Incorporating User Experiences to Improve Automated Detection of Toxic Content Online

Deepak Kumar – ESRG “Lightning” Talk

Content warning: threats of violence, identity attacks, insults

More Americans are being harassed online because of their race, religion, or sexuality

More Than One-Quarter of Americans Experience Severe Online Harassment, ADL Survey Finds

Survey shows members of marginalized groups experience more hate

1 in 3 Americans Suffered Severe Online Harassment in 2018

2018 really was more of a dumpster fire for online hate and harassment, ADL study finds

Roughly four-in-ten Americans have personally experienced online harassment, and 62% consider it a major problem. Many want technology firms to do more, but they are divided on how to balance free speech and safety issues online

Near-Term Research Trajectory

- SoK: Hate, Harassment, and the Changing Landscape of Online Abuse (Accepted to Oakland 2021)
- Measuring the Influence of Conflicting User Perspectives on Toxic Content Classification (In submission)
- Understanding the Relationships between Abusers and Targets of Online Abuse (working on it)
- Evaluating Personalized Models for Automated Toxicity Detection (future work)

Near-Term Research Trajectory

- SoK: Hate, Harassment, and the Changing Landscape of Online Abuse (Accepted to Oakland 2021)
- **Measuring the Influence of Conflicting User Perspectives on Toxic Content Classification (In submission)**
- Understanding the Relationships between Abusers and Targets of Online Abuse (working on it)
- Evaluating Personalized Models for Automated Toxicity Detection (future work)

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.

Google's comment-ranking system will be a hit with the alt-right

The company's API for scoring toxicity in online discussions already behaves like a racist hand dryer.

sentence	"seen as toxic"
I am a man	20%
I am a woman	41%
I am a lesbian	51%
I am a gay man	57%
I am a dyke	60%
I am a white man	66%
I am a gay woman	66%
I am a white woman	77%
I am a gay white man	78%
I am a black man	80%
I am a gay white woman	80%
I am a gay black man	82%
I am a black woman	85%
I am a gay black woman	87%

Where and how do current toxicity detection systems fail?*

*and how can we make them better?

Do automated toxicity systems generalize to a diverse set of users?

Measuring User Perspectives on Toxic Content

- Conducted a survey of 17,280 unique participants (through MTurk) requesting their perspectives on 20 comments online, demographic questions, and experiences with toxic content online
  - Asked if content was toxic (Likert scale) and if so, to place into one of 5 subcategories of abuse: Insult, Profanity, Identity Attack, Threat, Sexual Harassment
- Each comment was rated by N=5 raters, to balance cost with soliciting many ratings per comment to measure differences
  - In total rated 107,620 comments

Sourcing and Sampling Toxic Content

- Aggregated a corpus of 550K comments from Twitter, Reddit, 4chan
- Each comment was labeled through Google's Perspective API, which provides a score from 0 – 1 on toxicity
- Wanted to prioritize comments with *mixed* agreement across raters to understand why and where perspectives were different
    - Conducted a pilot study (N = 200) rated 800 comments, 80 from each stride, to identify comments with least agreement

Stride	% Agreement	% Final Dataset
0.0 – 0.1	90%	5%
0.1 – 0.2	81.8%	5%
0.2 – 0.3	80%	5%
0.3 – 0.4	76.4%	10%
0.4 – 0.5	71.4%	10%
0.5 – 0.6	65.2%	15%
0.6 – 0.7	68.3%	15%
0.7 – 0.8	65.2%	20%
0.8 – 0.9	76.4%	10%
0.9 – 1.0	80%	5%

Overall Perceived Toxicity

- 53% of comments have a median rating of “Not at all toxic” across all five raters, while 1% of comments have a median rating of “Extremely toxic”
  - *“They’re like the polar opposite of limp-writes smug douchebag homo’s [sic] and liberals who buy a Prius and think they’re better than everyone else.”*
- 39.1% of comments were rated either “Slightly toxic” or “Moderately toxic”
  - *“Kids with hoodies are going to be our future criminals”*
  - *“Women can’t be responsible for hiring people. It is foolish to entrust hormonal women with the most important part of the company.”*

