Identity Mention Biases in Toxic Comment Classification

Katherine Hann Maddie Kalil Sarah Payne
{khann22, mkalil, paynesa}@seas.upenn.edu

Toxic Comment Classification

- Area of growing research in recent years
- Used by many social media platforms
- Cyberbullying prevention and detection
- Binary or multi-classification

Accepted Comments Word Cloud

Rejected Comments Word Cloud

The Problem

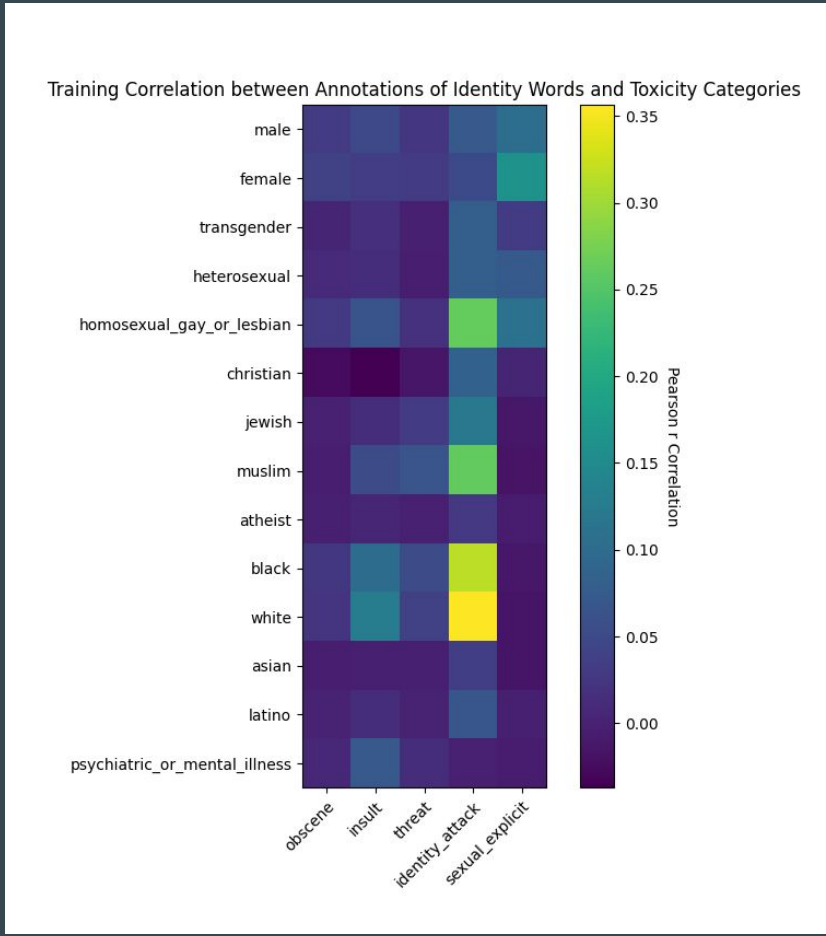
- Toxic comment classification, just like many other forms of NLP, can be biased
- Annotation biases
  - Sap et al. 2019 find that annotators are more likely to annotate African American Vernacular English (AAVE) Tweets as toxic
- Author dialect biases
  - Sap et al. 2019 demonstrated that Tweets occurring in AAVE are far more likely to be classified as toxic than those in other dialects
- **Identity mention biases**
  - Borkan et al. 2019 showed that comments with mentions of certain identities are more likely to be classified as toxic

Our Data

- We use Wikipedia comments for training and evaluation
  - Jigsaw task on Kaggle: unintended bias in toxic comment classification
  - Toxicity classes: toxicity, severe_toxicity, obscene, threat, insult, identity_attack, sexual_explicit
    - We do not consider severe_toxicity due to a lack of positive examples
  - Categories of identities: gender, sexual orientation, religious, racial, and disability
    - We do not consider any identity with less than 100 positive examples

Training Data Correlation

- Following Sap et al. 2019, we calculate the correlations between a comment being annotated as a certain identity and a toxicity score
- Mainly minority identities have high correlations, but so does white and identity_attack



