Conversation—AI

Lucy Vasserman

Contents

We're a unit within Alphabet that builds technology to make the world safer. Our team tackles a range of global security issues including defending against digital attacks, mitigating the rise of online hate and harassment, countering online extremism, and fighting censorship.

Conversation AI Effort

Mission
Protect voices in conversation

Our work
API, tools, and research

Perspective

Problem

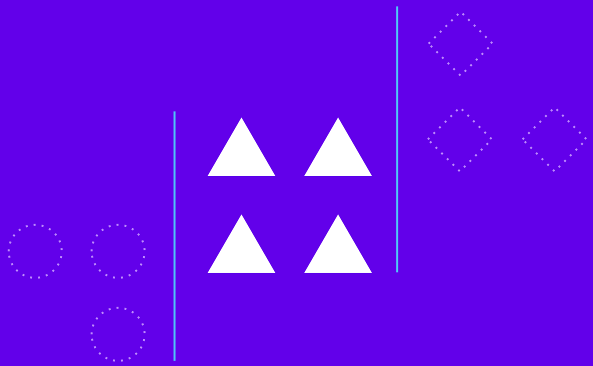
Abuse and toxicity have led people to give up on conversations.

Problem

Voices are silenced

People stop expressing themselves and the loudest voices shout other out of the room .

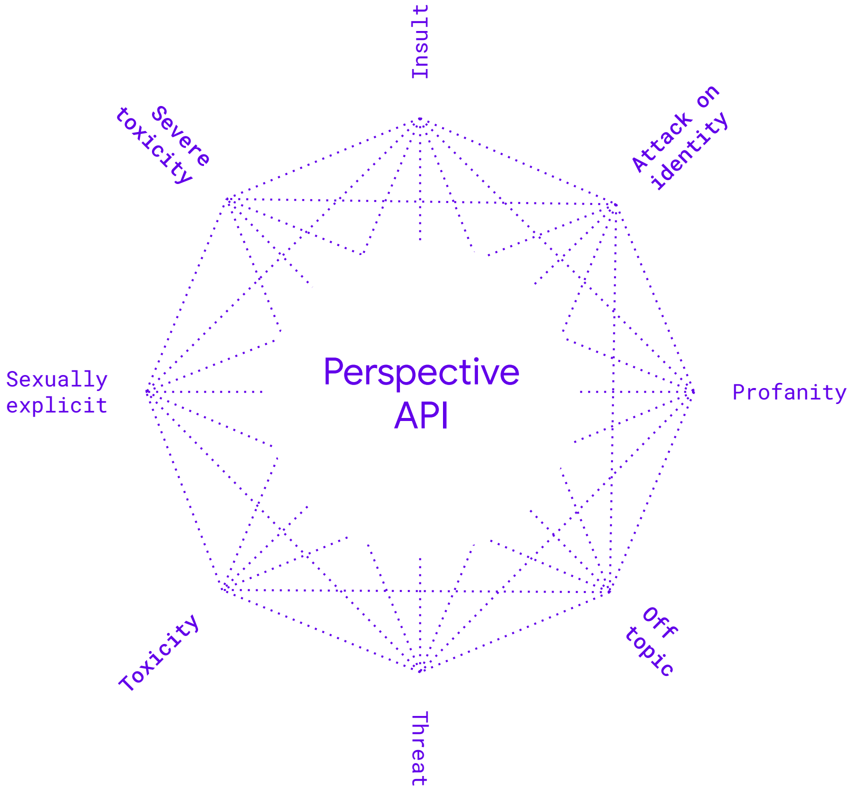
People are siloed



By optimizing for likes/shares platforms create filter bubbles so that people who disagree don't interact, or they shut down comments and discussion all together.

“Shut up. You’re an idiot!”

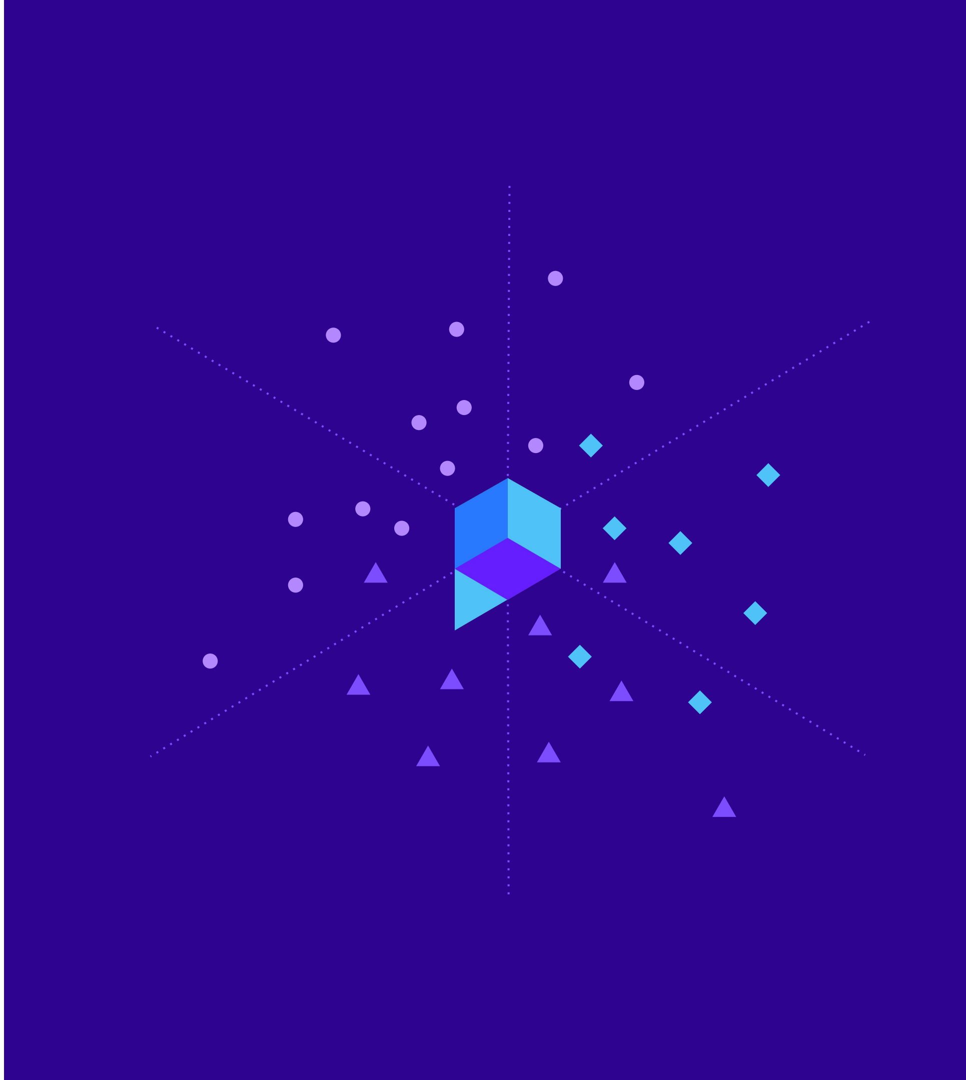
Toxicity	**0.9**
Severe Toxicity	0.1
Insult	0.4
Attack on Identity	0.2
Threat	0.1
Profanity	0.4
Sexually Explicit	0.1