Subcategories of Toxic Comments

- Asked participants to place toxic comments into one of several subcategories
  - Participants could select multiple categories per comment

Category	Overall	Slightly Toxic	Moderately Toxic	Very Toxic	Extremely Toxic
Insult	67%	76%	85%	89%	89%
Profanity	52%	59%	69%	74%	78%
Identity Attack	51%	57%	70%	79%	85%
Threat	31%	30%	44%	54%	59%
Sexual Harassment	18%	18%	27%	34%	39%

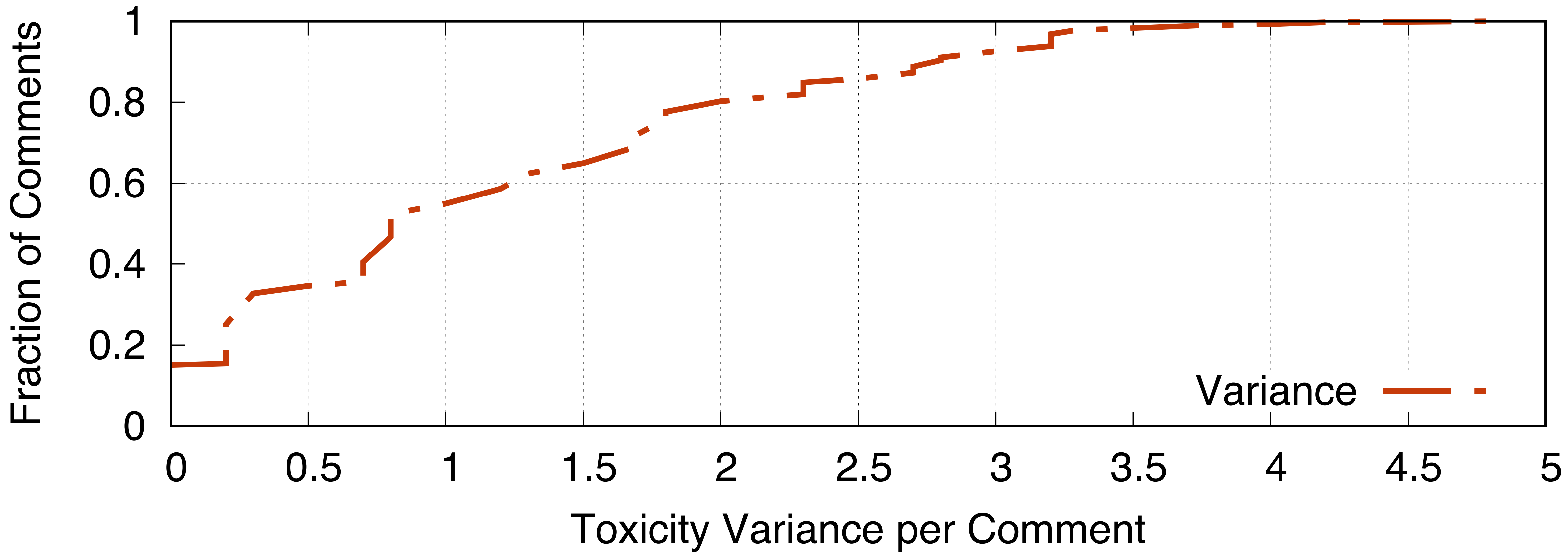
Subcategories of Toxic Comments

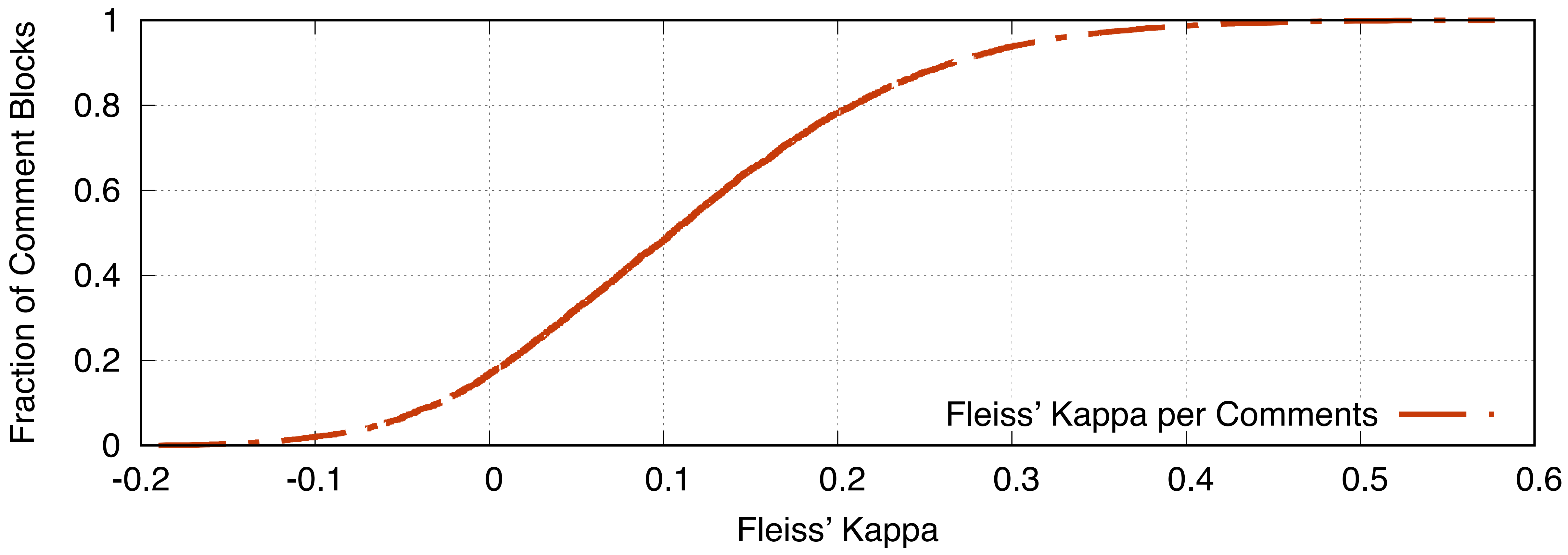
- Asked participants to place toxic comments into one of several subcategories
  - Participants could select multiple categories per comment
- Most common types of toxic content are insults, profanity, identity attack
  - Identity attacks more prevalent for “Extremely toxic” comments
- Threats, sexual harassment are more regularly rated “extremely toxic” compared to other categories

Category	Overall	Slightly Toxic	Moderately Toxic	Very Toxic	Extremely Toxic
Insult	67%	76%	85%	89%	89%
Profanity	52%	59%	69%	74%	78%
Identity Attack	51%	57%	70%	79%	85%
Threat	31%	30%	44%	54%	59%
Sexual Harassment	18%	18%	27%	34%	39%

Disagreement Between Raters

- Interested in understanding how often participants disagree, and why => variance of ratings
- Only 15% of comments have a variance of 0.0 (perfect agreement), while 7.5% of comments have a variance of 3.0 or higher
- Raters also frequently disagreed about subcategories of harassment (Fleiss’ Kappa)





What factors explain disagreements between raters?

Modeling Participant Decision Making

- Treat each rating task as a Bernoulli trail where labeling a comment as "Moderately toxic" or higher is toxic (1) and all the ratings are benign (0)
- Model the frequency of success as a quasi-Binomial distribution, with categorical parameters drawn from demographic questions
- Compute odds that certain demographic groups perceive toxic content at higher rates