Training Data PPMI

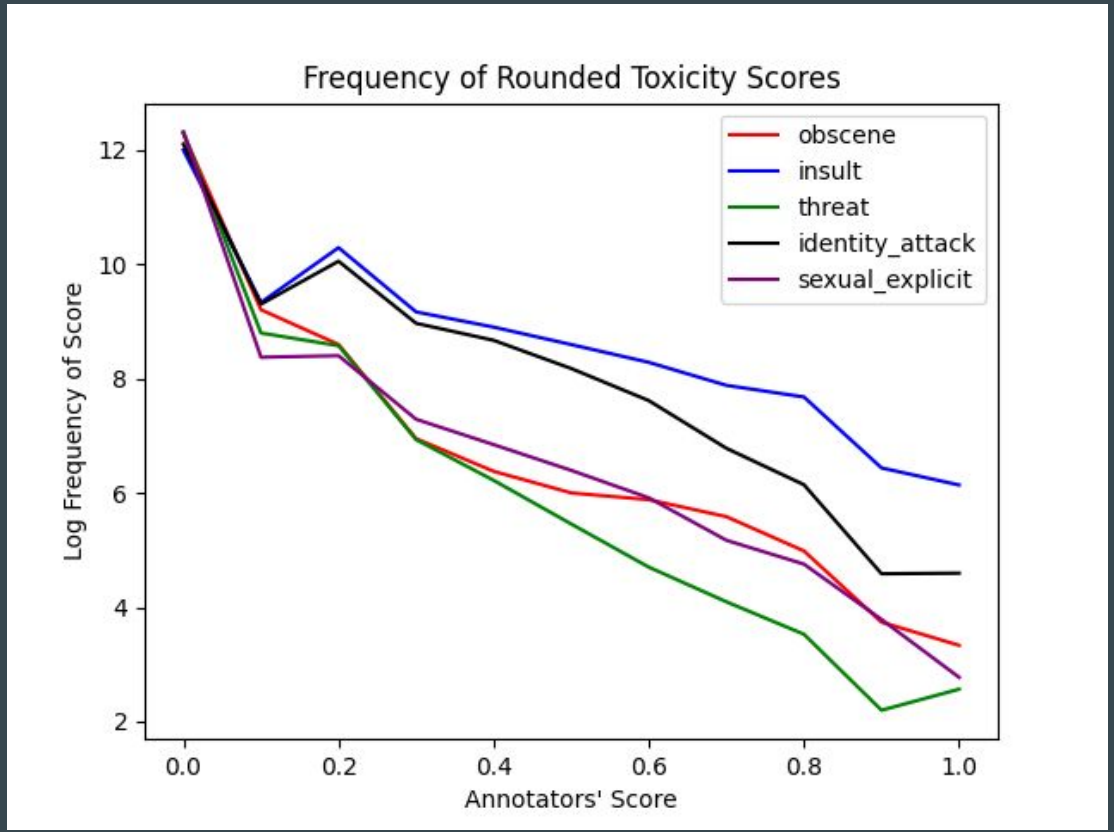
- High PPMI for muslim identity mention with toxicity label, and female identity with sexual_explicit label

Positive PMI of Training Data

	toxicity	obscene	insult	threat	identity_attack	sexual_explicit
male	0.069	0.388	0.22	0.234	0	0.618
female	0.004	0.583	0.252	0.177	0	0.941
transgender	0	0	0	0	0.183	0
other_gender	0	0	0.798	0	0	0
heterosexual	0	0	0	0	0.263	0.886
homosexual_gay_or_lesbian	0	0	0	0	0.309	0.229
bisexual	0	0	0	0	0.138	0.578
other_sexual_orientation	0	0	0	0	0.278	0
christian	0.689	0	0.05	0	0	0
jewish	0	0	0	0.545	0.237	0
muslim	1.023	0	0	0.326	0.3	0
hindu	0	0	0.025	0.104	0.08	0
buddhist	0	0	0	0	0.269	0
atheist	0	0.263	0.142	0	0	0
other_religion	0	0	0	0	0.26	0
black	0	0	0	0	0.308	0
white	0	0	0	0	0.187	0
asian	0	0.358	0.119	0.005	0	0
latino	0	0	0	0	0.25	0
other_race_or_ethnicity	0	0	0	0	0.364	0
physical_disability	0	0	0.511	0	0	0
intellectual_or_learning_disability	0	0	0.798	0	0	0
psychiatric_or_mental_illness	0	0	0.649	0	0	0
other_disability	0	0	0.798	0	0	0

Training data Binarization

- Many evaluation metrics require binary labels
- Data authors recommend 0.5 as threshold
- 0.1 looks possible but does not work as well



Our Model

- We evaluate Unitary AI's BERT-utilizing model
  - RoBERTa-based
  - Fine tuned on Kaggle training data
  - Multi-headed self-attention with 12 heads

Overall Prediction AUC

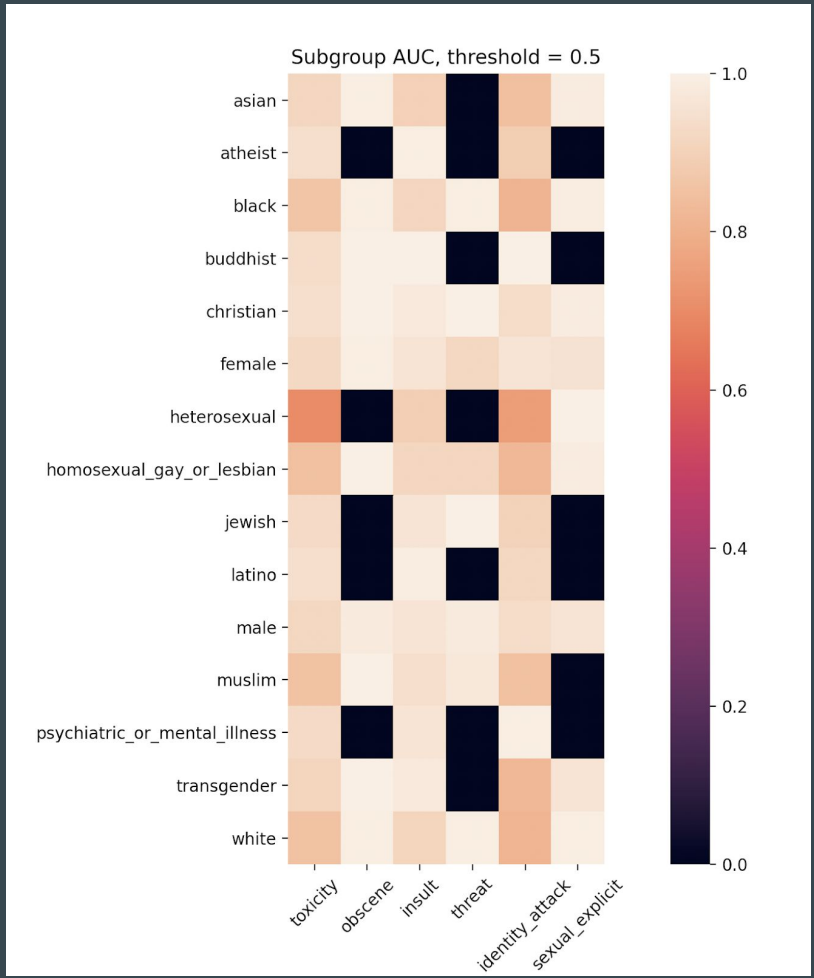
Toxicity Type	AUC
Toxicity	0.94900
Obscene	0.99474
Threat	0.97261
Insult	0.97092
Identity_attack	0.96248
Sexual_explicit	0.98904

