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[†]
[◁]*Stanford University* [○]*Google* [†]*University of Illinois at Urbana-Champaign*

Abstract

In this work, we demonstrate how existing classifiers for identifying toxic comments online fail to generalize to the diverse concerns of Internet users. We survey 17,280 participants to understand how user expectations for what constitutes toxic content differ across demographics, beliefs, and personal experiences. We find that groups historically at-risk of harassment—such as people who identify as LGBTQ+ or young adults—are more likely to to flag a random comment drawn from Reddit, Twitter, or 4chan as toxic, as are people who have personally experienced harassment in the past. Based on our findings, we show how current one-size-fits-all toxicity classification algorithms, like the Perspective API from Jigsaw, can improve in accuracy by 86% on average through personalized model tuning. Ultimately, we highlight current pitfalls and new design directions that can improve the equity and efficacy of toxic content classifiers for all users.

1 Introduction

Online hate and harassment is a pernicious threat facing 48% of Internet users [52]. In response to this growing challenge, online platforms have developed automated tools to take action against toxic content (e.g., hate speech, threats, identity attacks). Examples include Yahoo's abusive language classifier trained on crowdsourced labels attached to news comments [43], Google Jigsaw's Perspective API, which is trained on Wikipedia moderation verdicts for abuse as well as samples from other online communities [35, 57], and Instagram's recent classifier that detects harassing comments posted as a reply to photos [32].

Although platforms have used these classifiers to address toxic content in direct violation of their policies [41], a variety of content that is not toxic enough to violate policy may still cause harm to Internet users [1]. These "gray areas" stem from the fact that users may disagree about what constitutes toxic content online based on their lived experiences, cultural perspective, political views towards free speech, or access to appropriate context [26, 50]. While prior research has demonstrated that certain groups are more at-risk of experiencing online hate and harassment [45, 52], no study has investigated how users from diverse backgrounds interpret online toxicity or how their views on what content they would like to see online differ. Understanding these nuanced differences is an important first step to designing harassment defenses for diverse Internet users.

In this work, we investigate divergent user interpretations of toxic content and identify whether current classifiers can be tuned to accommodate a diversity of perspectives. At the core of our study, we develop a survey instrument that asks 17,280 participants to rate and label the toxicity of 20 random comments drawn from 107,620 Twitter, Reddit, and 4chan comments. In tandem, we collect demographic data and log participants' previous exposure and experiences with online harassment. Taken together, our survey instrument provides access to a diverse set of perspectives on why people deem certain comments as toxic. We explore this data in three steps: we investigate user ratings of toxic content in aggregate, we identify the factors that result in identical comments receiving divergent ratings, and finally, we demonstrate how modern classifiers can better accommodate differing user perspectives.

Participants frequently disagree on whether comments are toxic. In aggregate, participants labeled 53% of our dataset as "not toxic", 39% as "slightly" or "moderately toxic" and the remaining 8% as "very" or "extremely toxic". However, 85% of comments exhibited some form of disagreement, including whether participants were comfortable seeing the comment on any online platform. Even when participants uniformly agree that a comment is toxic, they disagree about the subcategory the comment belonged to (e.g., a threat versus an insult). As such, a la carte models that isolate individual classes of toxic content—for instance, identity-based attacks [55]—may fail

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.

to adequately meet the needs of a user base with diverse perspectives on toxic content.

A variety of factors influence how users perceive toxicity. We find that a participant's personal experience with harassment, whether the participant belongs to an at-risk group frequently targeted by harassment [13,45], and a participant's attitudes towards filtering online discourse all correlate with rating a comment as toxic or not. For example, holding all other factors constant, the odds that a participant rates a comment as toxic increase 1.64 times if they identify as LGBTQ+. Alternatively, these odds decrease by 0.78 times for users who regularly witness others targeted by toxic content, potentially due to desensitization. Combined, no single demographic variable or experience defines how participants interpret toxic content, underscoring the need for diverse raters in data labeling and model construction.

Finally, we investigate how we might leverage current state-of-the-art classifiers to enable diverse user perspectives of toxic content online. We focus on Jigsaw's Perspective API [23] and Instagram's comment nudge [32]. As a baseline, we find for content that Perspective deemed 90% likely to be toxic, only 50% of our participants agreed. Similarly, Instagram's classifier flagged only 27% of comments that a majority of our participants rated as toxic. We propose potential improvements based on *personalized tuning*—finding a threshold for the classifier that is set based on individual responses or in larger demographic groups. These improvements achieve an 86% boost in accuracy per individual and a 22% improvement in accuracy per demographic cohort, highlighting personalized modeling as a future direction in toxicity classification.

We conclude with a discussion of how to overcome the limitations of crowdsourced labeling and one-size-fits-all classification that we identified through our work. To this end, we have shared our results with Jigsaw and have released our labeled dataset[1] to enable other researchers to reproduce our analysis, build new classifiers, and further explore how different individuals perceive toxic behavior online.

2 Background & Related Work

2.1 What is toxic content?

We use the term *toxic content* as an umbrella for identity-based attacks such as racism on social media [2,21,55], bullying in online gaming or replies to posts [36,50], trolling [10], threats of violence, sexual harassment, and more [47, 52]. These attacks represent a subset of abuse stemming from *hate and harassment*, a broader threat that encompasses any activity where an attacker attempts to inflict emotional harm on a target (e.g., stalking, doxxing, sextortion, and intimate partner violence) [11,52]. Unlike spam, phishing, or related abuse classification problems that can rely on expert raters, toxic content is an inherently subjective problem. For the purposes of our study, we focus exclusively on text-based toxic content, but attacks may also extend to images and videos [58].

Previous studies have shown that some demographic cohorts in the United States are more likely to receive and report toxic content than others [12,45]. For example, a survey by Pew found that men were more likely to report experiencing offensive name calling and physical threats, while women were more likely to experience sexual harassment [45]. Beyond gender, Black adults were found to report higher rates of name calling and purposeful embarrassment [45], while people who identify as LGBTQ+ were three times as likely to report offensive name calling, physical threats, and sexual harassment [6, 13]. Similarly detailed demographic studies from various global perspectives are not yet available. In order to ensure that automated detection works for all people, including at-risk groups, we argue that it is critical to first understand how different people perceive toxic content and how perceptions generalize across Internet users.

2.2 Detecting toxic content

Security researchers and practitioners have proposed a multitude of blocklist-based, machine learning, and natural language processing techniques to detect toxic content. The simplest of these approaches rely on manually curated lists of abusive words or users, such as HateBase's corpus of hate speech related terms [27], or BlockTogether's list of abusive Twitter accounts [34]. These provide targeted protections against exact matches of terms or known abusers, but fail to generalize to other types of toxic content, or in the context of blocklists, anonymous posts.

