

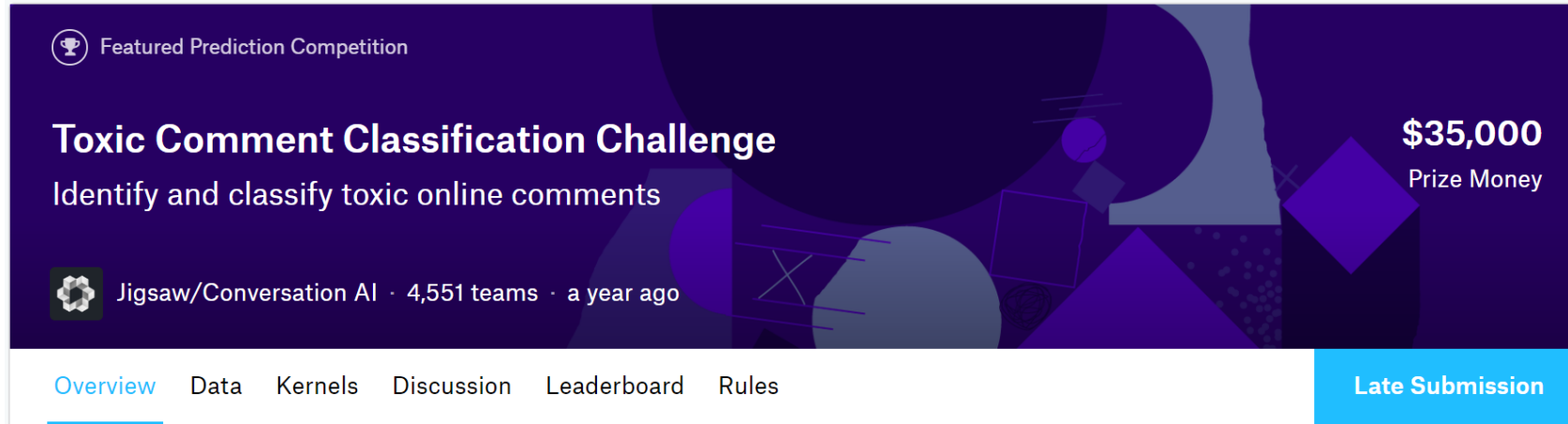
# Toxic Comment Classification

David Braslow

May 8, 2019

# Overview


2018



Featured Prediction Competition

## Toxic Comment Classification Challenge

Identify and classify toxic online comments

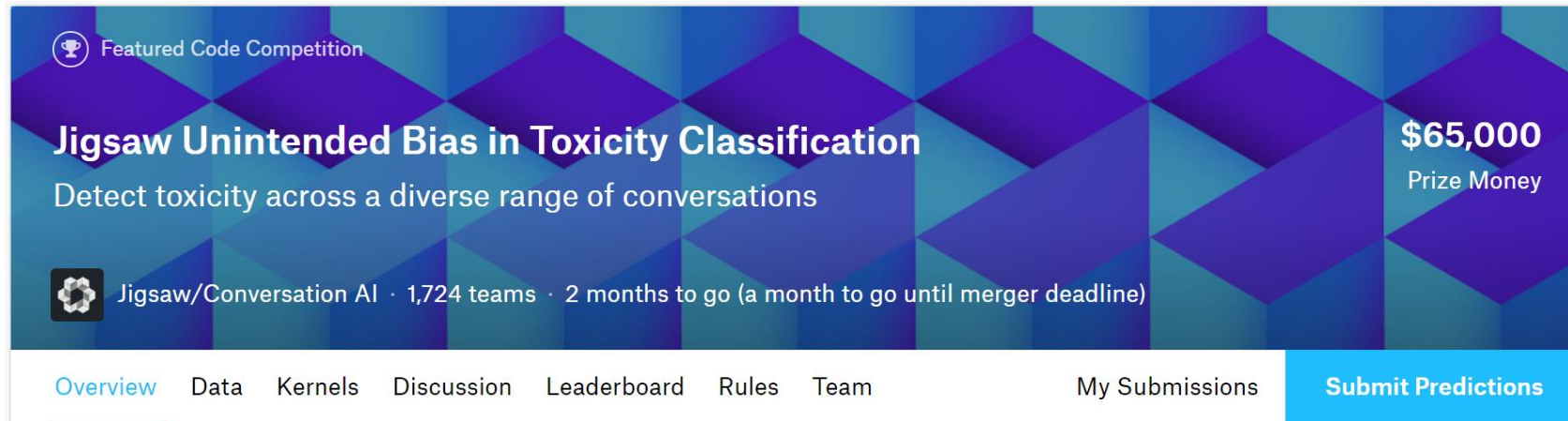
 Jigsaw/Conversation AI · 4,551 teams · a year ago

