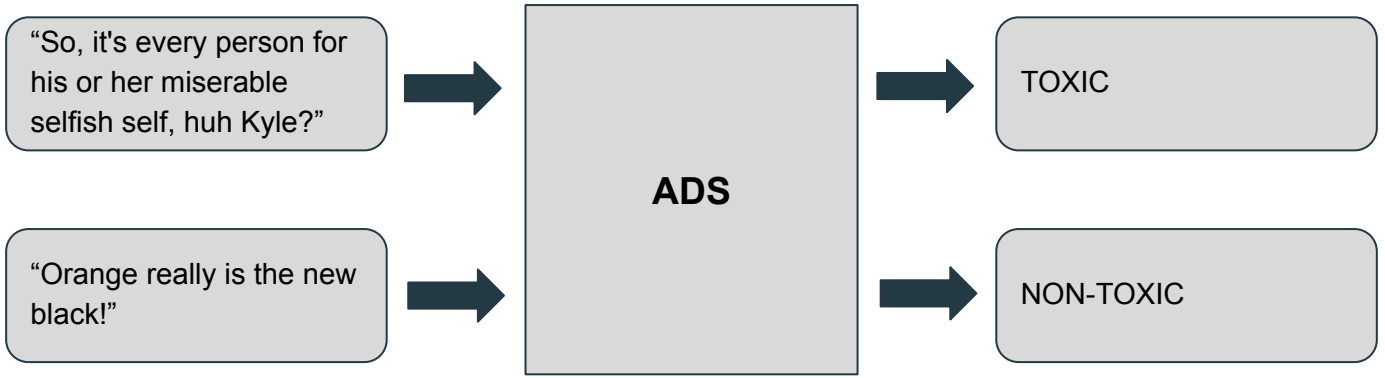
ADS Audit: Unintended Bias in Toxicity Classification

Chloe Zheng, Preston Harry
DS-GA 1017 Spring 2022

Background

Kaggle Competition: Jigsaw Unintended Bias in Toxicity Classification

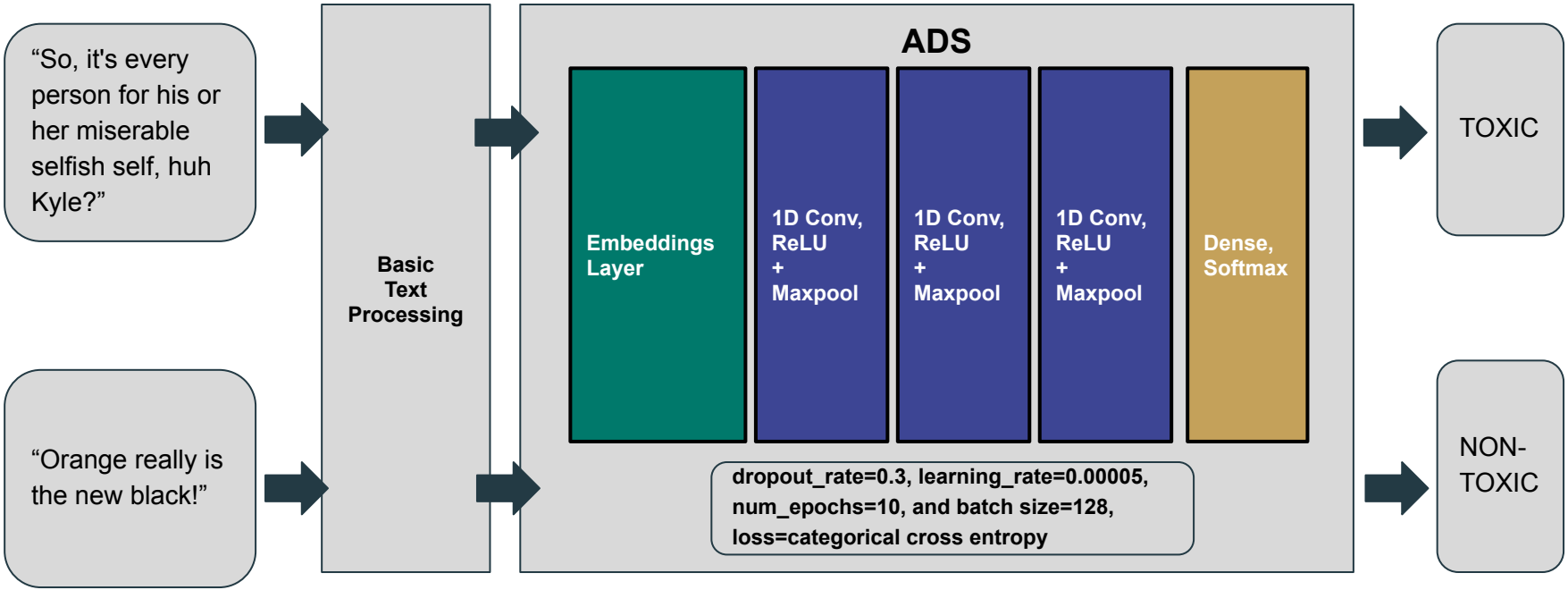
- Goal: implement a model that classifies toxicity of comments while minimizing unintended bias with respect to mentions of identities
- Proposed Solution: Benchmark Kernel