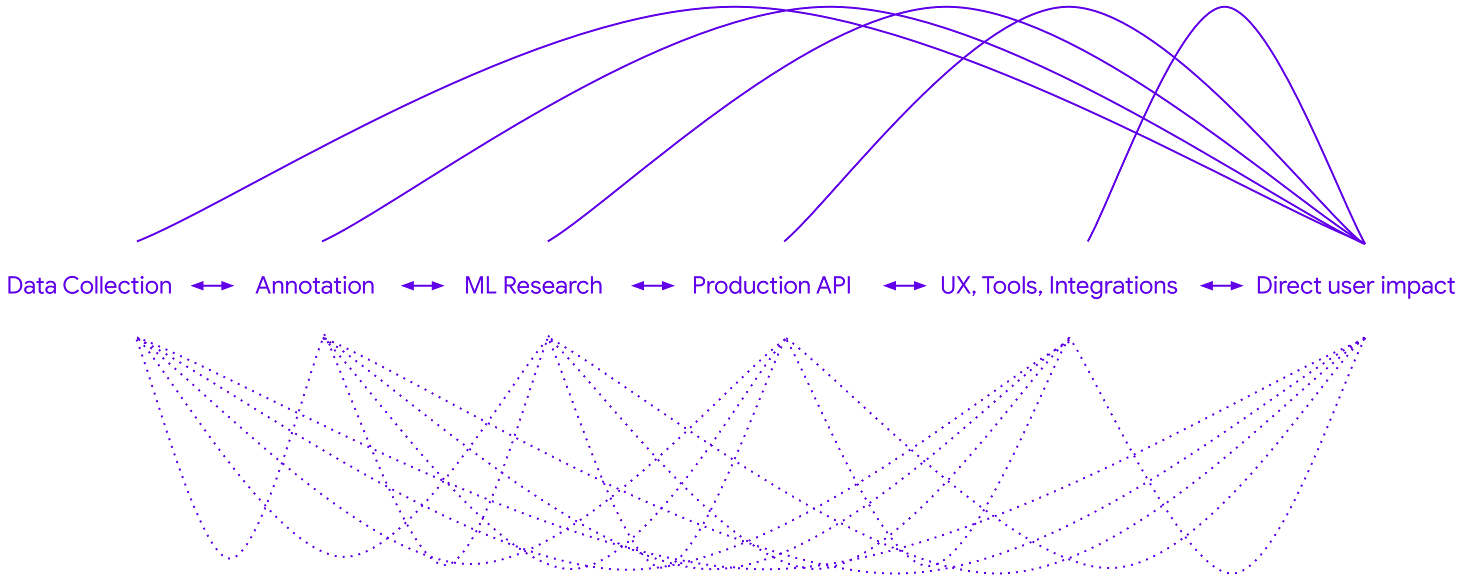
Perspective

Perspective aims to classify the emotional impact of language.

Is this a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion?



Outputs


Data Collection
Annotation
ML Research
Production API
UX, Tools, Integrations
Direct user impact

Success Metrics

Participation

Increase in the number of voices in a discussion

Diversity

Increase diversity of voices in a discussion

Toxicity

Reduce the prevalence of toxic comments online

Action

Increase in action against toxicity across ecosystem

Values

Community

Topic Neutrality

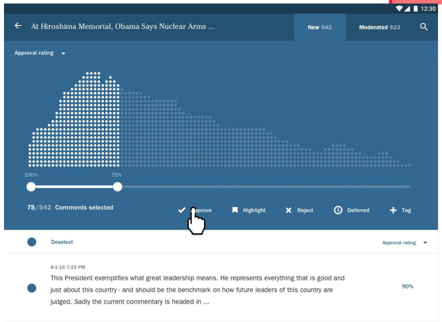
Transparency

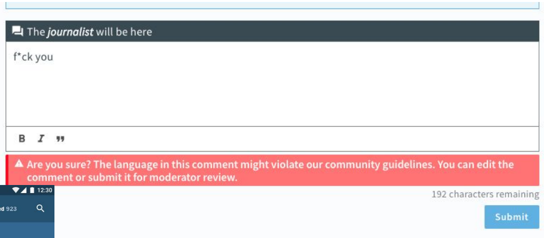
Privacy

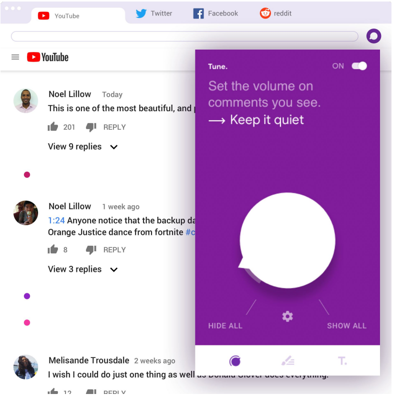
Inclusivity

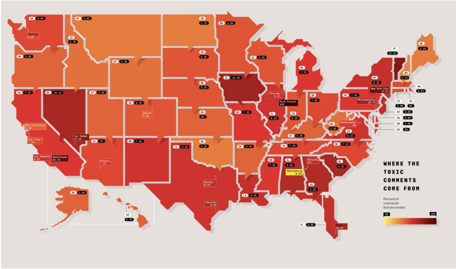
What we build

Experiences









Moderation

Help community managers set rules and review comments faster.

Authorship

Help people understand the impact of what they are writing.

Readership

Help people discover the conversations that interest them.

Visual trends

Help creators build data visualizations to better understand conversations at scale.