Evaluating Bias in the Model

- Five measures for understanding identity-mention bias in the model, as used by Borkan et al:
  - 3 types of AUC
  - 2 types of AEG
- Values that cannot be calculated will be blacked out in heat maps
- Lexical Swapping with WinoBias
- Data Fuzzing

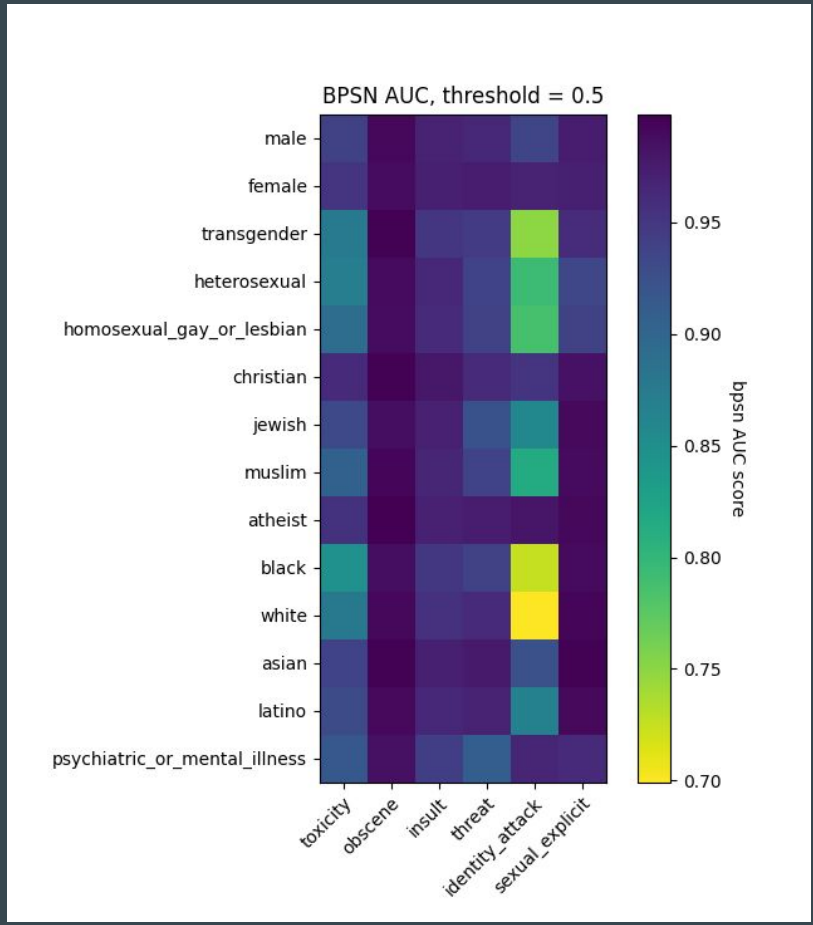
Subgroup AUC

- Calculating AUC on comments from a single identity
- Scores were near 1 for all groups for which a score could be computed
- Subgroup AUC by itself may not effectively identity unintended bias in a model



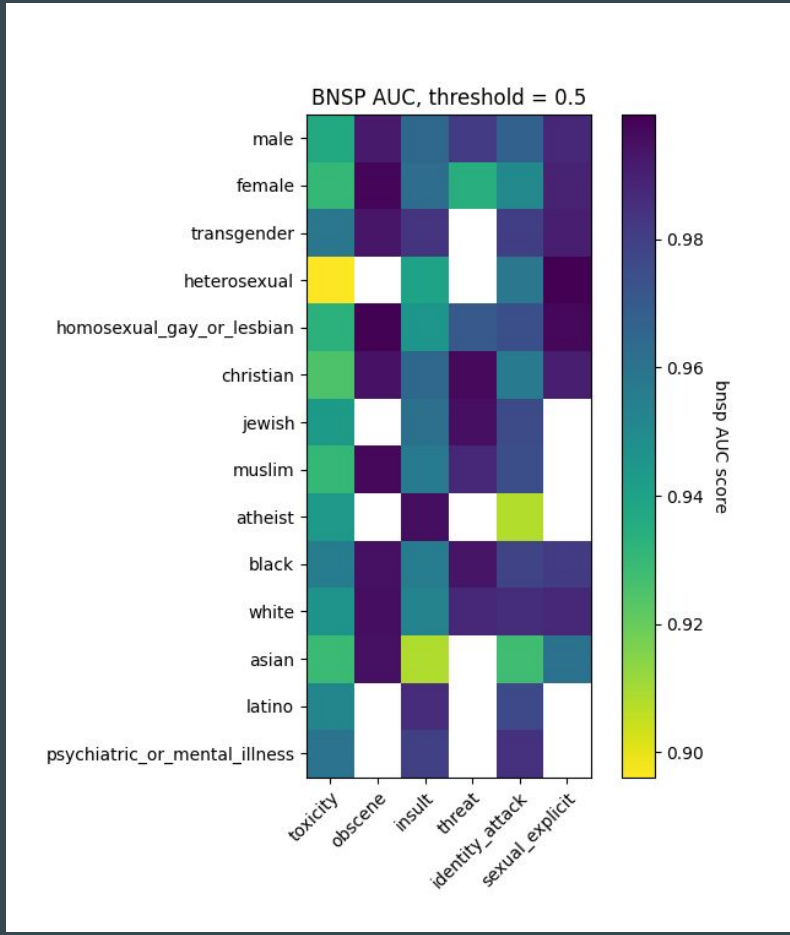
Background Positive Subgroup Negative AUC