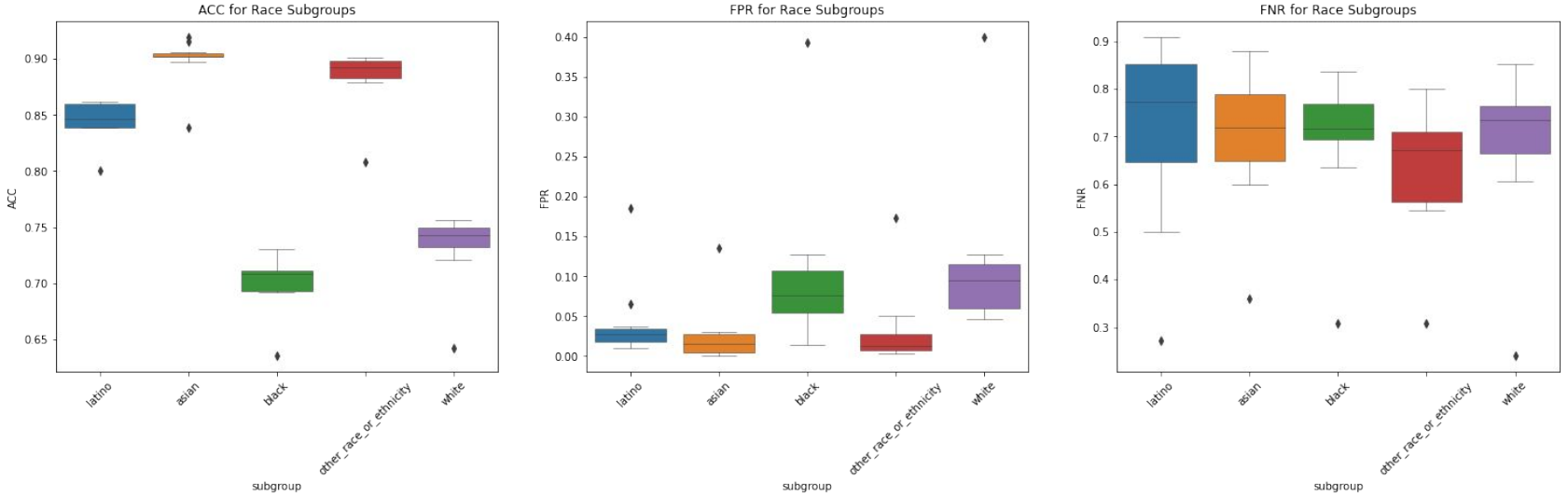
Data

- Civil Comments
  - Comment text
  - Binary classification
- Jigsaw Data Pre-processing
  - Annotators
    - Type of toxicity: Insult, Obscene, Threat, etc.
    - Attribute features: Race, Religion, Sexuality, Gender, Disability
  - Annotations are for Bias Evaluation
- Basic Text Processing
  - Lowercase, remove punctuation
  - Comment text represented as numeric vector where each element is mapped to each word of the comment