Transparency

Transparency

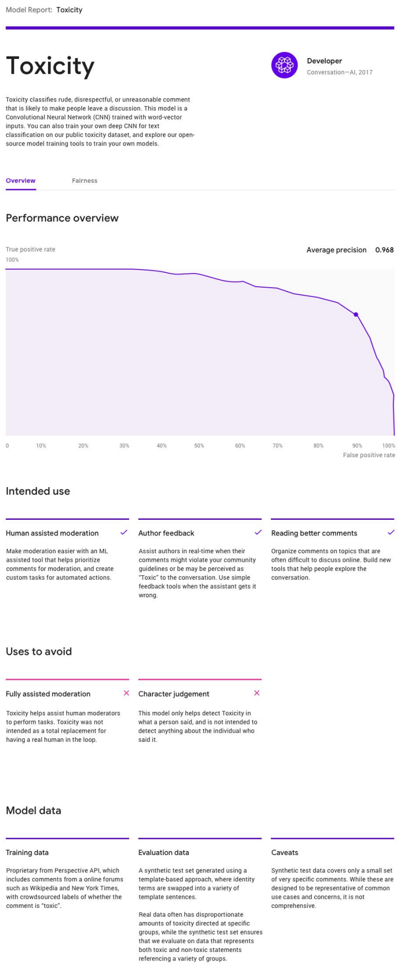
Public Demo

Having an easy to use public demo has enabled us to find and fix problems

Model Cards

A **Model Card** is a documentation framework that outlines:

- Evaluation results
- Intended usage
- Insight into training processes



Model Report: Toxicity

Toxicity

Developer
Conversation—AI, 2017

Toxicity classifies rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion. This model is a Convolutional Neural Network (CNN) trained with word-vector inputs. You can also train your own deep CNN for text classification on our public toxicity dataset, and explore our open-source model training tools to train your own models.

Overview | Fairness

Performance overview

True positive rate
100%
Average precision 0.968
0 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
False positive rate

Intended use

Human assisted moderation ✓
Make moderation easier with an ML assisted tool that helps prioritize comments for moderation, and create custom tasks for automated actions.

Author feedback ✓
Assist authors in real-time when their comments might violate your community guidelines or be may be perceived as "Toxic" to the conversation. Use simple feedback tools when the assistant gets it wrong.

Reading better comments ✓
Organize comments on topics that are often difficult to discuss online. Build new tools that help people explore the conversation.

Uses to avoid

Fully assisted moderation ✕
Toxicity helps assist human moderators to perform tasks. Toxicity was not intended as a total replacement for having a real human in the loop.

Character judgement ✕
This model only helps detect Toxicity in what a person said, and is not intended to detect anything about the individual who said it.

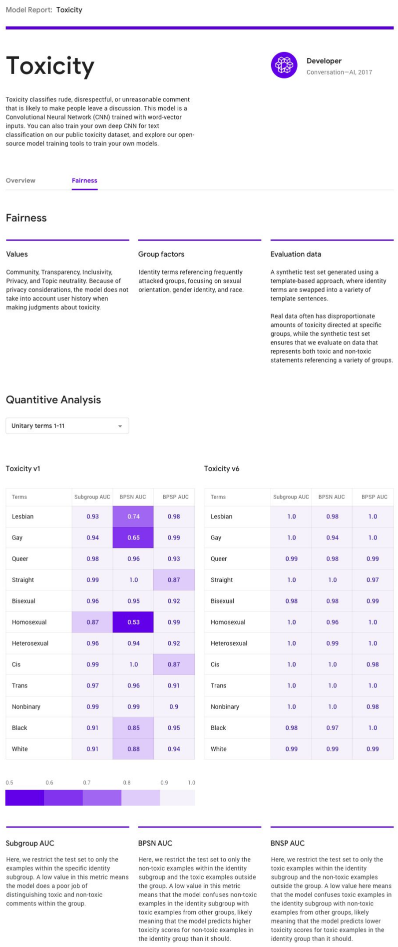
Model data

Training data
Proprietary from Perspective API, which includes comments from a online forums such as Wikipedia and New York Times, with crowdsourced labels of whether the comment is "toxic".

Evaluation data
A synthetic test set generated using a template-based approach, where identity terms are swapped into a variety of template sentences.

Real data often has disproportionate amounts of toxicity directed at specific groups, while the synthetic test set ensures that we evaluate on data that represents both toxic and non-toxic statements referencing a variety of groups.

Caveats
Synthetic test data covers only a small set of very specific comments. While these are designed to be representative of common use cases and concerns, it is not comprehensive.



Model Report: Toxicity

Toxicity

Developer
Conversation—AI, 2017

Toxicity classifies rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion. This model is a Convolutional Neural Network (CNN) trained with word-vector inputs. You can also train your own deep CNN for text classification on our public toxicity dataset, and explore our open-source model training tools to train your own models.

Overview | Fairness

Fairness

Values
Community, Transparency, Inclusivity, Privacy, and Topic neutrality. Because of privacy considerations, the model does not take into account user history when making judgments about toxicity.

Group factors
Identity terms referencing frequently attacked groups, focusing on sexual orientation, gender identity, and race.

Evaluation data
A synthetic test set generated using a template-based approach, where identity terms are swapped into a variety of template sentences.

Real data often has disproportionate amounts of toxicity directed at specific groups, while the synthetic test set ensures that we evaluate on data that represents both toxic and non-toxic statements referencing a variety of groups.

Quantitive Analysis

Unitary terms 1-11

Toxicity v1

Terms	Subgroup AUC	BPSN AUC	BPSP AUC
Lesbian	0.93	0.74	0.98
Gay	0.94	0.65	0.99
Queer	0.98	0.96	0.93
Straight	0.99	1.0	0.87
Bisexual	0.96	0.95	0.92
Homosexual	0.87	0.53	0.99
Heterosexual	0.96	0.94	0.92
Cis	0.99	1.0	0.87
Trans	0.97	0.96	0.91
Nonbinary	0.99	0.99	0.9
Black	0.91	0.85	0.95
White	0.91	0.88	0.94

Toxicity v6

Terms	Subgroup AUC	BPSN AUC	BPSP AUC
Lesbian	1.0	0.98	1.0
Gay	1.0	0.94	1.0
Queer	0.99	0.98	0.99
Straight	1.0	1.0	0.97
Bisexual	0.98	0.98	0.99
Homosexual	1.0	0.96	1.0
Heterosexual	1.0	0.99	1.0
Cis	1.0	1.0	0.98
Trans	1.0	1.0	1.0
Nonbinary	1.0	1.0	0.98
Black	0.98	0.97	1.0
White	0.99	0.99	0.99

0.5 0.6 0.7 0.8 0.9 1.0

Subgroup AUC
Here, we restrict the test set to only the examples within the specific identity subgroup. A low value in this metric means the model does a poor job of distinguishing toxic and non-toxic comments within the group.