Demographic	Treatment	Reference	Odds
Gender	Female	Male	0.952
	Non-binary	Male	0.707
Age	18-24	35-44	1.238*
	25-34	35-44	1.227*
	45-54	35-44	0.972
	55-64	35-44	0.980
	65+	35-44	0.977
Race	Minority	Non-minority	1.126*
LGBTQ+	LGBTQ+	Not LGBTQ+	1.644*
Political affiliation	Conservative	Liberal	1.024
	Independent	Liberal	0.901*
Importance of religion	Not too important	Not important	1.216*
	Somewhat important	Not important	1.572*
	Very important	Not important	1.840*
Parent	Is a parent	Not a parent	1.330*
Education	College	High school	1.139*
	Advanced degree	High school	1.365*
Impact of technology on society	Very negative	Neutral	0.803*
	Somewhat negative	Neutral	0.870
	Somewhat positive	Neutral	0.970
	Very positive	Neutral	1.142*
Toxic content a problem?	Rarely	Not a problem	1.030
	Occasionally	Not a problem	0.958
	Frequently	Not a problem	1.029
	Very frequently	Not a problem	1.125*
Party most responsible	Law enforcement	Bystander	1.282*
	Receiver	Bystander	0.716*
	Platform	Bystander	0.706*
	Sender	Bystander	0.619*
Witnessed toxic content	Yes	No	0.780*
Target of toxic content	Yes	No	1.483*

Modeling Participant Decision Making

- Treat each rating task as a Bernoulli trail where labeling a comment as "Moderately toxic" or higher is toxic (1) and all the ratings are benign (0)

- Model the frequency of success as a quasi-Binomial distribution, with categorical parameters drawn from demographic questions

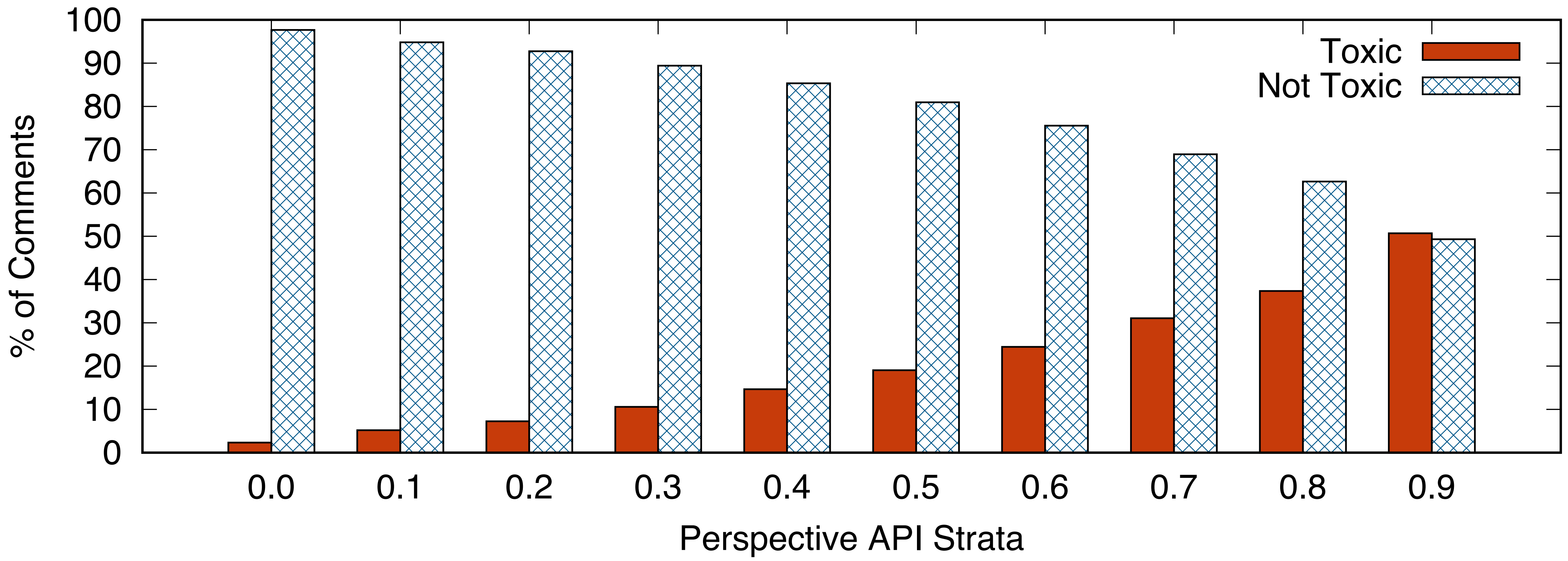
- Compute odds that certain demographic groups perceive toxic content at higher rates