More sophisticated machine learning models include Yahoo's regression model trained on a corpus of roughly 300,000 abusive comments with crowdsourced labels that included hate speech, derogatory messages, and profanity [43]. Using a variety of NLP-based features, they found their classifier could achieve an AUC of 0.90, though domain-specific language and concept drift (e.g., changes in abusive terms) degraded performance over time. Since then, a variety of models have incorporated crowdsourced labels such as Wikipedia moderation decisions [18,57], in-game conversations [4], and social media posts [9,15,17,19,51,54] to varying degrees of success. In another example, Founta et al. leveraged HateBase to build crowdsourced sublabels from participants for abusive tweets, and then characterized a sample of Twitter data [22]. Related approaches have examined how to take a model trained for one community and apply it to a separate community or site to avoid the cost of generating a labeled training set [8]. Finally, several studies have focused on latent annotator bias in datasets [46,56] and also demonstrated that disagreements between raters for social tasks may explain why classifiers excel on benchmarks but suffer in practice [24].

[1] https://data.esrg.stanford.edu/study/toxicity-perspectives

Prominent models deployed at-scale today include Jigsaw's Perspective API, a deep learning classifier for detecting toxic comments which is used by the New York Times, Disqus, and other news sites for moderating toxic comments [23]. Similarly, Instagram recently deployed a model for nudging users away from posting comments that the classifier perceives as harassment due to similar abusive text being reported in the past [32]. We evaluate how these models generalize across users in Section 6.

2.3 Other intervention strategies

While our work focuses on how best to train classifiers to automatically detect toxic content, researchers have also considered a variety of other strategies for moderating toxic content. One example is building mechanisms into online platforms to escalate conflicts to community tribunals who are empowered to remove toxic content and take action against abusive users [40]. Other examples include enabling bystanders to simply report toxic content [16], or providing family and friends with tools to assist in moderating toxic content on behalf of a target [5, 39]. All of these techniques leverage community and context to overcome the limitations of automated classification, but alone may fail to scale to the hundreds of millions of interactions that happen online every day. Additionally, these systems cannot relieve moderators of the emotional burden of reviewing toxic content [42].

2.4 Differentiation from prior work

Prior work in evaluating automated toxicity classifiers has focused on either investigating underlying bias in training data, such as flagging comments with the word "gay" as hateful [14, 18], or shown that classifiers are easily manipulated by substituting "offensive" words while retaining semantic meaning [33]. The focus of our work is to first, understand how perspectives of toxic content change based on individual experiences, and second, evaluate the impact these experiences have on automated toxic content detection (Section 6). Prior work identified certain groups to be at higher risk of online harassment [13, 45], however, no work has shown whether these experiences lead to differences in perception of toxic content online. Closest to this is work by Cowan et al. who investigated perceptions of hate speech against three target groups on college campuses. However, their study is limited in scale (N $<$ 500) and not specific to an online context; our work focuses on a broader set of participants, focuses on several categories of toxic content, and is more representative of online discussion. Furthermore, we investigate if implementing a personalized filter—one that better captures the sentiment of participants by their individual experiences—can improve toxicity detection.

Theme raised by participants	**Offensive** N=72 $\kappa = 0.95$	**Hateful** N=74 $\kappa = 0.98$	**Toxic** N=79 $\kappa = 0.9$
Insulting, demeaning, or derogatory	42–44%	55%	58–62%
Identity attack, hate speech, or racist	33–35%	39–41%	33–34%
Profane or obscene	21%	12%	19%
Threatening or intimidating	11%	11%	16%
Not constructive or off-topic	3%	0%	9–11%
None of the above	29–32%	20–22%	19–20%

Table 1: **Interpretation of the Terms: Offensive, Hateful, and Toxic**—We find the term toxic resulted in the broadest interpretation for our rating task.

3 Methods

3.1 Survey instrument

Our survey consisted of three parts: pre-exercise questions about the participant's attitude towards technology and toxic content, an exercise where the participant rated 20 comments from social media and community forums as toxic or not, and finally, demographic and attention check questions. We provide our full survey instrument in the Appendix. Our study was approved by our institution's IRB.

Selecting terminology and comprehension. As a preliminary step, we first determined what terminology to use for our rating task. An inherent challenge here is the ambiguity of the term *toxic content* or *hate and harassment* and a lack of consensus across researchers and industry [44].

In the absence of common best practices, we ran a pilot study with $N = 300$ participants recruited from Mechanical Turk to identify the terminology we should use in our survey instrument. We asked each participant the open ended question: "When you see a post or comment, what do you look for to decide if it's $< x >$?", where x was one of "hateful", "offensive", or "toxic". We recruited $N = 100$ participants per survey variant. We did not use the term "abusive" as not to overload its meaning with other online abuse such as for-profit cybercrime or unsafe content including drugs or self-harm. After filtering for attention checks, we received a total of $N = 225$ responses.

We reviewed each response and identified five emergent themes, detailed in Table 1. Two independent raters coded every response according to these themes, with multiple themes possible per response. Coding achieved an interrater agreement Cohen's kappa $\kappa > 0.9$ for all three variants, indicating strong agreement.[2] We found that participants most often interpreted "offensive" to mean comments that were insulting, profane, or an identity-based attack. Participants even more narrowly construed the term "hateful" to mean comments that

[2] In the event that a rater ascribed multiple themes to a single open ended response, we required both raters to select the same set of themes to constitute agreement.

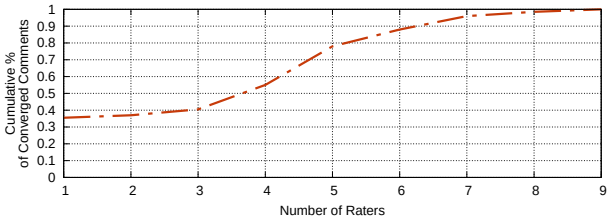

Figure 1: **Toxicity Convergence**—We observe an inflection point where after five participants rate a comment, the benefit of additional perspectives falls off.

involved an identity-related attack or insult. On the other hand, "toxic" encompassed the largest set of themes, where participants also considered whether a comment was constructive or off-topic, and whether a comment was threatening. Based on our findings, we adopted "toxic" as our final survey term to describe our rating task to participants.

Determining the number of ratings per comment. Our survey instrument had to satisfy two competing goals: capturing a diverse enough set of ratings to measure divergence among participants while also maximizing the number of comments rated by participants to produce a meaningfully-sized evaluation corpus. In order to identify how many ratings we should solicit per comment, we ran a pilot survey where 100 participants rated a fixed set of 200 manually curated comments. Each comment was rated by 10 unique participants. Participants selected their rating on a five-point Likert scale ranging from "Not at all toxic" to "Extremely toxic".