**\$35,000**  
Prize Money

[Overview](#) [Data](#) [Kernels](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Late Submission](#)

Detect toxic comments


2019



Featured Code Competition

## Jigsaw Unintended Bias in Toxicity Classification

Detect toxicity across a diverse range of conversations

 Jigsaw/Conversation AI · 1,724 teams · 2 months to go (a month to go until merger deadline)

**\$65,000**  
Prize Money

[Overview](#) [Data](#) [Kernels](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Team](#) [My Submissions](#) [Submit Predictions](#)

Detect toxic comments — *and* minimize unintended model bias

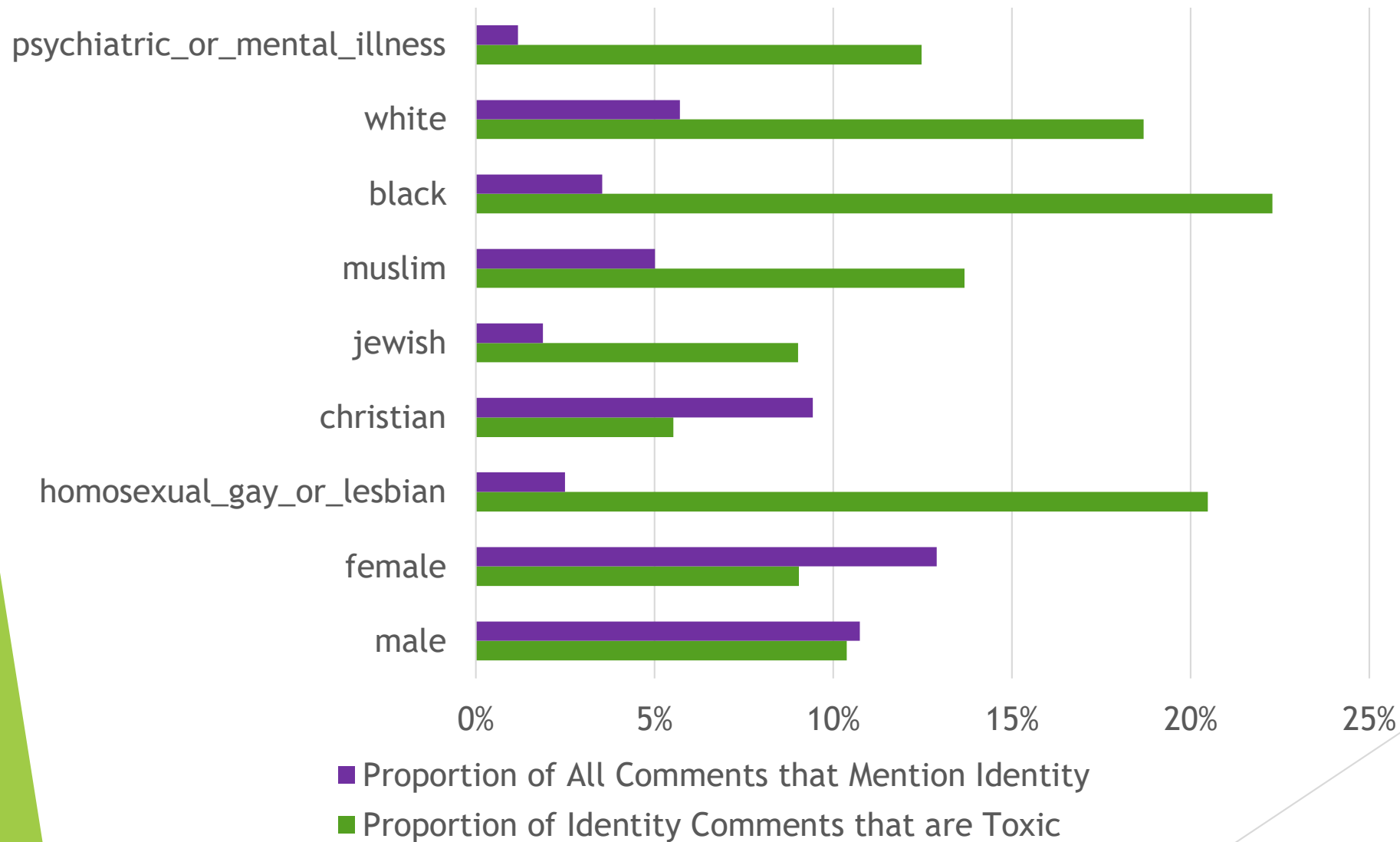
# Identity Attributes

## Toxicity @1

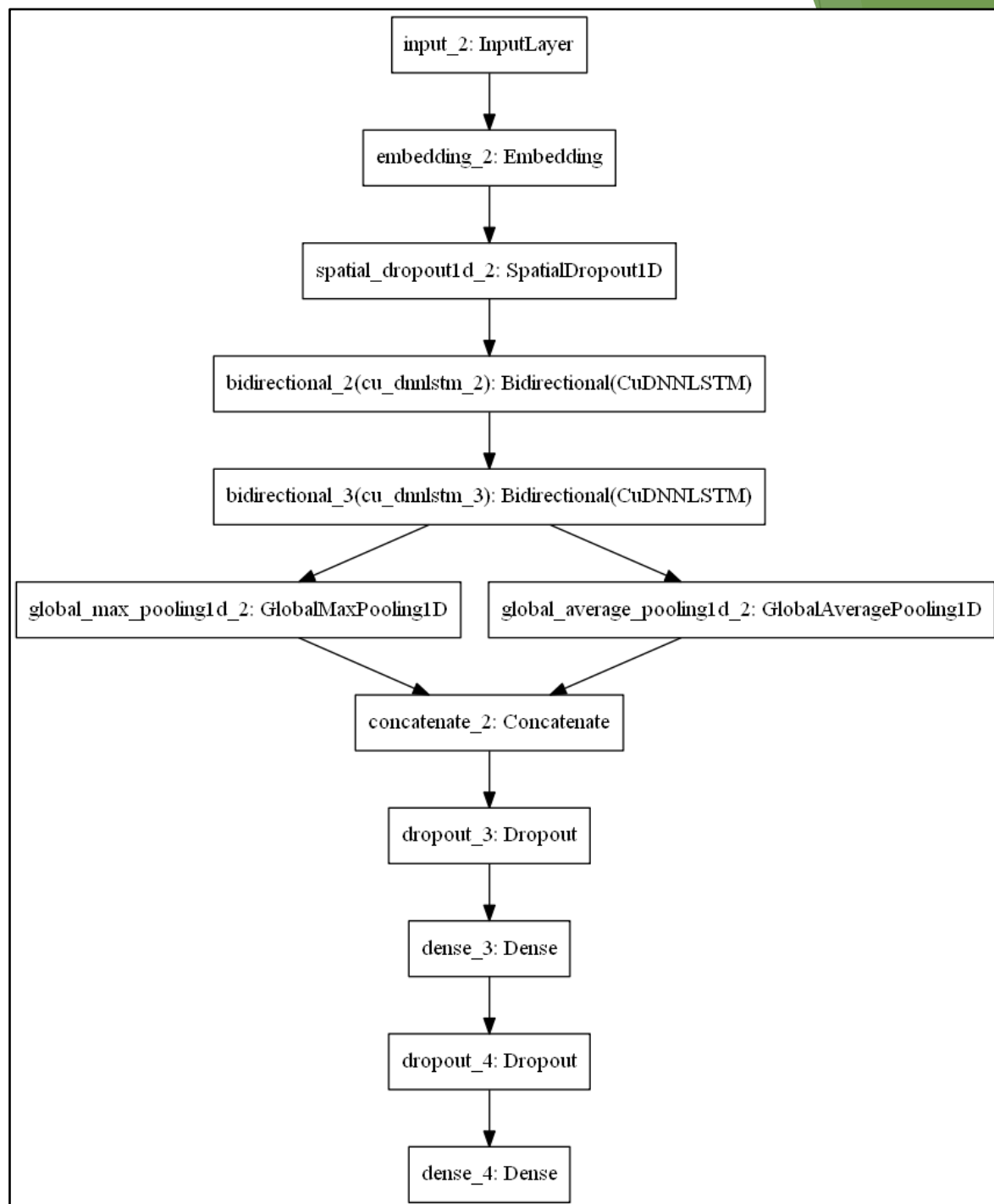
Identity groups	Subgroup AUC	BPSN AUC	BNSP AUC
lesbian	0.93	0.74	0.98
gay	0.94	0.65	0.99
queer	0.98	0.96	0.93
straight	0.99	1.00	0.87
bisexual	0.96	0.95	0.92
homosexual	0.87	0.53	0.99
heterosexual	0.96	0.94	0.92
cis	0.99	1.00	0.87
trans	0.97	0.96	0.91
nonbinary	0.99	0.99	0.90
black	0.91	0.85	0.95
white	0.91	0.88	0.94