BPSN AUC
Here, we restrict the test set to only the non-toxic examples within the identity subgroup and the toxic examples outside the group. A low value in this metric means that the model confuses non-toxic examples in the identity subgroup with toxic examples from other groups, likely meaning that the model predicts higher toxicity scores for non-toxic examples in the identity group than it should.

BNSP AUC
Here, we restrict the test set to only the toxic examples within the identity subgroup and the non-toxic examples outside the group. A low value here means that the model confuses toxic examples in the identity subgroup with non-toxic examples from other groups, likely meaning that the model predicts lower toxicity scores for toxic examples in the identity group than it should.

[The False Positive - Medium Blog](#)
[Conversation AI - Jigsaw](#)

Unintended Bias

False "toxic" positives

A naively trained model will have some strong unintended biases illustrated by these false-positive examples...

Comment	Toxicity score
The Gay and Lesbian Film Festival starts today.	0.82
Being transgender is independent of sexual orientation.	0.52
A Muslim is someone who follows or practices Islam.	0.46

Bias Mitigation

Bias caused by dataset imbalance

- Frequently attacked identities are overrepresented in toxic comments
- Length matters

Add *assumed non-toxic data* from Wikipedia articles to fix the imbalance.

- Original dataset had 127,820 examples
- 4,620 non-toxic examples added

	Comment Length				
Term	**20-59**	**60-179**	**180-539**	**540-1619**	**1620-4859**
ALL	17%	12%	7%	5%	5%
gay	88%	77%	51%	30%	19%
queer	75%	83%	45%	56%	0%
homosexual	78%	72%	43%	16%	15%
black	50%	30%	12%	8%	4%
white	20%	24%	16%	12%	2%
wikipedia	39%	20%	14%	11%	7%
atheist	0%	20%	9%	6%	0%
lesbian	33%	50%	42%	21%	0%
feminist	0%	20%	25%	0%	0%
islam	50%	43%	12%	12%	0%
muslim	0%	25%	21%	12%	17%
race	20%	25%	12%	10%	6%
news	0%	1%	4%	3%	3%
daughter	0%	7%	0%	7%	0%

How can we measure unintended bias?

Definitions

- **Unintended bias** exists if the model performance varies across different subgroups
- **Subgroups** are the identities mentioned in the text (not the identities of the author or recipient)

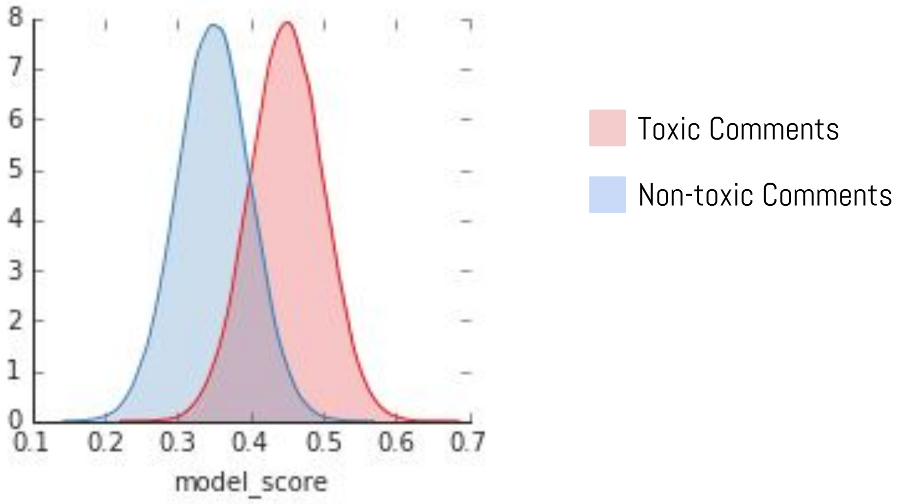
Metrics

- Metrics should be threshold independent

Measuring Overall Model Performance - AUC

How good is the model at distinguishing toxic from non-toxic examples? (ROC-AUC)

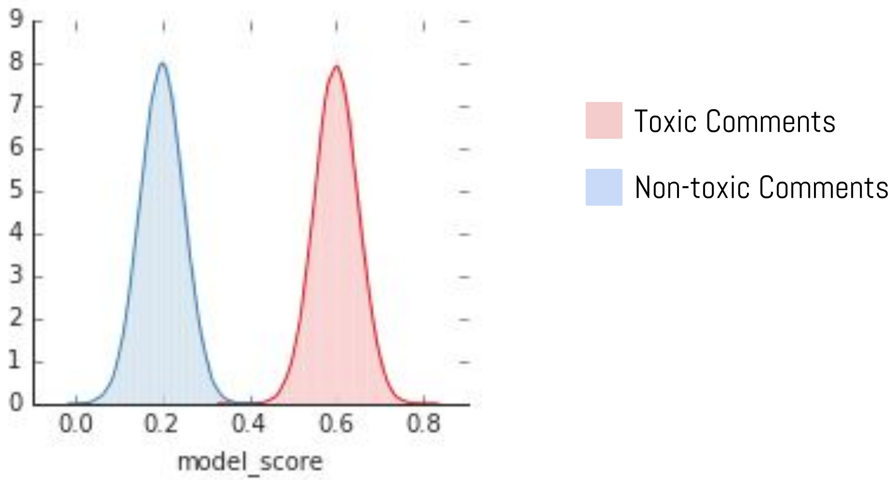
AUC (for a given test set) = Given two randomly chosen examples, one in-class (e.g. one is toxic and the other is not), AUC is the probability that the model will give the in-class example the higher score.