We then measured how quickly each comment's ratings converged to its average toxicity score. In this context, we define the average toxicity score to be the fraction of participants that labeled a comment as "Moderately toxic" or greater per comment (the top three ratings of our Likert scale). The global average is the average across all raters for each comment. We then measured the number of participants required for the running average rating to fall within 10% of the global average toxicity score per comment.

Figure 1 shows a CDF of the number of ratings required for convergence for our pilot data. With only 2 ratings, 37% of comments had converged to their final distribution. However, with five ratings, we found 78% of the comments had converged to their final distribution with each incremental participant adding only marginal improvements toward the global average. As we needed to balance soliciting as many ratings as possible per comment with the cost of doing so, we selected five participants to rate each comment for this study.

3.2 Sourcing potentially toxic content

We sourced an initial corpus of 549,058 comments from Twitter, Reddit, and 4chan for our study. We selected these platforms as they represent a diverse cross-section of Internet users, are conversation driven, and contain varying degrees of toxic behavior [3, 7, 28]. All data was collected between December 2019 and August 2020. While our dataset does not capture all types of conversations—such as private discussions via messaging apps or "walled gardens" like Facebook—our collection strategy avoids privacy constraints that would otherwise prevent sharing content with random participants on crowdsourcing platforms.

Stride	Aggregate Rating	% Agreement	% Final Dataset
0.0—0.1	Not toxic	90%	5%
0.1—0.2	Not toxic	81.8%	5%
0.2—0.3	Not toxic	80%	5%
0.3—0.4	Not toxic	76.4%	10%
0.4—0.5	Not toxic	71.4%	10%
0.5—0.6	Not toxic	65.2%	15%
0.6—0.7	Not toxic	68.3%	15%
0.7—0.8	Toxic	65.2%	20%
0.8—0.9	Toxic	76.4%	10%
0.9—1.0	Toxic	80%	5%

Table 2: **Interrater Agreement per Stride**—Although raters agree broadly for comments with either low or high toxicity scores, raters show minimal agreement when a comment is scored between 0.5—0.8. As such, we oversample these ranges for our dataset.

Given the class imbalance inherent to each site, where benign content far outweighs toxic content (with the exception of perhaps 4chan), a purely random sampling approach would be prohibitively expensive to gather crowdsourced labels for a sufficiently large volume of toxic content. Instead, we leveraged the Perspective API `TOXICITY` model (discussed in detail in Section 2) to build a stratified sample of potentially toxic content.[3] The API takes as input a sample of text and returns a score between 0 and 1, describing the likelihood that an audience would perceive the text to be toxic.

In order to identify which score ranges correlated with the largest rating disagreement among participants, we ran a pilot survey where 200 participants rated 800 comments, with 80 comments sourced from each 0.1-stride between 0 and 1. For example, we selected 80 comments with a toxicity score of 0—0.1, 80 comments with a score of 0.1—0.2, and so on. Five independent participants rated each individual comment. We then measured the interrater agreement for each stride as shown in Table 2. We found that participants broadly agreed on comments that had a `TOXICITY` score of < 0.3 or > 0.9, with the least agreement when a comment had a score of between of 0.5 and 0.6 and between 0.7 and 0.8. A comment with a score of 0.5 might look like:

> "I'm so sick of this mess. The Dems are not good because the Repubs are bad. The Repubs are not good when the Dems are bad. The enemy of your enemy can still be your enemy. #BothPartiesSuck"

[3] Instagram does not provide a public API, thus we did not consider it when building our dataset.

Table 2 shows the final distribution of comments we include per stride. Our dataset preferentially includes comments with lower interrater agreement, however, we note that at least 5% of comments are sampled from each API stride. Our data distribution by source is 67% Twitter comments, 15% Reddit comments, and 18% 4chan comments. We note our final dataset contains at least 16,000 comments per platform. Our sampling skews towards Twitter as we wanted to guarantee a fixed ratio of comments per stride while maintaining a large corpus (N $>$ 100,000) but were limited by fraction of comments available in each stride from 4chan and Reddit.

3.3 Recruitment and validation

We recruited participants for our final survey through Amazon Mechanical Turk to "Participate in a survey about content online". Previous studies have validated the use of Mechanical Turk in security and privacy contexts [48]. Given the scale of this work, we needed to balance overall cost, fair compensation, and the goal of attracting a large and diverse sample of workers across MTurk. After piloting, we decided to pay $1 for completion. Participants took a median of 13 minutes to complete the task. We only recruited participants with at least a 95% approval rating [49] and restricted participants to residents of the United States. All participants were over the age of 18. As our survey instrument collects potentially sensitive demographic information (gender, sexual orientation, race, and more), we provided an option to decline every demographic question. As mentioned previously, our survey was approved by our IRB.

In order to validate a participant's responses, we relied on an attention check question at the end of the survey that asked participants to recall what term we had used throughout the survey (i.e., toxic). Additionally, we included an open ended question asking participants to describe how they define toxic content (akin to our pilot) and set a manually identified threshold on this response. We solicited new participants until we reached our $n = 5$ threshold per comment. Our final dataset consists of 17,280 participants and 107,620 rated comments.

Table 3 outlines the demographic distribution of our participant pool. Participants were evenly split across men and women, with a median age range of 25–34. Most participants identified as White, non-Hispanic (71%), and did not identify as a member of the LGBTQ+ community (81%). Attitudes towards religion were mixed with most participants either deeming religion not important (32%) or very important (31%). Political attitudes were mixed across Liberal, Independent, and Conservative participants. Our participants also split evenly between parents and non-parents. Our sample does not perfectly align to the US Census demographics for all demographic cohorts [53]. However, our modeling results in Section 5 control per demographic cohort and will stay consistent even if some cohorts are over or under sampled. Overall, our recruitment provided access to a variety of groups that historically are more likely to be the targets of toxic content. Our dataset is available at https://data.esrg.stanford.edu/study/toxicity-perspectives.

Demographic	**Cohort**	**% Respondents**
Gender	Male	46%
	Female	52%
	Nonbinary	1%
Age	18 – 24	12%
	25 – 34	40%
	35 – 44	25%
	45 – 54	13%
	55 – 64	7%
	65+	3%
Race & Ethnicity	Non-minority	71%
	Minority	29%
LGBTQ+ status	Not LGBTQ+	81%
	LGBTQ+	16%
Religion importance	Not important	32%
	Not too important	12%
	Somewhat important	23%
	Very important	31%
Political attitude	Liberal	40%
	Independent	27%
	Conservative	27%
Parent	Yes	52%
	No	47%

Table 3: **Demographics of Respondents**—Our recruitment strategy provided access to a diverse set of raters, including members of communities that are historically at-risk. Not all percentages sum to 100% due to some participants declining to provide demographic information.

