Designing Toxic Content Classification for a Diversity of Perspectives

Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, Michael Bailey

How Google's Jigsaw Is Trying to Detoxify the Internet

Can Facebook Use AI to Fight Online Abuse?

The task of detecting abusive posts and comments on social media is not entirely technological

Instagram to use artificial intelligence to detect bullying in photos

The move highlights efforts from tech companies to use automation to moderate their platforms.

07-17-20

Twitter automatically flags more than half of all tweets that violate its rules

07-17-20

Twitter automatically flags more than half of all tweets that violate its rules

Twitter still failing women over online violence and abuse

Users may disagree about what constitutes toxic content online, leading to “gray areas” in automated classification

How do users from diverse backgrounds interpret toxic content online?

Survey Participants (US only)

20K
participants

100K
comments

Survey Participants (US only)

20K
participants

27%
minorities

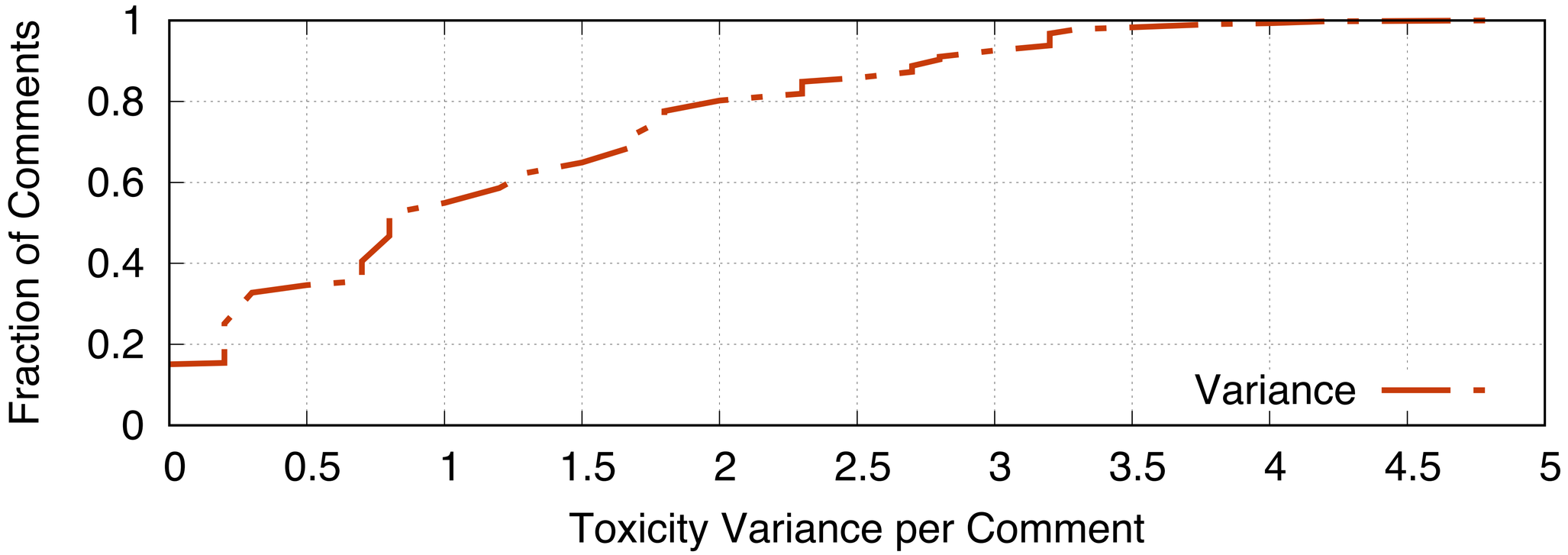
13%
LGBTQ+

100K
comments

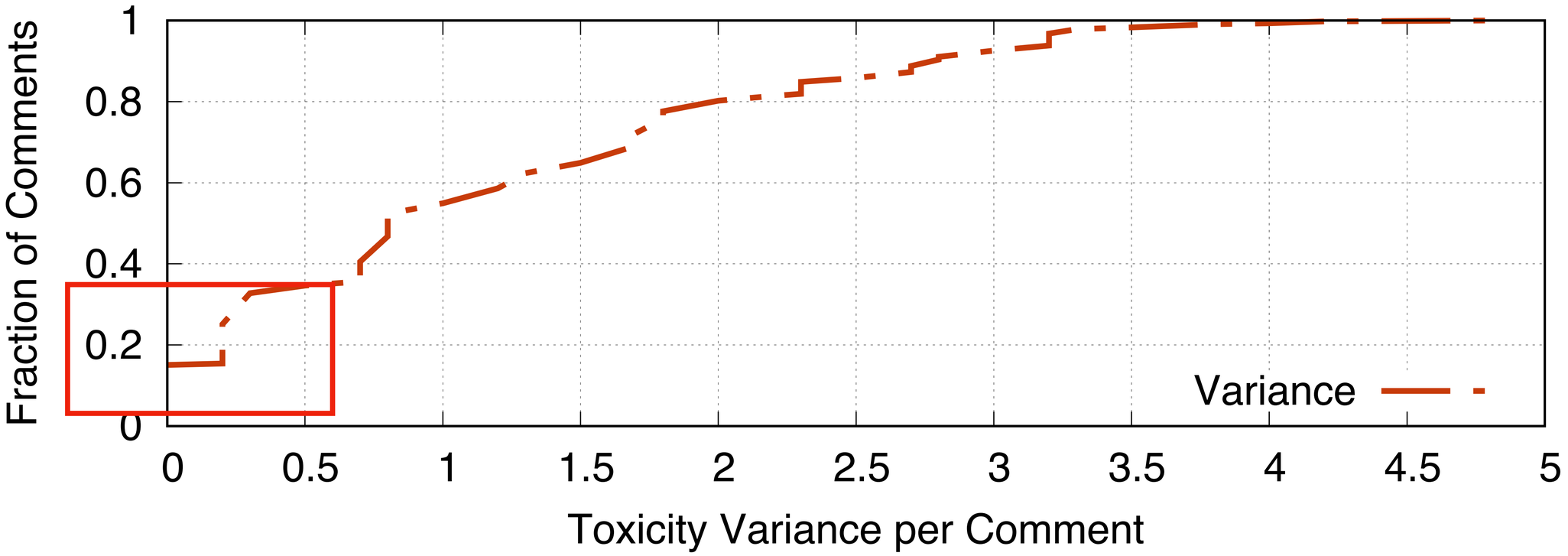
51%
religious

50%
parents

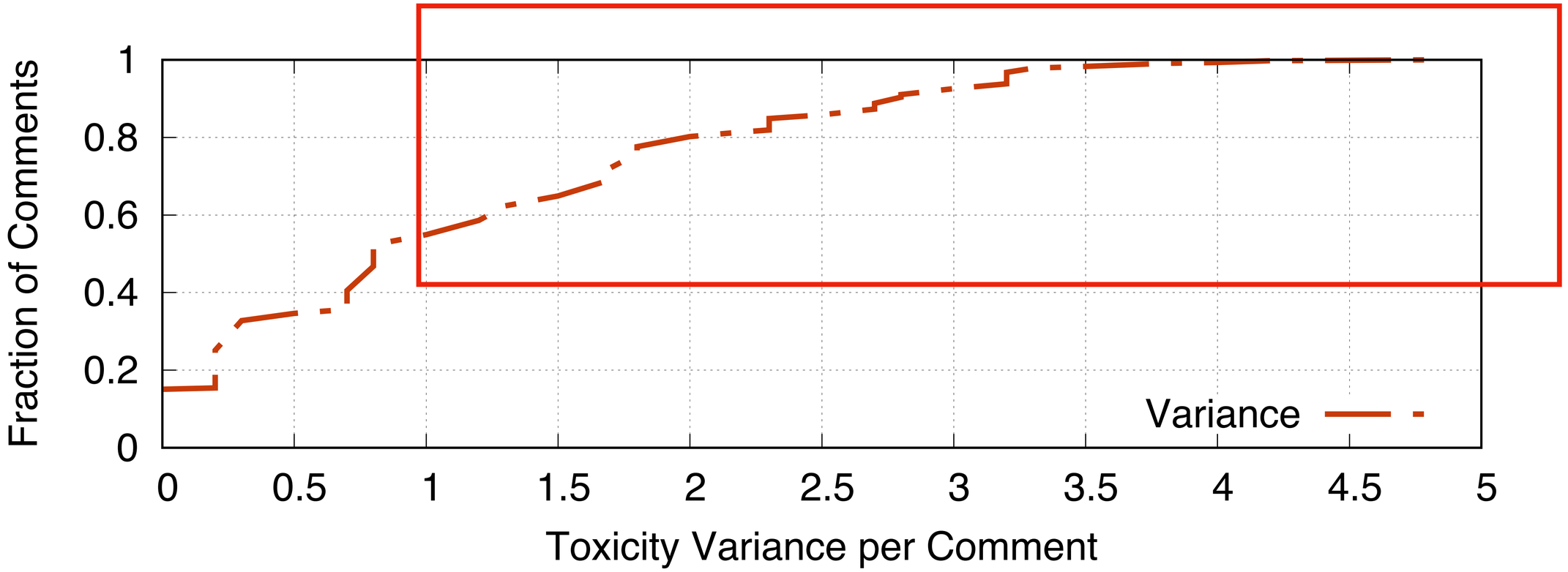
Disagreement Between Raters



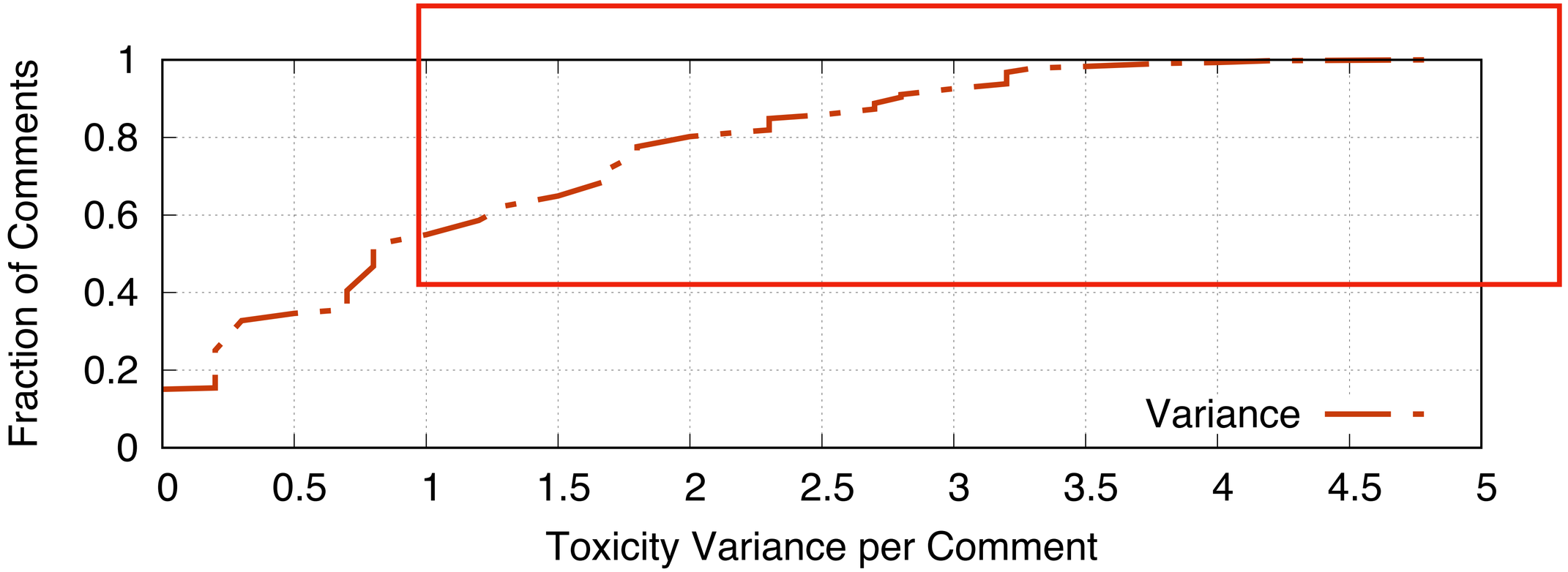
Disagreement Between Raters



Disagreement Between Raters



Disagreement Between Raters



Participants regularly disagreed on whether a comment was toxic!

What factors might explain disagreement between raters?

Identities, Experiences, and Toxicity

1.2x
18-24 vs. 35-44

1.1x
minorities vs. non-minority

1.6x
LGBTQ+ vs. non-LGBTQ+

1.3x
parent vs. non-parent

Identities, Experiences, and Toxicity

1.2x
18-24 vs. 35-44

1.1x
minorities vs. non-minority

0.8x
witnessed toxic content

1.6x
LGBTQ+ vs. non-LGBTQ+

1.3x
parent vs. non-parent

1.5x
target of toxic content

Personalized abuse protections can help account for diverse perspectives in toxic content classification

Fine Tuning Toxicity Classifiers

0.35

avg. precision

0.37

avg. accuracy

Fine Tuning Toxicity Classifiers

0.35
avg. precision

0.6
avg. precision

personalized tuning

0.37
avg. accuracy

0.68
avg. accuracy

Current automated toxicity classifiers fail to generalize to users with varied lived experiences.

Building new anti-abuse defenses must take into account the diverse perspectives of Internet users.