- Sensitive to false positives
- Low BPSN AUC for marginalized identities with “identity_attack”
  - Low for “white” -- often these comments contain white supremacist or “reverse racism” language
  - Low for “heterosexual” -- often these comments contain homophobic langauge



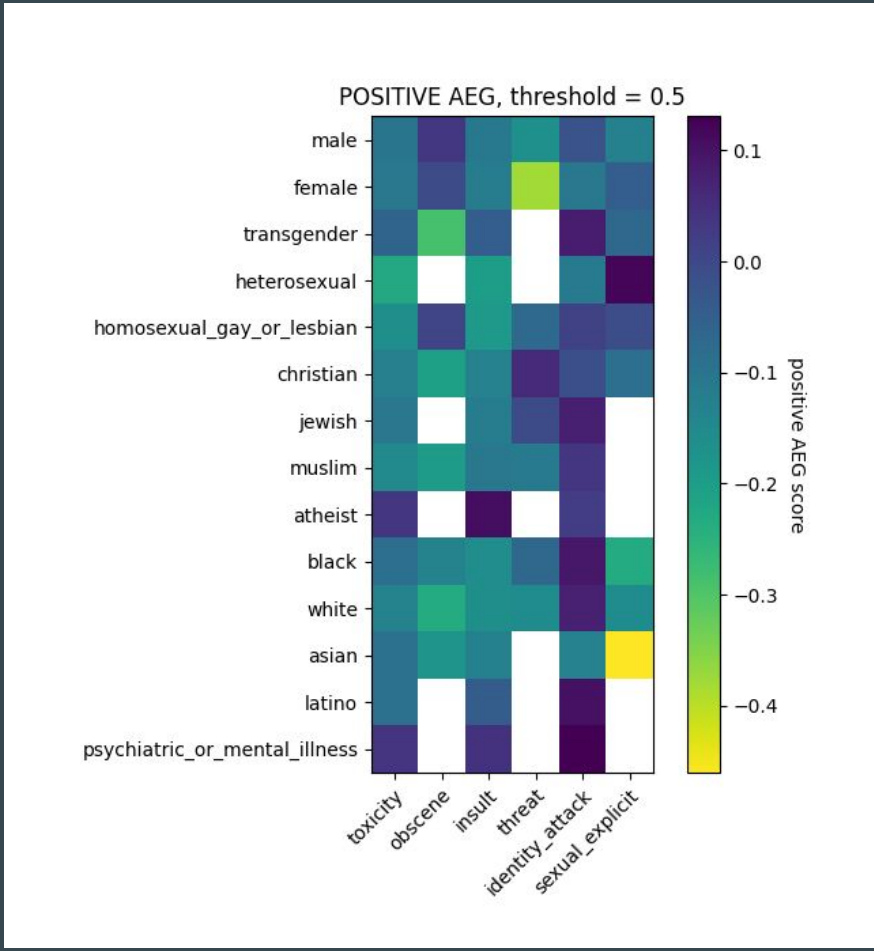
Background Negative Subgroup Positive AUC

- Sensitive to false negatives, which we see across a variety of identities
- Areas with higher AUC (fewer false negatives) are mainly marginalized identities
- Lower AUC for Asian and heterosexual
  - Asian-- very scattered comment topics
  - Heterosexual -- “identity_attack” tends to be homophobic but elsewhere more scattered and focused on the Church