3.4 Ethical considerations

Given that our experiments expose participants to potentially toxic content, on the Mechanical Turk description screen we included an initial warning that described our rating task and the potential harms that might arise from participating. We stated:

> Risks related to this research include feeling targeted or potentially hurt by viewing potentially toxic comments and recalling negative experiences in the past regarding your personal experience with toxic comments online.

At this point, participants could choose to accept the rating task or simply move on without any exposure. After accepting the task, participants consented to a longer agreement, that again reminded participants that they would be exposed to toxic content multiple times. Additionally, our stratified sampling approach avoided most egregious toxic content as detected by existing automated classifiers, where there was unlikely to be any disagreement. This is in line with multiple prior studies that rely on crowdsourcing for toxic content

judgements [4, 18, 35]. Overall, participants voluntarily saw a small number of potentially toxic comments in a short session, most of which were rated to be only moderately toxic and which are most in line with conversations that broadly occur on the Internet.

4 Toxicity, Filtering, and Removal Decisions

We examine how often participants deem a comment toxic and the frequency that participants disagreed in rating the severity of toxicity per comment. Additionally, we explore what classes of toxic content (e.g., sexual harassment, profanity) participants were most aligned in recognizing and ultimately their personal beliefs of whether such content should be allowed online.

4.1 Overall perceived comment toxicity

Each comment in our dataset includes five independent toxicity ratings drawn from a Likert scale ranging from "Not at all toxic" to "Extremely toxic." We considered two strategies for aggregating these ratings into an overall non-binary toxicity score per comment. The first, *max rating*, selects the maximum toxicity rating across all participants for a comment. The second, *median rating*, selects the median rating across all participants. Figure 2 shows the distribution of toxicity scores via these two metrics. Although the maximum rating achieves a relatively even distribution across our Likert scale, this often results from a single outlier among the five raters inflating the measured toxicity, compared to the median rating. As such, we opted for the median rating and use it throughout this work unless otherwise noted.

Overall, 53% of comments in our dataset have a median rating of "Not at all toxic", while only 1% of comments have a median rating of "Extremely toxic". An example of an "Extremely toxic" comment from 4chan was:

> "They're like the polar opposite of limp-wristed smug douchebag homo's [sic] and liberals who buy a Prius and think they're better than everyone else".

Comments that were rated either "Slightly toxic" or "Moderately toxic" make up 39.1% of our labeled dataset. Examples from this category include a Reddit comment like "Kids with hoodies are going to be our future criminals," and 4chan comments like "Women can't be responsible for hiring people. It is foolish to entrust hormonal women to the most important part of the company." These examples show how even mildly toxic comments contain racism and sexism—higher ratings merely represent the perceived intensity of the harassment involved.

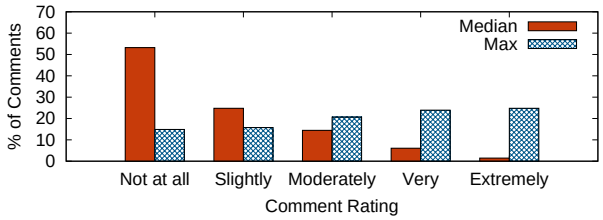

Figure 2: **Median and Maximum Toxicity Per Comment**–The median toxicity rating per comment better captures the central tendency of raters compared to the max rating per comment.

4.2 Subcategories of toxic comments

When participants rated a comment as "Slightly toxic" or higher, they also provided an explanatory label for the toxic behaviors present in the comment. Categories included "Insult," "Identity attack," "Profanity," "Threat," and "Sexual harassment." Participants could select multiple such labels per comment. As an example, the following comment was labeled as both a threat and an identity attack: "I think all you Deep State participants have done Enough [sic] damage to this country.....go jump off the Brooklyn Bridge!". We derived these labels from the themes surfaced by participants in our pilot study, adding "sexual harassment" as an additional theme and removing "off topic" due to the lack of context provided to participants (see Section 3). We refer readers to the Appendix for the detailed instructions that we provided to participants on how to differentiate these categories.

We present a breakdown of the perceived classes of toxic comments in our dataset in Table 4. Each column represents the fraction of comments rated at each toxicity level that fell into each subcategory. Overall, insults are the most common type of toxic comment (67%), followed by profanity (52%), and identity attacks (51%). This is not necessarily an indication that these are the most common toxic behaviors for sites in our sample, but rather these are the toxic behaviors that raters identified. Participants also perceive different sublabels as more or less toxic. For example, 85% of "Extremely toxic" comments involve an identity attack, whereas the same is true for only 57% of comments rated "Slightly toxic" or lower. We also investigate the reverse—which is the fraction of comments in each sublabel that fall into each toxicity level, and find that participants perceive threats and sexual harassment as "Extremely toxic" (3.3%, 3.7% of comments respectively) at a higher rate than identity attacks (2.9%), profanity (2.6%), and insults (2.3%).

4.3 Frequency and intensity of disagreement

While our overall score provides guidance on whether a plurality of participants view a comment as toxic or not, in practice we are interested in how often participants disagree and why. For example, of all comments with a median toxicity of "Not

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%

Table 4: **Categories of Toxic Content Recognized by Participants**–Participants were most likely to perceive content as insulting or containing an identity attack, whereas sexual harassment and threats of violence or rape were less frequent.

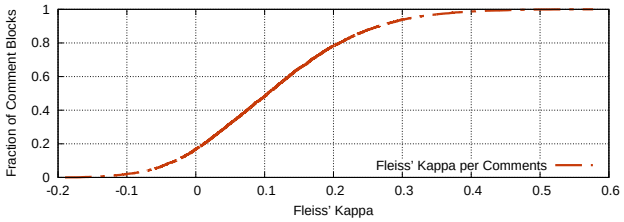

Figure 3: **Interrater Agreement for Subcategory Selection**–Agreement between subcategories of raters is low, with a median interrater agreement score κ of 0.10. The highest agreement between raters only reached 0.57 (moderate agreement), highlighting the difference between rater definitions of subcategories of toxic content.

at all toxic", only 28% have uniform agreement among all five raters. In order to measure diverging perspectives, we calculated the variance of toxicity ratings for each comment. To do this, we treated each rating as an ordinal value between 0 and 4. A variance of 0 indicates perfect agreement for a comment. The maximum variance of 4.8 indicates two competing groups (e.g., two "Extremely toxic," three "Not at all toxic"). We opted for variance over other multi-rater agreement metrics like Krippendorf's alpha or Intra Class Correlation as we are interested in disagreement on individual comments, not between raters.