Demographic	Treatment	Reference	Odds
Gender	Female	Male	0.952
	Non-binary	Male	0.707
Age	18-24	35-44	1.238*
	25-34	35-44	1.227*
	45-54	35-44	0.972
	55-64	35-44	0.980
	65+	35-44	0.977
Race	Minority	Non-minority	1.126*
LGBTQ+	LGBTQ+	Not LGBTQ+	1.644*
Political affiliation	Conservative	Liberal	1.024
	Independent	Liberal	0.901*
Importance of religion	Not too important	Not important	1.216*
	Somewhat important	Not important	1.572*
	Very important	Not important	1.840*
Parent	Is a parent	Not a parent	1.330*
Education	College	High school	1.139*
	Advanced degree	High school	1.365*
Impact of technology on society	Very negative	Neutral	0.803*
	Somewhat negative	Neutral	0.870
	Somewhat positive	Neutral	0.970
	Very positive	Neutral	1.142*
Toxic content a problem?	Rarely	Not a problem	1.030
	Occasionally	Not a problem	0.958
	Frequently	Not a problem	1.029
	Very frequently	Not a problem	1.125*
Party most responsible	Law enforcement	Bystander	1.282*
	Receiver	Bystander	0.716*
	Platform	Bystander	0.706*
	Sender	Bystander	0.619*
Witnessed toxic content	Yes	No	0.780*
Target of toxic content	Yes	No	1.483*

Can we use these results to improve automated toxicity classifiers?

Benchmarking Toxicity Classifiers: Accuracy

- We don't present accuracy in aggregate since our sampling mechanism is inherently biased
    - Instead, we present performance by stride, "toxic" score is > 0.75 or > "Moderately toxic"
- Only a *weak* correlation between participant's Likert ratings and the Perspective API score ($r = 0.39$, $p < 0.01$)
- At highest stride (>0.9), accuracy of classifier was 51%



Benchmarking Toxicity Classifiers: RMSE

- To compute how far away toxicity scores were from Perspective, we computed RMSE
    - Averaged and normalized ratings per comment, compared to Perspective rating
- Error increases as strides increase, suggesting that the API struggles to match ground truth at high decision thresholds

Stride	Accuracy	RMSE
0.0 – 0.1	0.98	0.12
0.1 – 0.2	0.95	0.14
0.2 – 0.3	0.93	0.18
0.3 – 0.4	0.90	0.24
0.4 – 0.5	0.85	0.30
0.5 – 0.6	0.81	0.36
0.6 – 0.7	0.76	0.42
0.7 – 0.8	0.50	0.48
0.8 – 0.9	0.37	0.55
0.9 – 1.0	0.51	0.55