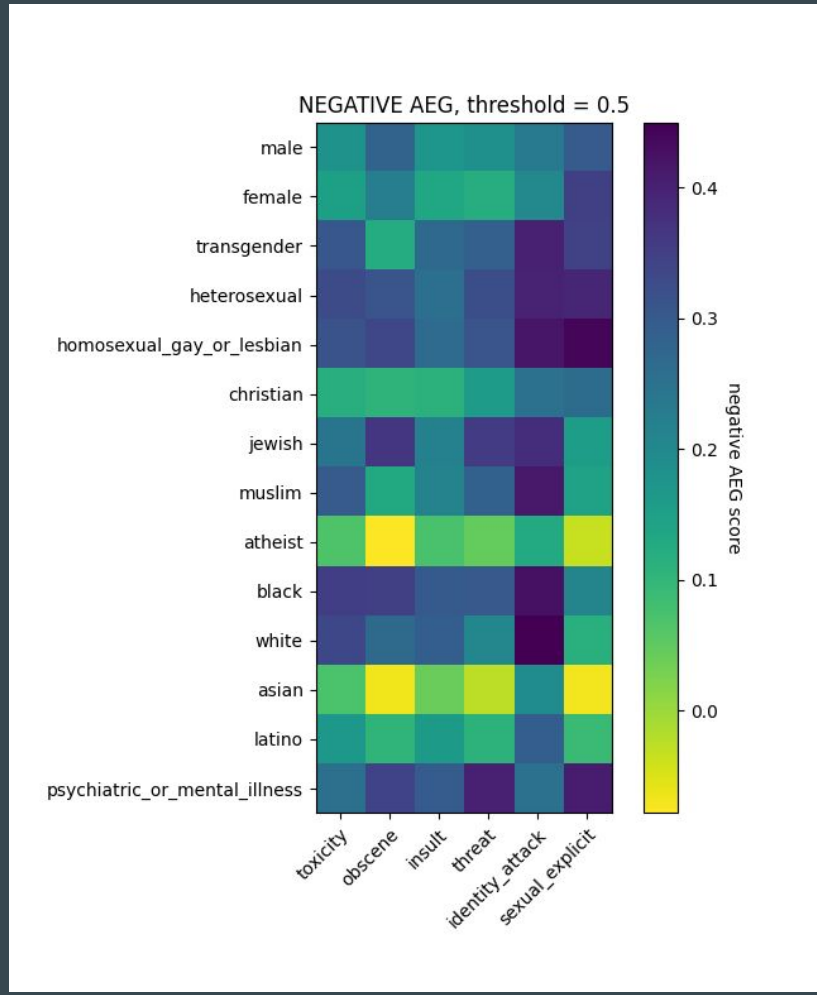
Positive AEG

- Compare positive examples from the subgroup with positive examples from the background
  - Goal of low separability
- ~75% of results are < |0.15|, meaning that most subgroups' TPRs have little bias when compared with their background
- Values < -0.3, reporting higher TPR in the subgroup than background
  - Ex: asian/sexual_explicit
    - Maybe not a true indicator of model bias
  - Ex: transgender/obscene, female/threat
    - Possibly due to stereotyping by annotators -- not reflected in correlations



Negative AEG

- Compare negative examples from the subgroup with negative examples from the background
- Higher values overall
- Low BPSN AUC correlated with high negative AEG
  - Indicates an upward shift (towards toxicity) of model scores for this non-toxic item
  - Ex: white/identity_attack
- High negative AEG without BPSN agreement
  - Still an upward shift in the model's scores for relatively non-toxic items, but not enough to cause mis-orderings with toxic items
  - Ex: homosexual_gay_or_lesbian/sexual_explicit