Measuring Overall Model Performance - AUC

How good is the model at distinguishing toxic from non-toxic examples? (ROC-AUC)

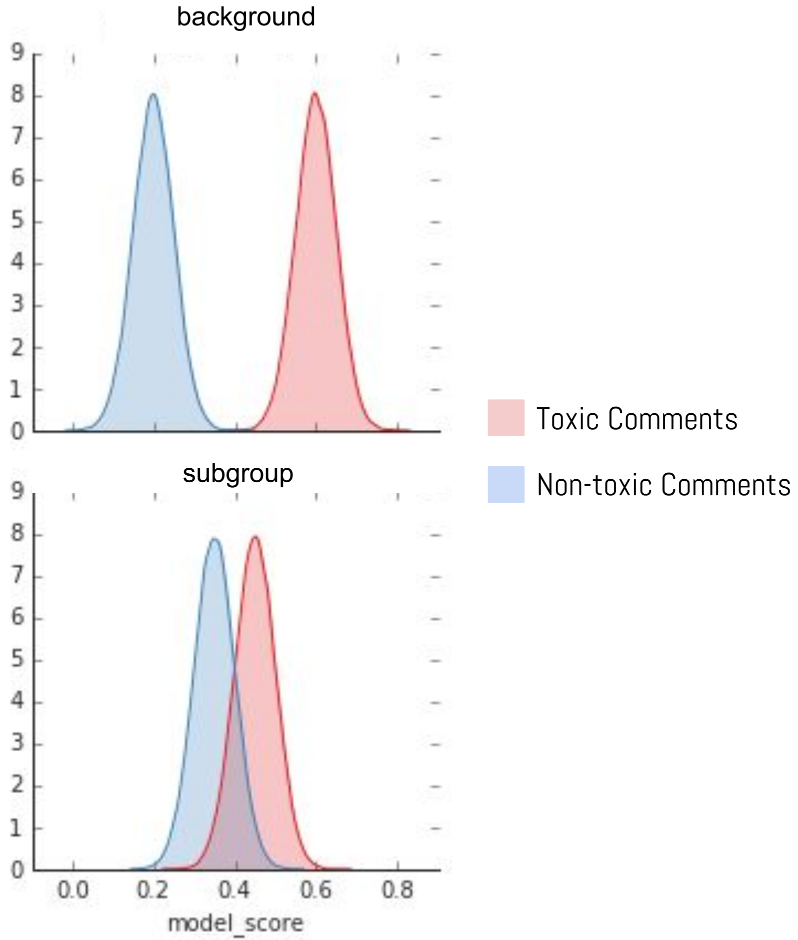
AUC (for a given test set) = Given two randomly chosen examples, one in-class (e.g. one is toxic and the other is not), AUC is the probability that the model will give the in-class example the higher score.



Subgroup AUC

Measures low subgroup performance.

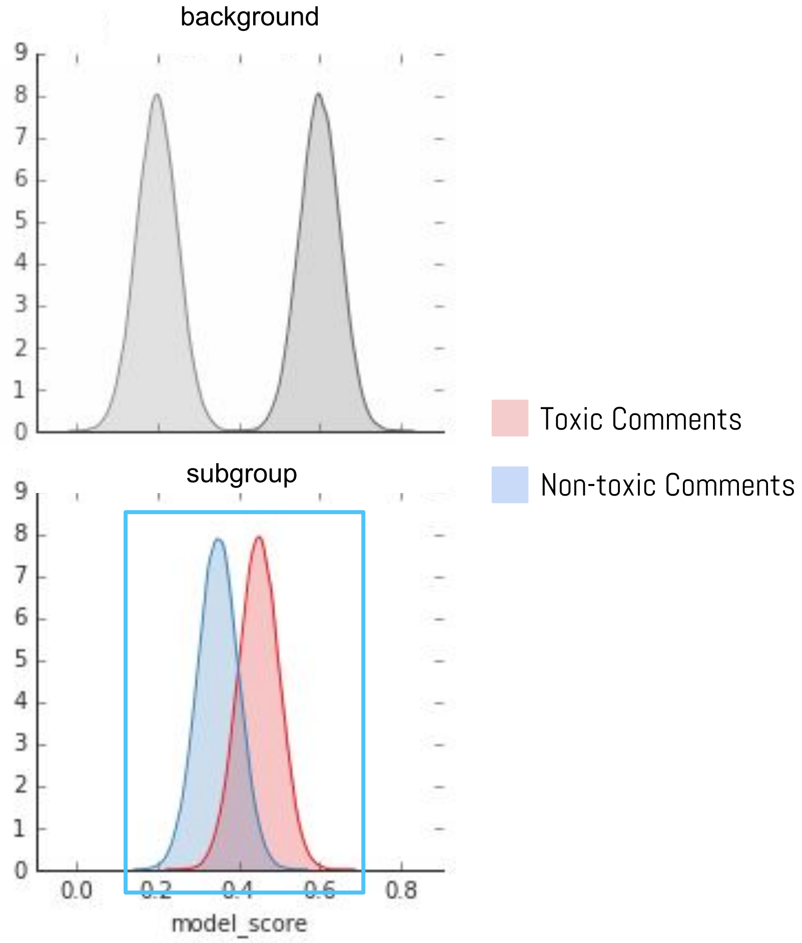
Detects if the model performs worse on subgroup comments than it does on comments overall.



Subgroup AUC

Measures low subgroup performance.

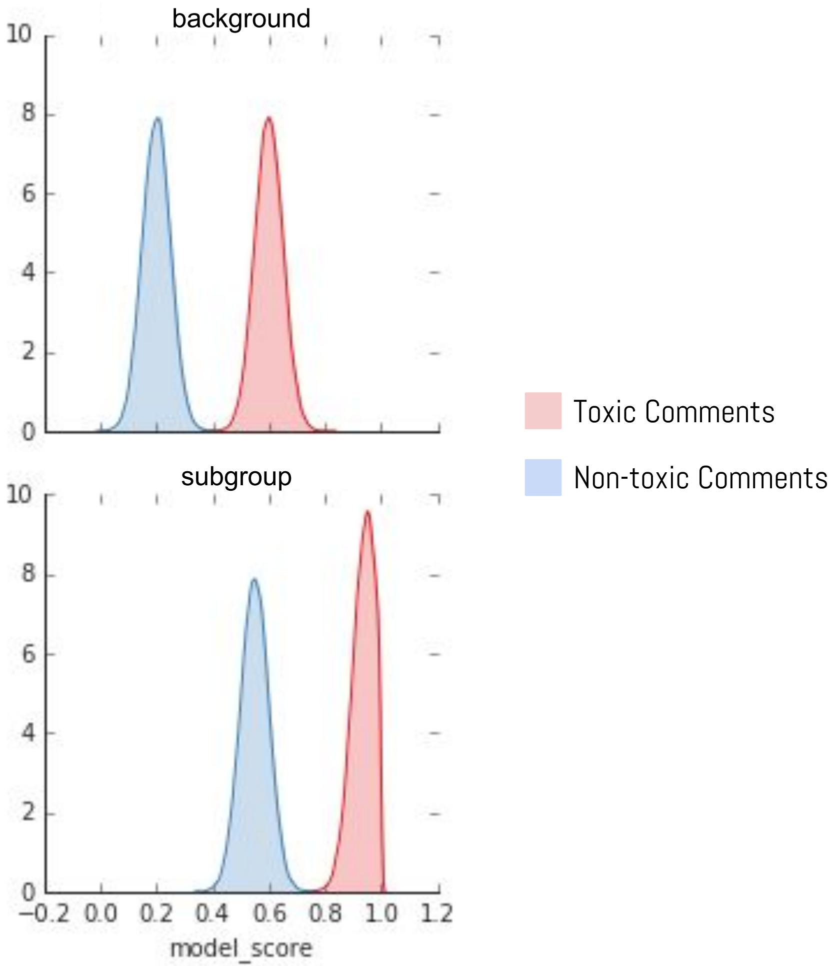
Detects if the model performs worse on subgroup comments than it does on comments overall.



