise the prediction calculation method

Thank you!