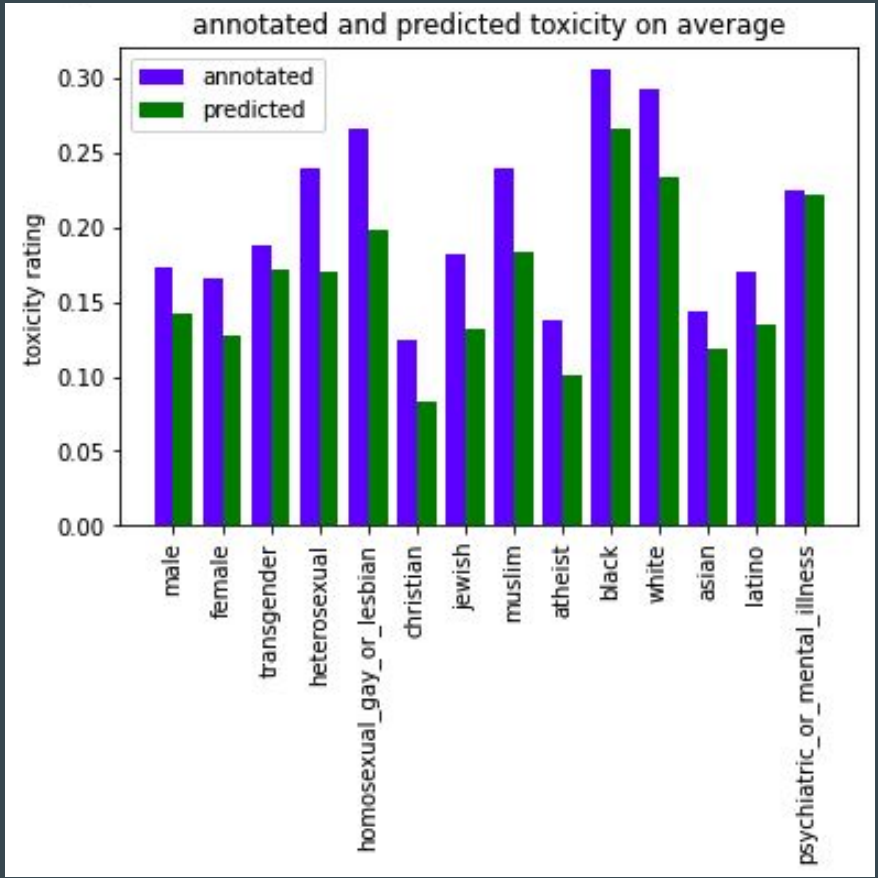


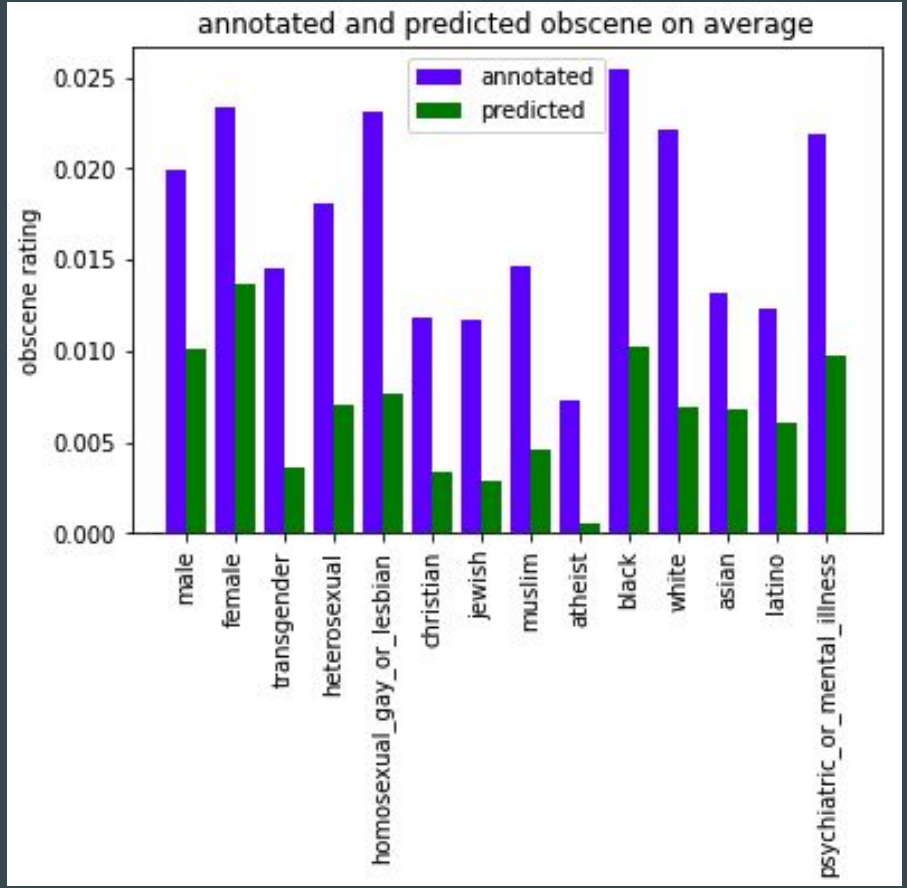
More Information on the Predictions

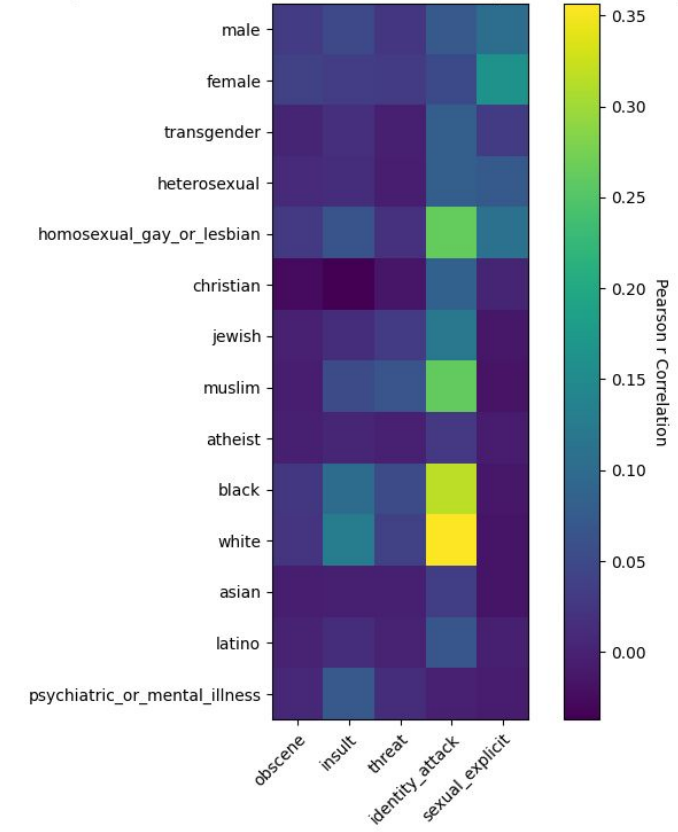
- Is the model developing bias for toxicity of identity groups that were not biased in the training data? And not for others that were biased in the training data?
  - Maybe, but the predictions just seem to be lower overall in toxicity than the annotations

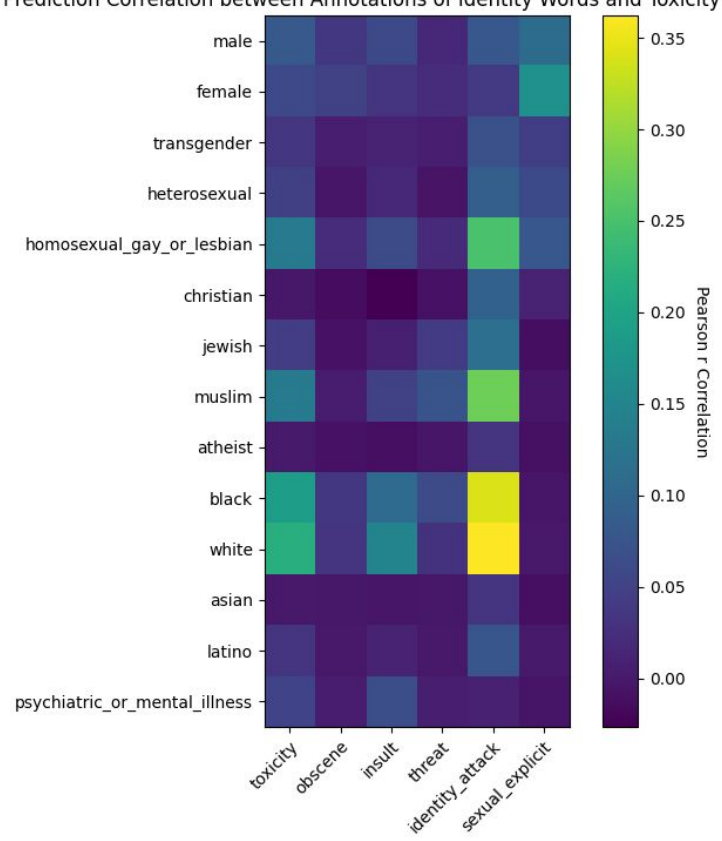
Mean Toxicity Annotated and Predicted Scores