Only 15% of comments have a variance of 0, indicating all participants rated the comment identically. In aggregate, the median variance of all rated comments is 0.8. However, the spread of scores for comments rated as at least "Slightly toxic" is larger, with a median variance of 1.3 per comment. As an example, the comment from Twitter:

> "At least REDACTED served, unlike you, a weirdo making memes online all day like a little lunatic."

had a variance of 1.3, with two raters finding the comment "Very toxic", one rater finding the comment "Moderately toxic", one finding the comment "Slightly toxic", and one rater not finding the comment toxic at all. In contrast, 7.5% of comments have a variance of 3.0 or greater, indicating widespread disagreement. For example, the comment from Twitter:

> "So you don't want money.... Just free college, loan forgiveness, and (and I'm not sure how this is relevant) healthcare for veterans? I presume you believe only blacks were slaves? Also, your last sentence implies you believe all blacks were slaves..."

had a variance of 3.2. Only 0.03% of comments have a variance of 4.8, which is the maximum amount.

Even when participants agree that a comment has some degree of toxicity, they may still differ on why they feel a comment is toxic. Of comments that participants uniformly deemed toxic, just 0.4% had identical categories assigned by all five participants. We quantify the degree of category disagreement across our dataset using Fleiss' Kappa κ. This score assesses how well a fixed number of raters place a subject into one of several nominal categories—in our case, selecting the same set of categories (e.g., sexual harassment, insult) per comment. In order to arrive at an estimate, we first calculated the κ per block of comments[4] and then calculated the global average.

The best group of five raters achieved a $\kappa = 0.57$, indicating only moderate agreement [37]. The median group of raters achieved a $\kappa = 0.10$, indicating low agreement. These findings illustrate that participants are in general, more likely to agree on toxicity ratings than on the justification for their decision.

4.4 Filtering and removal recommendations

Apart from the perceived toxicity of comments, we also asked participants to make a decision for whether they personally would want to see each comment (e.g., personalized filtering), and whether the comment should be allowed online at all (e.g., global filtering). Of comments rated "Slightly toxic" or higher, participants reported they would personally not want to see 37% of comments. We did not observe a strong distinction between personal filtering and global filtering. In the event a participant felt personal filtering was appropriate, they also felt that the comment should not be allowed online generally 70% of the time. In the most extreme case, 30% of participants would *never* remove a comment from an online platform—even for participants that rated at least one comment as "Extremely toxic" (as 10% of that 30% of our participants did). That participants can recognize harassment but decline intervention represents one of the fundamental conflicts between tackling toxic content online and unfettered free speech.

We observe similar, competing perspectives when it comes to who participants feel is the most responsible for addressing

[4]Each set of five participants are guaranteed to rate the same twenty comments in a random order, which enables us to compare kappa values across participants per block of comments.

toxic content online. As part of our pre-exercise questions, we asked participants whether they felt toxic content was a problem and what party was most responsible for addressing toxic posts or comments online. 42% of participants felt toxic content was very frequently or frequently a problem. Another 51% felt it was rarely or occasionally a problem, while 5% felt it was not an issue at all. Additionally, 47% of participants felt the onus of addressing toxic content was on the user who sent the comment, compared to 27% of participants who felt that the hosting platform held the most responsibility. This rift in beliefs—both for toxic content being an issue online, and what party is responsible for solving it—represents a challenge moving forward for tackling harassment online.

5 Competing Perspectives of Toxicity

Given the frequency of disagreement among raters on what constitutes toxic content, we explore potential explanatory variables stemming from a participant's personal experiences, demographics, and opinions on whether toxic content is a societal problem.

5.1 Modeling participant decision making

We treat each rating task per participant as a Bernoulli trial where a rating of "Moderately toxic" or higher indicates the participant found a comment toxic (e.g., a successful event, or 1), and all other ratings as benign (e.g., failure, or 0). We then model the frequency of success across all labeling tasks as a quasi-Binomial distribution $Y_i(n_i, \pi_i, \phi)$ using a logarithmic link function. The model's parameters consist of categorical variables related to a participant's age, gender, political affiliation, religious beliefs, LGBTQ+ affiliation, education, race and ethnicity, and parental status. The model also incorporates whether a participant has previously witnessed toxic content online or personally been the target of toxic content, whether the participant thinks toxic content is an issue, and who is most responsible for addressing toxic content.

Table 5 contains the results of our model. We report the model's weights as the odds that a participant with a specific trait or belief—after holding all other traits constant—will rate a comment randomly drawn from our corpus as toxic. All results noted with an asterisk are statistically significant with $p < 0.01$. While not shown in a table, we repeat the same modeling process to also understand if any factors influence a participant categorizing a toxic comment as any of our five subcategories of toxic content. We report the full parameters of our models in the Appendix. We discuss the results of our full analysis in detail below.

5.2 Influence of personal experiences

Overall, 77% of participants reported having witnessed toxic content while online. This aligns with a prior Pew study of personal experiences with online harassment, which observed

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 & Ethnicity	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*

Table 5: **Demographics, Experiences, and Opinions**—We report the change in likelihood that a participant will flag a random comment as toxic, given a specific trait, in terms of odds. All values noted with an asterisk are significant with $p < 0.01$. See Appendix for model weights and exact significance values.

73% of Americans have observed online harassment [45]. Conversely, 29% of participants in our study reported having been the target of toxic content.[5] Both of these experiences exhibit a statistically significant influence on toxicity ratings. Prior personal experience with being the target of toxic content increases the odds of rating new content as toxic by 1.483 times. These participants potentially empathize with others who might be emotionally harmed by toxic content, and as such, take a stronger stance on what behavior constitutes harassment. Conversely, prior experience with witnessing toxic content decreases the odds of rating new content as toxic by 0.780 times. These participants potentially view

[5] Participants answered both of these questions after the labeling task, which means their answers may have been colored by the perceived toxicity, or lack thereof, of the comments they labeled.

new toxic content through a comparative lens, excusing abusive behavior that does not rise to the level of severity the participant previously encountered. Our findings illustrate the importance of understanding the experience of people who have been targets of harassment as well as highlights the risk of desensitization.

5.3 Influence of demographics

Gender. We find no statistically significant differences between the odds that non-binary, female, and male participants rate a comment as toxic. Furthermore, female and male participants have nearly identical rates for identifying each subcategory of toxic content. One exception is that the odds of a male participant identifying a comment as threatening compared to female participants increases by 1.158 times. One potential explanation is that men report higher rates of physical threats and name calling compared to women [45], and may be more sensitive to those categories of toxic content.

Age. We find that young participants in particular are more likely to flag comments as toxic compared to older participants. Specifically, the odds of rating a comment as toxic by people ages 18–34 increases 1.227–1.238 times compared to participants aged 35–44. When comparing people 35–44 and groups of older adults, we find no statistically significant difference between successive age groups. One possibility is that younger participants may be more represented on the sites we sample from, and thus familiar with the slang or style of attacks present. In line with previous studies [13, 45], participants between the ages of 18–34 also experienced online harassment at higher rates (27%–30% versus 20–24%), which may shape their opinion and sensitivity to toxic content.