Implementation

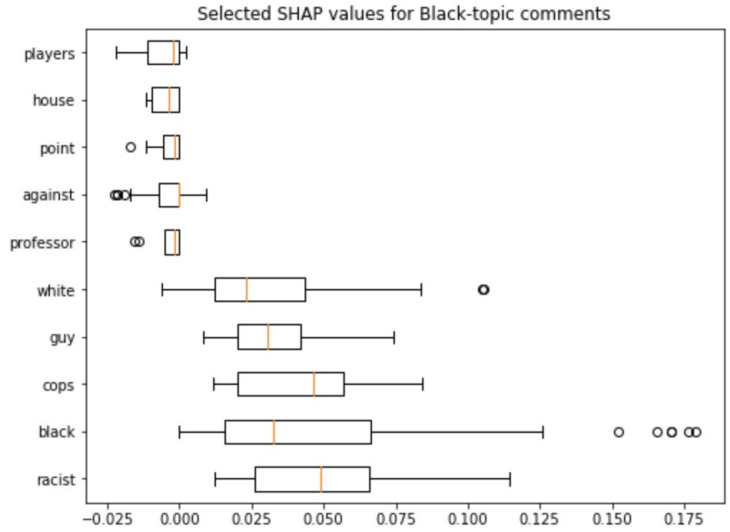


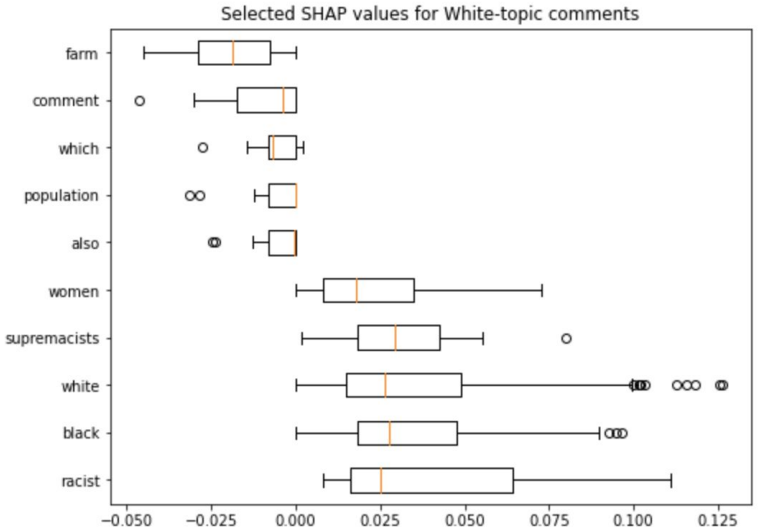
Evaluation (FPR/FNR)



- Low accuracy and higher FPR for Black- and White- related comments
- High FNR for all subgroups–likely due to low target base rate (8%)

Evaluation (SHAP)



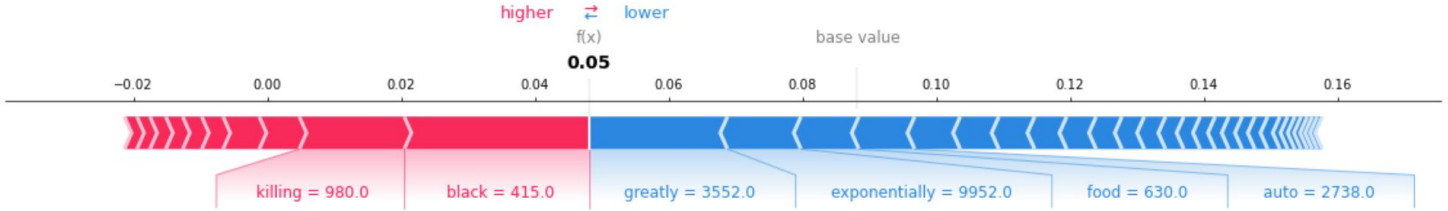


- Identity-related terms like “black,” “white,” and “women” have some of the highest SHAP values among frequently used terms.
  - Positive value = more “toxic”
- The ADS appears to disproportionately associate identity-referencing terms with toxicity

Evaluation (SHAP)

Comment: “Poverty increased, food stamp increases, debts and deficits increased greatly, black killing blacks increased exponentially in certain areas, auto debt record highs, student debt record highs, Middle East conflict never stopped. The saddest part is he got a peace prize for jack sh&&. That's obama basically.”

True label: Toxic **Predicted label**: Non-toxic



Comment: “It's pretty black and white, convicted and known terrorist who killed people gets paid $10.5M for it.”

True label: Non-Toxic **Predicted label**: Toxic

Summary

- Data Suitability
  - Annotator bias
  - Lack of representation of identity features
- Fairness of Implementation
  - Toxicity associated with racial identity features
- Potential Deployment would be *irresponsible*
- Next Steps
  - Consider an alternative dataset that is more transparent
  - Re-annotate dataset - automated and manual methods
  - Advanced text pre-processing
  - Group fairness metrics - FPR, FNR, SHAP