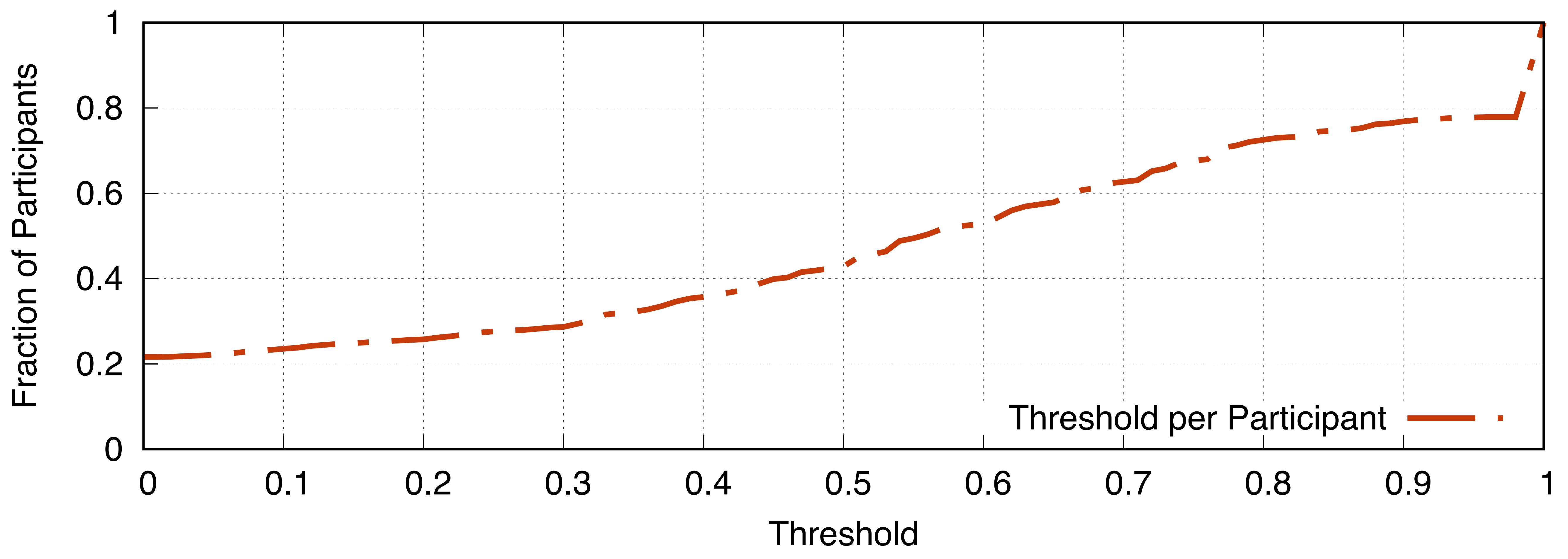
Tuning Toxicity Classifiers

- Recent work from Jigsaw suggests that monolithic toxicity classifiers can be improved simply by <u>tuning</u> the threshold at which it operates
  - Towards “personalized” toxicity detection systems
- In aggregate, a monolithic classifier achieves the best performance at a threshold of 0.49, which achieves a precision of 0.35 and an accuracy of 0.37

Can we do better?

Tuning Toxicity Classifiers – Individual Tuning

- We tuned the classifier for each participant
    - Looked for the threshold that maximizes the F1 score (precision + recall)
- 71.5% of participants saw an improvement in accuracy over the one-size-fits-all model!



Tuning Toxicity Classifiers – Cohort Tuning

- Tuned based on demographic cohort to identify performance improvements
- Demographic tuning in aggregate offers a smaller improvement over the aggregate classifier
  - Individual tuning is a more effective strategy than grouping raters from demographics into buckets

Demo	Max Precision		Max Accuracy	
	Value	% Change	Value	% Change
Religion	0.4	14.3%	0.41	10.8%
Politics	0.37	5.7%	0.37	0%
Age	0.44	25.7%	0.44	20.6%
Gender	0.39	11.4%	0.40	7.5%
Race	0.36	2.9%	0.36	-2.7%
Parent	0.37	5.7%	0.39	5.4%
LGBTQ+	0.36	2.9%	0.37	0%

Discussion + Next Steps

- Our work demonstrates how automated toxicity detection systems fail to generalize across a wide variety of users with varied lived experiences
  - One-size-fits-all model has poor accuracy; personalized tuning helps
- Idea: If content alone makes this problem hard, **what other features** are important in understanding the spread of toxic content online?
- Idea: **Can fully personalized models** be an effective strategy for mitigating the harm of toxic content online?

Near-Term Research Trajectory

- SoK: Hate, Harassment, and the Changing Landscape of Online Abuse (Accepted to Oakland 2021)
- Measuring the Influence of Conflicting User Perspectives on Toxic Content Classification (In submission)
- **Understanding the Relationships between Abusers and Targets of Online Abuse (working on it)**
- **Evaluating Personalized Models for Automated Toxicity Detection (future work)**