LGBTQ+. A participant's LGBTQ+ identity plays a strong role in toxicity ratings. Identifying as LGBTQ+ increases the odds of rating a comment as toxic by 1.644 times compared to participants who do not. Furthermore, LGBTQ+ participants were far more likely to assign all subcategories to toxic comments—with threats showing the largest increase in odds (1.865 times). LGBTQ+ participants are a historically at-risk cohort for online harassment [13] and so may be cognizant of toxic behaviors, biases, and language that other participants fail to identify.

Importance of religion. Religion has one of the strongest influences on how participants perceive toxic content. In particular, religion being "Very important" to a participant increases the odds they rate a comment as toxic by 1.840 times. This impact still holds even when a participant reports that religion is "Not too important", where the odds of rating a comment as toxic increase by 1.216 times. Similarly, religious participants were far more likely assign all subcategories to toxic comments—with profanity and threats showing the largest increase in odds (1.604–1.878 times).

Parents. There is a small but statistically significant difference between the perspectives of parents and non-parents. Being a parent increases the odds of rating a toxic as comment by 1.330 times. Being a parent also increased the odds of flagging sexually harassment (1.298 times) and profanity (1.158 times). These differences are potentially influenced by content that parents do not want their children to see online.

Race and Ethnicity. We find that belonging to a racial or ethnic minority plays only a small role in influencing perspectives of toxic content, amounting to an increase in odds of 1.126 times compared to non-minority participants. Previous studies have shown that minorities and non-minorities experience similar rates of online harassment, but that when harassment occurs, people self-report it is more likely a result of their race or ethnicity [45].

Education and political affiliation. Compared to participants with only a high school education, the odds participants with advanced degrees labeled comments as toxic increases 1.365 times, however, we find no similar relationship to those with college degrees but no advanced degrees. Finally, we find that a participant's political affiliation also has a small impact on the odds of identifying toxic content. Notably, identifying as an independent decreases the odds of flagging toxic content online by 0.901 times compared to liberal participants. These variations may stem from the underlying content and discussions present in our dataset.

5.4 Influence of technology beliefs

Finally, we examine how attitudes towards technology and toxic content online influence toxicity ratings. We find that, when participants feel that toxic content is "Very frequently" a problem, the odds they flag content as toxic increases by 1.125 times compared to others who feel toxic content is "Not a problem". Similarly, when participants feel that technology's role in peoples' lives remains "Very positive", the odds they flag content as toxic increases by 1.142 times compared to neutral participants. These participants potentially have a lower threshold for what they deem to be toxic behavior, or feel a greater obligation to address toxic content.

6 Benchmarking Toxicity Classifiers

Given the influence of personal experiences on toxicity ratings, we next analyze how well widely-deployed automated detection systems from Jigsaw and Instagram currently perform in aggregate, per demographic cohort, and per individual.

6.1 Perspective API

Overall performance. As previously discussed, our dataset uses stratified sampling to oversample potentially toxic comments with the highest rates of disagreement among

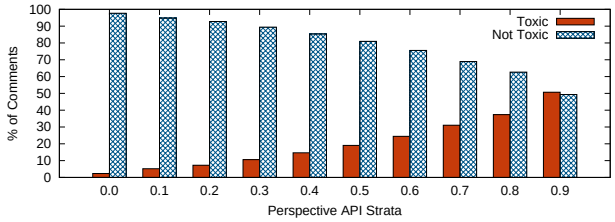

Figure 4: **Toxic/Benign Comment Distribution per Perspective API Stride**—Higher Perspective API scores correlate with a larger fraction of toxic content, however, the fraction of toxic content per stride never exceeds the fraction of benign content.

participants. We omit the vast majority of benign content on Twitter, Reddit, and 4chan that would otherwise be present in a random sample. As such, it is misleading to compare standard performance metrics (e.g., accuracy, precision-recall) across our entire dataset. We control for this bias by considering the accuracy of the Perspective API per *stride* of our sampling. As part of this, we convert every comment's rating distribution into a binary verdict. We treat every comment with a median Likert score of "Moderately toxic" or higher as toxic and all other comments as benign. To compute accuracy, we deem a perspective score of > 0.75 as toxic and all other comments as benign.

Figure 4 shows the fraction of toxic and benign content at each stride of our dataset. The 0.1 stride includes all comments the Perspective API gave a 0–10% likelihood of being toxic, whereas the 0.9 stride includes all comments with a 90-100% likelihood of being toxic. While higher Perspective API scores have monotonically increasing degrees of perceived toxicity, the fraction of toxic content per stride is almost always smaller than fraction of benign content, with the exception of the highest stride, where the labels are roughly equal. Overall, we find only a weak correlation between our participant's Likert ratings and the Perspective API ($r = 0.39$, $p = 0.0$). In line with this, the accuracy for comments in the highest Perspective API stride is only 51%, indicating our participants disagreed with the Perspective rating in 49% of cases. As such, it appears that the Perspective API favors false positives over false negatives. Such a balance is better suited for re-ranking or informing moderation decisions as opposed to outright filtering.

Identifying divergent comments. To provide qualitative examples of divergence between participants and the Perspective API, we isolated comments with the largest intensity of disagreement between the Perspective API and our participants to understand the underlying cause. This is best captured by a root mean square error (RMSE). In particular, we first calculated the average rating across all participants for each comment which we then normalized to a 0–1 scale. We relied on the original 0–1 score of the Perspective API for comparison. We find that error is smallest for lower scores of the

Stride	% Comments	Accuracy	RMSE
0.0	5.0	0.98	0.12
0.1	5.0	0.95	0.14
0.2	5.0	0.93	0.18
0.3	10.0	0.90	0.24
0.4	10.0	0.85	0.30
0.5	15.0	0.81	0.36
0.6	15.0	0.76	0.42
0.7	20.0	0.50	0.48
0.8	10.0	0.37	0.55
0.9	5.0	0.51	0.55

Table 6: **Accuracy and Root Mean Squared Error per Perspective API Stride**—As Perspective API scores increase, we observe a larger divergence between what participants construe as toxic content versus the classifier. For comments that Perspective rates as 90% likely to be toxic, our participants agreed only 45% of the time.

Perspective API, indicating both the classifier and our participants agree on what is not toxic. However, error increases as strides increase, with error maximizing at the highest stride ($>= 0.9$) as shown in Table 6. This result is consistent with our observations with accuracy and indicates that the API struggles to match ground truth at high decision thresholds.

We then selected the top 200 comments with the highest error between participants and the Perspective API. Of these, we selected 100 where the Perspective API skewed higher, and 100 where the Perspective API skewed lower. Two independent coders reviewed each comment, came together to form a codebook of high-level themes, and then independently rated every comment before resolving any disagreement. Examples comments where Perspective and our participants disagreed on toxicity ratings can be found in the Appendix, Table 8 and Table 9.