Background Positive Subgroup Negative (BPSN) AUC

Measures subgroup shifts to the right

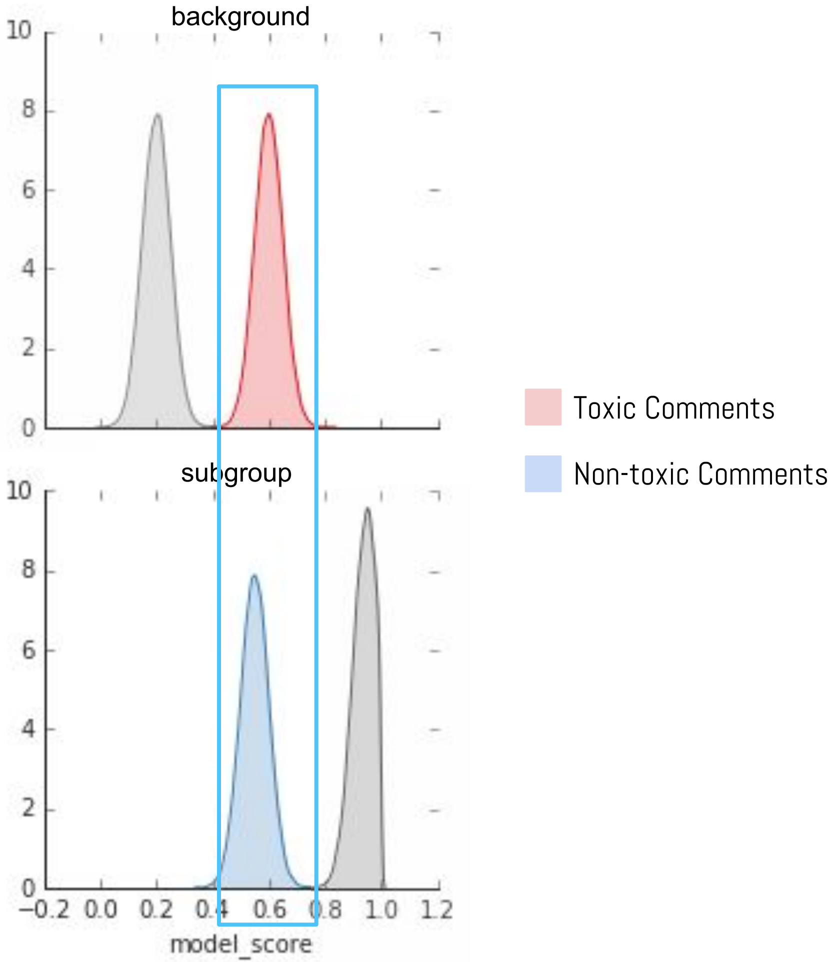
Detects if the model systematically scores comments from the subgroup higher.



Background Positive Subgroup Negative (BPSN) AUC

Measures subgroup shifts to the right

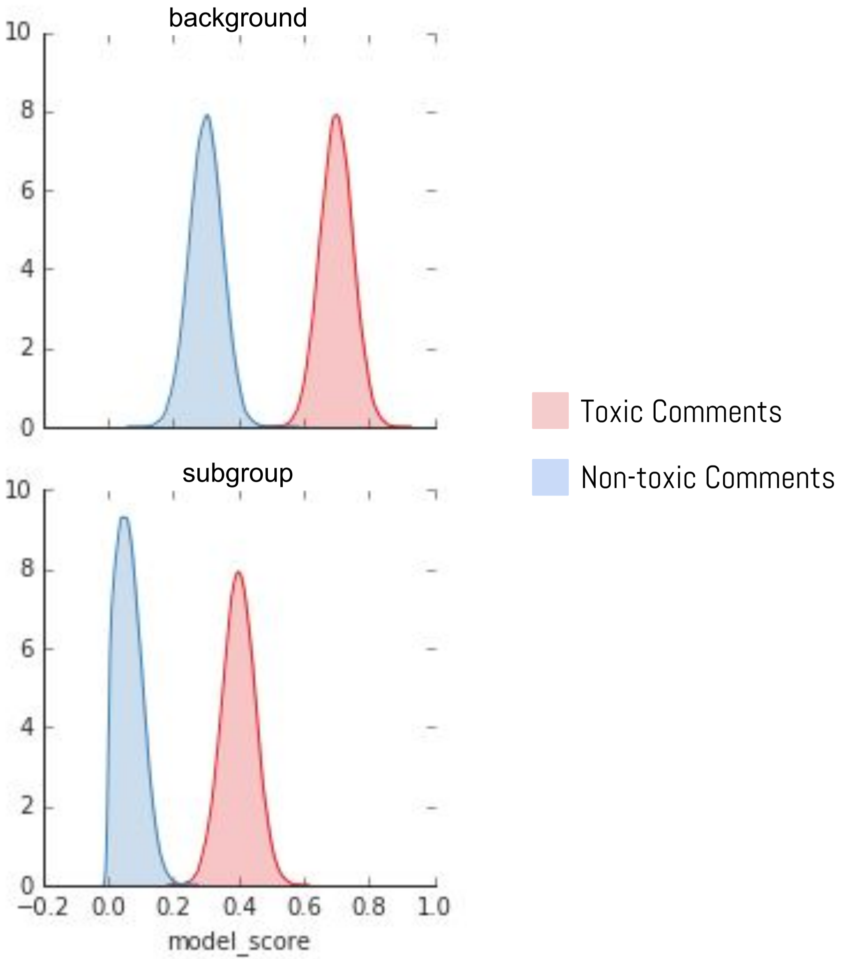
Detects if the model systematically scores comments from the subgroup higher.