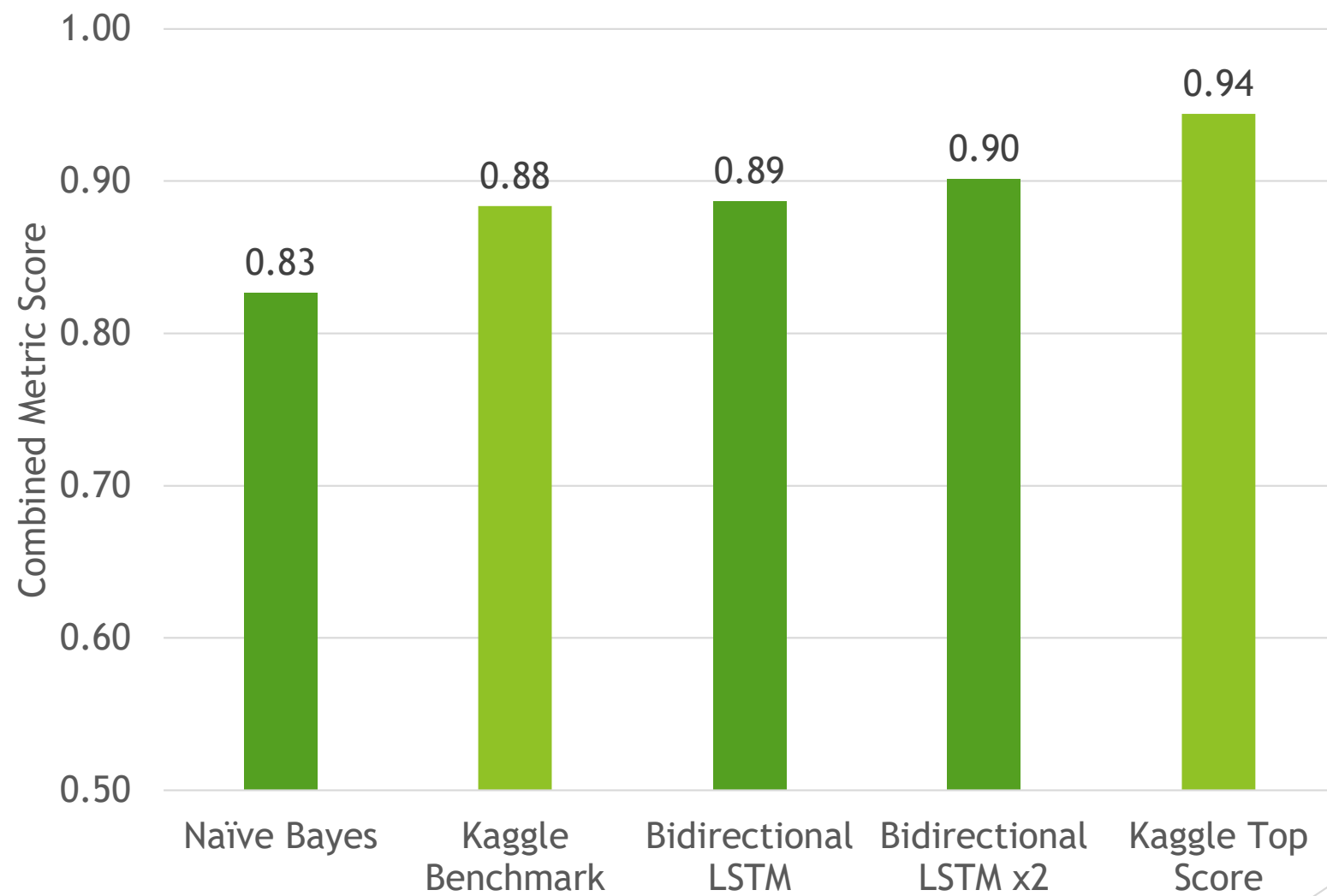
# Data Exploration



# Model Architecture



# Results



# Subgroup Results

	Subgroup AUC	BPSN AUC	BNSP AUC
male	0.88	0.92	0.93
female	0.94	0.92	0.96
homosexual_gay_or_lesbian	0.84	0.78	0.98
christian	0.90	0.95	0.89
jewish	0.88	0.91	0.93
muslim	0.82	0.88	0.93
black	0.83	0.77	0.97
white	0.82	0.84	0.95
psychiatric_or_mental_illness	0.88	0.97	0.85

# Conclusion

I can predict comment toxicity well using Bidirectional neural networks

This will be useful for flagging comments for removal

Race-based toxicity is particularly challenging to identify

Model improvement is possible with more computing resources