In some cases, the Perspective API is more sensitive to toxic content than our participants. The majority of these comments included some form of profanity that participants did not find toxic (82%). For example, Perspective rates the comment "holy shit, you are perfect!" at 0.83 while every rater found the comment to be not at all toxic. Other sources of disagreement include differing sensitivities to negativity (6%), and scenarios where negative terms convey emphasis (4%). Examples include "Song sucks" and "The stupid tongue! :D Cutest thing" respectively. In effect, the classifier is incorrectly shaping its decision based on negative or profane keywords alone.

In contrast, our participants also occasionally deemed content more toxic than the Perspective API. Identity attacks against minorities, women, and the LGBTQ+ community made up 39% of comments missed by the Perspective API. Another 17% of comments contained political messages that participants deemed toxic. For example, participants rated the comment "YOU'RE JUST LIKE YOUR BOSS AND THE REST OF THE GOP A BUNCH OF!!!!! #CLOWNS" with

a toxicity of 0.75, compared to 0.48 from the Perspective API. Other themes included adult content (11%) and threats of violence or rape (9%). Additionally, despite Perspective API regularly flagging profanity, minor grammatical changes such as the lack of spaces in "nofuckingbody" resulted in score of 0.21, whereas adding spaces results in a score of 0.93. Researchers have abused this sensitivity to minor perturbations in text to construct adversarial examples that evade the Perspective API [25, 29].

Tuning classifiers to personal preferences. Our results indicate that a single definition of toxic content online does not capture the varied experiences and opinions of Internet users. As such, a one-size-fits-all model for abuse detection will likely not be able to capture the toxicity preferences of all participants. Recent work from Google Jigsaw has focused on allowing participants to "Tune" existing APIs to their own personal preferences, simply by adjusting the Perspective API to a specific threshold [38]. However, it is unclear how effective this tuning strategy can be to end-users and where this mechanism may fall short. We investigate the differences in accuracy and precision for the optimal threshold for each individual participant compared to the dataset in aggregate. Although our dataset is not a truly random sample of Internet comments, comparing personal thresholds to the aggregate still provides insight into the effectiveness of personal tuning.

To identify the optimal threshold for all ratings taken in aggregate, we first convert each comment rating into a binary label. A comment rating has a positive label if the participant personally did want to see the comment online, and a negative label if they did not. We then sweep over all Perspective API decision thresholds from 0–1 and identify the *lowest* threshold that maximizes the F1-score, which is the weighted average of the precision and recall. We find that the optimal perspective API threshold for the aggregate dataset ranges from 0.18–0.49, all of which achieve a precision of 0.35 and an accuracy of 0.37.

We perform the same analysis on an individual level, identifying a threshold that maximizes the F1 score for each participant. If a participant did not personally elect to remove any comments they encountered, we set their threshold to the maximum possible value (1.0). Figure 5 shows a distribution of thresholds per individual. For 21.6% of participants, their maximal threshold is 0.0, suggesting that labeling every comment as toxic maximizes both precision and recall. The median threshold is 0.61, resulting in an average precision of 0.6 and an average accuracy of 0.68, an increase in accuracy of 86% compared to a one-size-fits-all classifier. Per this personalized approach, 71.5% of participants saw an improvement in accuracy over the one-size-fits-all optimum accuracy. As such, more research is needed to understand how best to quickly personalize models and how to gather ongoing feedback in order to adjust model thresholds.

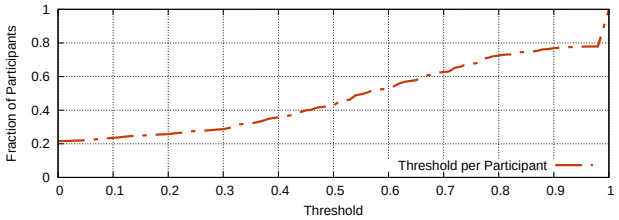

Figure 5: **Optimum Personalized Threshold per Participant**—The threshold that maximizes classifier accuracy per participant is mixed. 21.6% of participants are maximized at a threshold of 0.0, which amounts to labeling every comment as toxic. After tuning to personal thresholds, 71.5% of participants achieved an accuracy greater than the overall classifier.

	Max Precision		Max Accuracy	
Demographic	**Value**	**% Change**	**Value**	**% Change**
Religion	0.40	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%

Table 7: **Optimum Accuracy and Precision per Demographic**—We show the maximum accuracy and precision when tuning the Perspective API per demographic cohort, as well as the percentage change from the one-size-fits-all model. We find that cohort-based models perform marginally better in some categories, but fall short of performance improvement from personalized models.

Tuning classifiers to demographic preferences. Given differences between demographic cohorts (Section 5), we also investigate the performance benefits for tuning the Perspective model to broad demographic groups. Table 7 shows the maximum precision and accuracy when each independent demographic group is tuned for separately. We find that demographic tuning in aggregate offers a smaller improvement over the aggregate classifier compared to personalized tuning, with only a 0–20.6% increase in accuracy. Age-specific model thresholds provided the best performance gain. These results highlight that even within broad demographic groups, individual experiences and preferences take more importance when making toxicity determinations online. Any cohort-based model would need to account for multiple factors when designed and deployed.

6.2 Instagram nudges

In December 2019, Instagram rolled out a feature that nudges a user if they are about to post a comment similar to those that have been flagged in the past. As a small experiment, we also compare how well the Instagram classifier performs against

our ground truth data. We first sampled 200 comments—150 of the most egregious "toxic" comments which have a median toxicity rating of "Very toxic" or higher, and 50 "benign" comments that have a median toxicity rating less than "Slightly toxic". We then manually posted these comments to an Instagram account we controlled (with no audience), noting which comments triggered their classifier.

Of the toxic comments, just 41 (27%) triggered the Instagram classifier. These were mostly identity-based attacks (47%), followed by a mix of adult content (15%), profanity (7%), and threats (3%). Two expert raters attempted to label each comment, but we found no unifying themes that might explain why some toxic comments did not trigger detection. Categories reported by our participants for our toxic sample included insults (26%), profanity (22%), and identity attacks (21%). The classifier never triggered on a benign comment. As such, a significant gap remains in the classifier's ability to detect a wide variety of toxic comments.

7 Discussion

Based on our findings, we discuss potential best practices, pitfalls, and paths forward for improving toxic content classifiers to better serve a diversity of perspectives.