	Toxicity	Severe_toxicity	Obscene	Insult	Threat
Annotated	0.13281	0.006337	0.014191	0.09316	0.01086
Predicted	0.11836	0.0000922	0.007366	0.07714	0.00339


annotated and predicted toxicity on average
annotated
predicted
toxicity rating
0.00
0.05
0.10
0.15
0.20
0.25
0.30
male
female
transgender
heterosexual
homosexual_gay_or_lesbian
christian
jewish
muslim
atheist
black
white
asian
latino
psychiatric_or_mental_illness

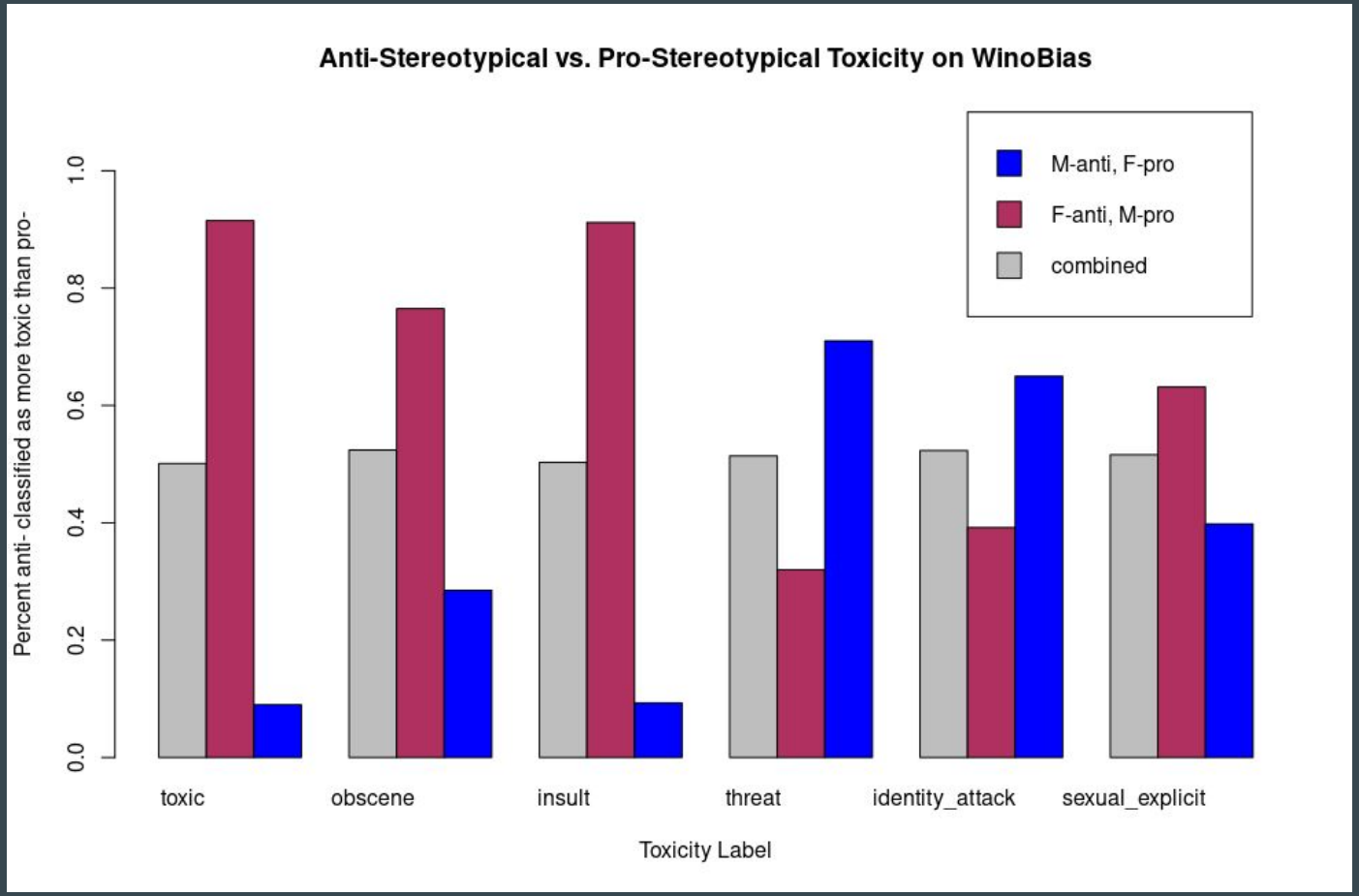

annotated and predicted obscene on average
annotated
predicted
obscene rating
0.000
0.005
0.010
0.015
0.020
0.025
male
female
transgender
heterosexual
homosexual_gay_or_lesbian
christian
jewish
muslim
atheist
black
white
asian
latino
psychiatric_or_mental_illness


Training Correlation between Annotations of Identity Words and Toxicity Categories
male
female
transgender
heterosexual
homosexual_gay_or_lesbian
christian
jewish
muslim
atheist
black
white
asian
latino
psychiatric_or_mental_illness
obscene
insult
threat
identity_attack
sexual_explicit
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Pearson r Correlation


Prediction Correlation between Annotations of Identity Words and Toxicity Categories
male
female
transgender
heterosexual
homosexual_gay_or_lesbian
christian
jewish
muslim
atheist
black
white
asian
latino
psychiatric_or_mental_illness
toxicity
obscene
insult
threat
identity_attack
sexual_explicit
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Pearson r Correlation