Background Positive Subgroup Negative (BPSN) AUC

Measures subgroup shifts to the left.

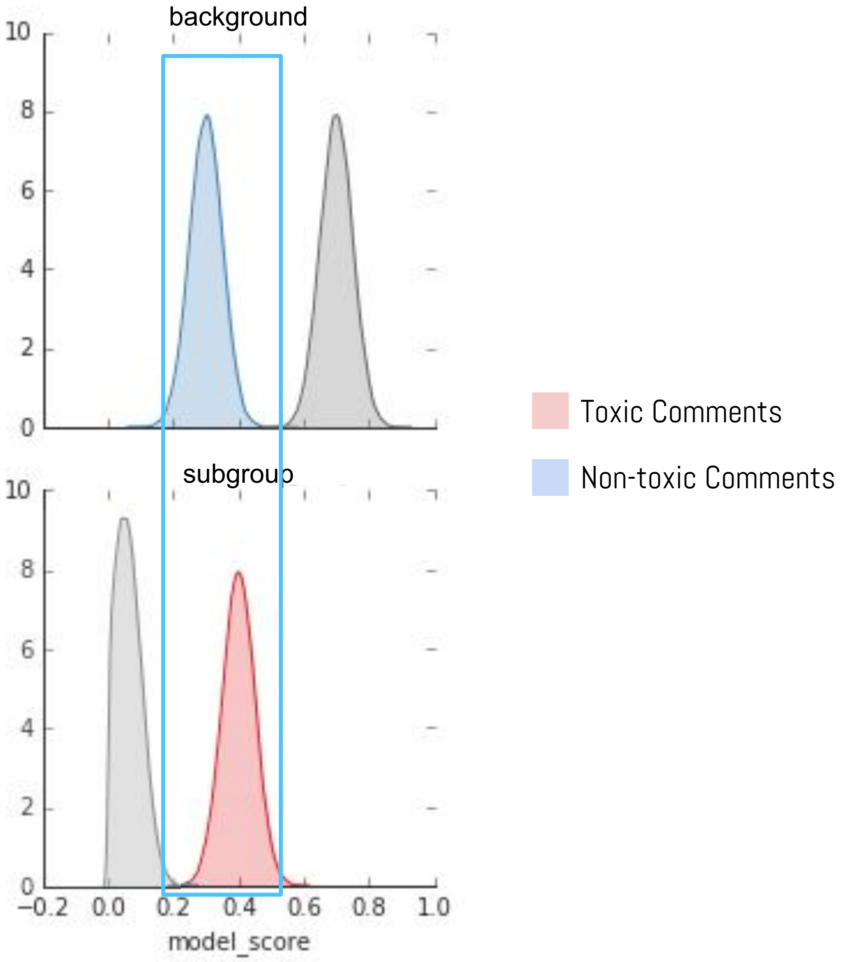
Detects if the model systematically scores comments from the subgroup lower.



Background Positive Subgroup Negative (BPSN) AUC

Measures subgroup shifts to the left.

Detects if the model systematically scores comments from the subgroup lower.



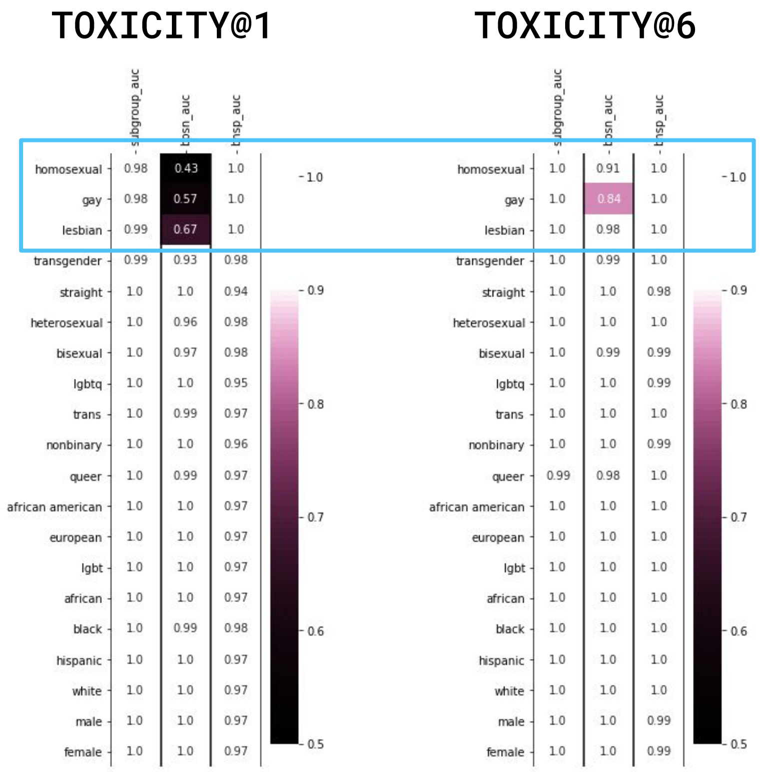
Evaluation on synthetic data

Synthetic data shows real improvement!

Comments are generated using simple templates

text: *"I am <identity>"*
label: *non-toxic*

text: *"I hate <identity>"*
label: *toxic*



Public dataset for bias research

~2 million comments released by Civil Comments platform

Annotated for toxicity (all)

Is this a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion?

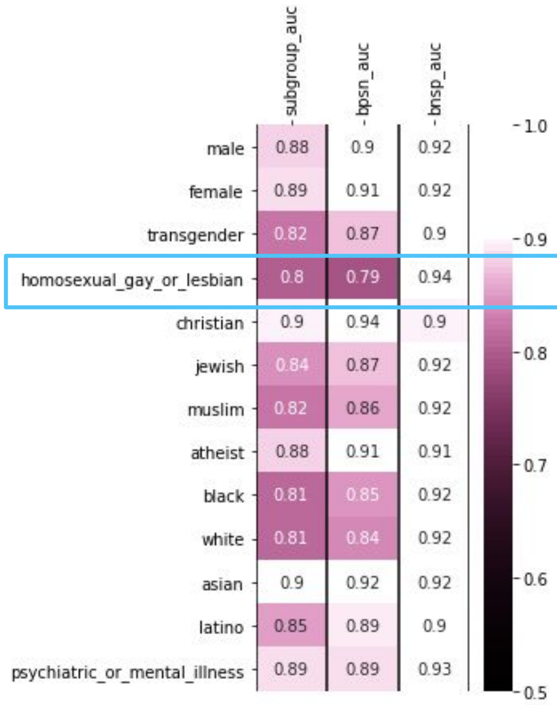
Annotated for identity content (~360k)

What genders are mentioned in this comment?
What races or ethnicities are mentioned in this comment?
etc...