Stereotypes & Lexical Sensitivity

- What are the lexical cues to bias?
  - Case study: gender
- WinoBias: pro- and anti- stereotypical sentence pairs for each gender
- If model performance depends on pro- or anti-stereotypical, model sensitive to stereotypes
- If main differences based on pronouns only, model sensitive to these as lexical cues


Anti-Stereotypical vs. Pro-Stereotypical Toxicity on WinoBias
M-anti, F-pro
F-anti, M-pro
combined
Percent anti- classified as more toxic than pro-
0.0
0.2
0.4
0.6
0.8
1.0
toxic
obscene
insult
threat
identity_attack
sexual_explicit
Toxicity Label

Data Fuzzing

- Replace keywords that identify identities in the data set with key words from other identities in the data set randomly
- Gain a better sense of how much of the model's bias towards toxicity for certain comments is attributable to identity mentions in those comments
- Creating the fuzzed data set:
  - Word tokenization, POS-tagging
  - Filter out all POS but adjective, noun, and noun plural, and a few irrelevant words
  - Form list of most common words for each identity
  - Replace identifier words in comments with other identifier words whose POS matches

Original Comment	*Fuzzed Comment*
Old white men ARE the swamp.	Old black women ARE the swamp.
say “Jews” not “Jewish”	say “muslims” not “muslim”
alt-white or all-white	alt-chinese or all-chinese

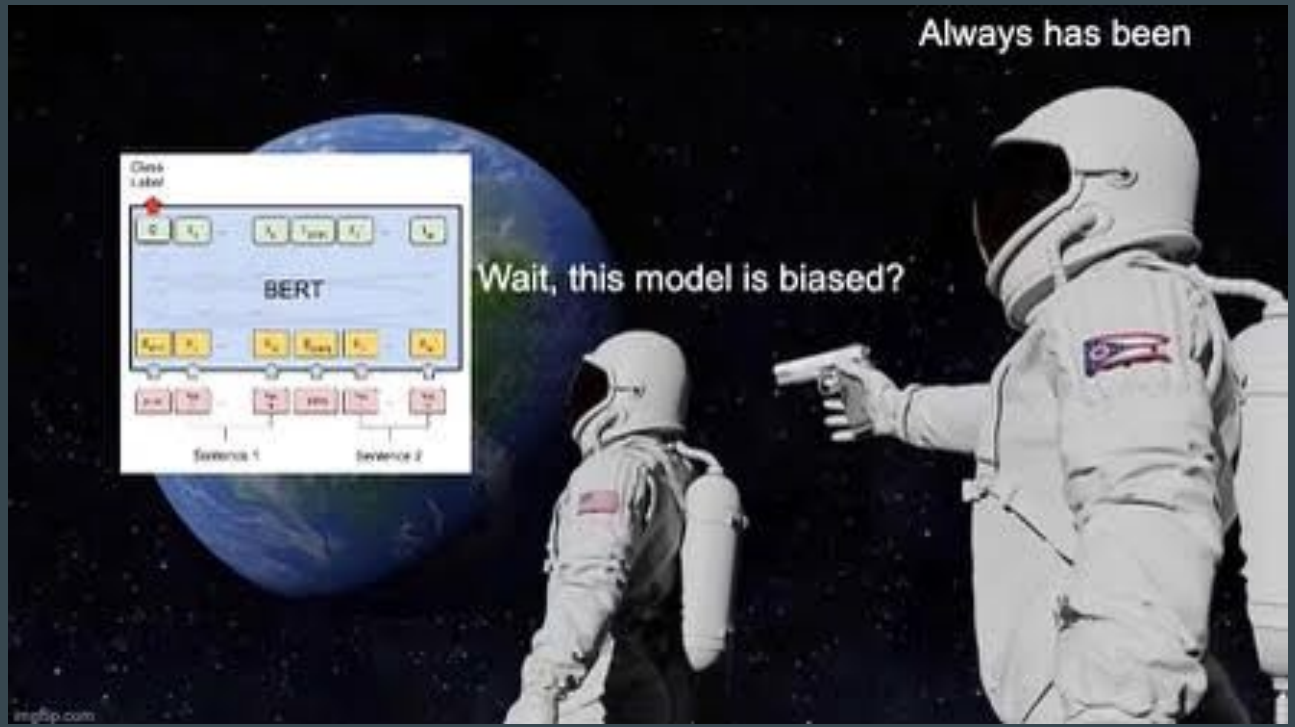
Data Fuzzing

Toxicity Type	AUC	Fuzzed AUC
Toxicity	0.94900	0.92918
Obscene	0.99474	0.99489
Threat	0.97261	0.96791
Insult	0.97092	0.96712
Identity_attack	0.96248	0.89466
Sexual_explicit	0.98904	0.98717

Conclusion

- We explored 7 different methods for identifying unintended bias in toxic comment classification methods
  - Borken et. al metrics (Subgroup AUC, BNSP/BPSN AUC, AEG) provide information about model performance on comments with certain identity mentions relative to the background
  - Data fuzzing techniques and stereotype bias prove information about lexical cues that may contribute to unintended model bias
  - We learned why non-marginalized identities may still have model bias (e.g. white supremacist comments)
  - We learned that lexical swapping provides additional bias information
- A comprehensive suite of metrics, along with an understanding of what information each metric encapsulates, will be useful for future research in mitigating bias in toxic comment classification and other NLP tasks

Thank You!!



Courtesy of Alex Yang

Appendix: Fuzzed Data Scores