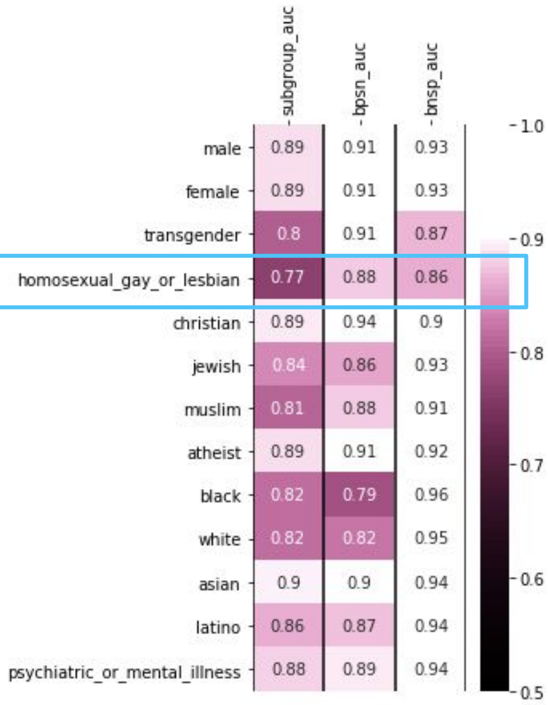
Evaluation on real data

Real data shows mixed results

TOXICITY@1



TOXICITY@6

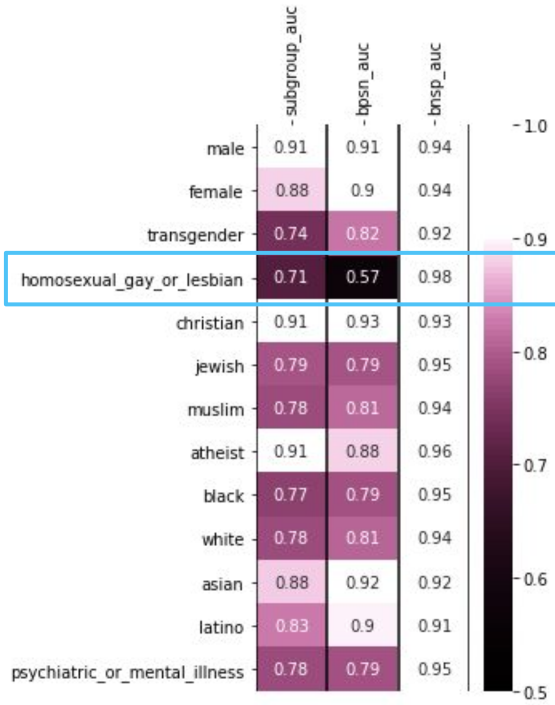


Evaluation on real data - short comments only

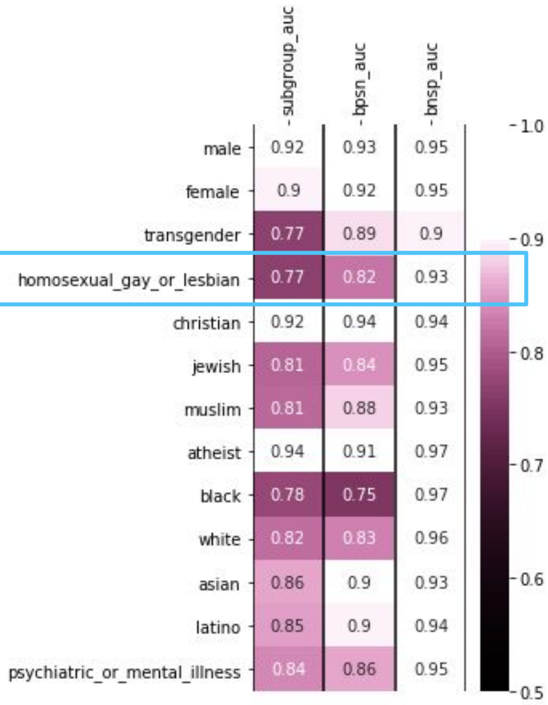
The unintended bias was worse for short comments.

Bias mitigation brought performance on short comments closer to overall performance, but bias still exists.

TOXICITY@1



TOXICITY@6



Kaggle Competition

Kaggle Competition

Data

2 million comments set from Civil Comments

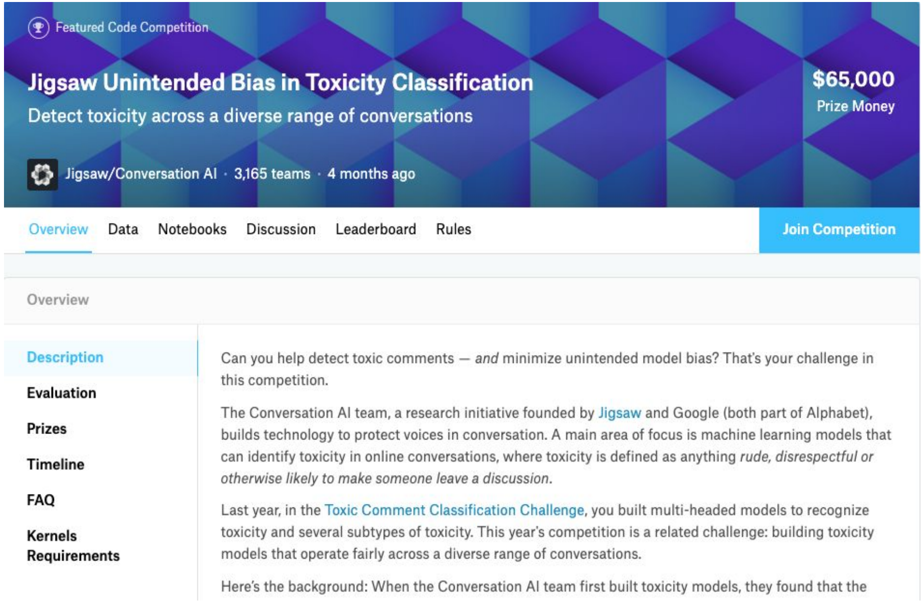
Evaluation

Generalized mean of three bias AUCs for all identities and overall AUC

Results

3k+ teams researching bias mitigation techniques

Winners used BERT models and identity-aware data weighting



Questions?