Best practices for crowdsourced labeling. During the development of our survey instrument, we were unable to identify any best practices for developing crowdsourcing instruments that gather toxic content ratings. Previous studies used disparate terminology including "abusive", "hateful", "offensive", and "toxic". For sublabeling tasks that involve categorizing toxic content into sexual harassment, identity-based attacks, or insults, we were unable to find terminology or a taxonomy that was evaluated for rater comprehension. Our experiments show that participants solicited from Mechanical Turk in the United States best understood the meaning of "toxic" compared to other terms, and that participants can identify at least five separate categories of toxic content. Given frequent rating disagreement between participants, we also found that five ratings per comment resulted in the best balance between minimizing crowdsourcing costs and achieving a high degree of accuracy. This rating methodology can serve as a future best practice when crowdsourcing labels for toxic content. Furthermore, our results are limited to participants solicited from Mechanical Turk, and should be validated with participants from other crowdsourced platforms.

Towards personalized definitions of toxicity. Our results suggest that personalized tuning of one-size-fits-all models greatly improves the accuracy per user compared to setting a global threshold for all users. In particular, we found that per-user models increased the accuracy of decisions by 86%. These results suggest the feasibility of relying on a general audience for training labels that users then tune to their personal preferences. However, increasing classifier performance beyond this point will remain a challenge without incorporating specific user feedback and examples. An intermediate approach, where models generalize to specific single-trait demographic cohorts rather than individuals, resulted in only a 0–20.6% improvement in accuracy, with age-specific models performing the best. In the absence of personalization or user feedback, platforms might consider increasingly sophisticated, community-based filters that take into account more than just one demographic trait.

Measuring toxicity using existing classifiers. Recent studies in toxic content have begun to leverage toxicity classifiers as a tool for measuring the prevalence of hate and harassment online, with additional post-processing via rater agreement [20, 30, 31]. Given the variations in classifier accuracy across demographic cohorts and types of sites, we caution against off-the-shelf usage of current classifiers without such post-processing or additional calibration. Even at a Perspective toxicity threshold of 0.9 or higher, our participants disagreed with the classifier's verdict in 50% of cases for the sites we measured.

Online Context. Our work does not incorporate the context that a comment is presented in. As such, it may be challenging for a participant to pinpoint if a comment is toxic versus simply sarcastic or joking. We selected this because toxicity detection systems classify text without additional context, and we wanted to evaluate them based on their current usage. Furthermore, users may have different responses to toxic content when they see such context in-situ (e.g., the toxic content may be targeted at an acquaintance). Some areas of future work include understanding how perspectives change if participants are provided with additional context when labeling, identifying if classifiers can be improved by adding context during training, and measuring participant responses to toxic content in-situ of browsing.

8 Conclusion

In this work, we built and deployed a survey instrument to 17,280 participants across the United States and asked them about their perspectives on toxic content online. We found that a participant's attitudes towards filtering toxic content varies across a multitude of factors: their demographic background, their personal experiences with harassment, and even their attitudes towards technology and the state of toxic content online. Given these influences, we showed how personalized tuning of independent thresholds for existing classifiers can improve the accuracy of toxic detection performance by 86% on average, pointing to personalized models as a future area of research in toxic content detection. We have released our labeled toxicity dataset to enable future work in this space and hope that our work presents paths forward for improving toxic content classification for a diverse set of users.

Acknowledgments

The material is based upon work supported by the National Science Foundation under grant #2030859 to the Computing Research Association for the CIFellows Project and through gifts from Google. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of their employers or the sponsors.

References

[1] Amnesty International. Twitter still failing women over online violence and abuse. https://www.amnesty.org/en/latest/news/2020/09/twitter-failing-women-over-online-violence-and-abuse/, 2020.

[2] Anti-Defamation League. Quantifying hate: A year of anti-semitism on twitter. https://www.adl.org/resources/reports/quantifying-hate-a-year-of-anti-semitism-on-twitter#methodology.

[3] R. Arthur. We analyzed more than 1 million comments on 4chan. hate speech there has spiked by 40% since 2015. https://www.vice.com/en_us/article/d3nbzy/we-analyzed-more-than-1-million-comments-on-4chan-hate-speech-there-has-spiked-by-40-since-2015.

[4] J. Blackburn and H. Kwak. Stfu noob!: predicting crowdsourced decisions on toxic behavior in online games. In *Proceedings of the 23rd international conference on World wide web*. ACM, 2014.

[5] L. Blackwell, J. Dimond, S. Schoenebeck, and C. Lampe. Classification and its consequences for online harassment: Design insights from heartmob. In *Proceedings of the ACM on Human-Computer Interaction*, 2017.

[6] L. Blackwell, J. Hardy, T. Ammari, T. Veinot, C. Lampe, and S. Schoenebeck. Lgbt parents and social media: Advocacy, privacy, and disclosure during shifting social movements. In *ACM CHI conference on human factors in computing systems*, 2016.

[7] A. Breland. Why reddit is losing its battle with online hate. https://www.motherjones.com/politics/2019/08/reddit-hate-content-moderation/.

[8] E. Chandrasekharan, M. Samory, A. Srinivasan, and E. Gilbert. The bag of communities: identifying abusive behavior online with preexisting internet data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017.

[9] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali. Mean birds: Detecting aggression and bullying on Twitter. In *ACM Web Science Conference*, 2017.

[10] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec. Antisocial behavior in online discussion communities. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[11] D. K. Citron. Addressing cyber harassment: An overview of hate crimes in cyberspace. *Journal of Law, Technology & the Internet*, 2014.

[12] G. Cowan and J. Mettrick. The effects of target variables and settting on perceptions of hate speech. *Journal of Applied Social Psychology*, 2002.

[13] Data & Society. Online harassment, digital abuse, and cyberstalking in america. https://datasociety.net/output/online-harassment-digital-abuse-cyberstalking/, 2016.

[14] T. Davidson, D. Bhattacharya, and I. Weber. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, 2019.

[15] T. Davidson, D. Warmsley, M. Macy, and I. Weber. Automated hate speech detection and the problem of offensive language. In *AAAI International Conference On Web and Social Media*, 2017.

[16] D. DiFranzo, S. H. Taylor, F. Kazerooni, O. D. Wherry, and N. N. Bazarova. Upstanding by design: Bystander intervention in cyberbullying. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018.

[17] K. Dinakar, R. Reichart, and H. Lieberman. Modeling the detection of textual cyberbullying. In *AAAI International Conference On Web and Social Media*, 2011.

[18] L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018.

[19] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati. Hate speech detection with comment embeddings. In *The Web Conference*, 2015.

[20] M. ElSherief, S. Nilizadeh, D. Nguyen, G. Vigna, and E. Belding. Peer to peer hate: Hate speech instigators and their targets. In *AAAI International Conference On Web and Social Media*, 2018.

[21] J. Finkelstein, S. Zannettou, B. Bradlyn, and J. Blackburn. A quantitative approach to understanding online antisemitism. In *Proceedings of the AAAI International Conference on Web and Social Media*, 2020.

[22] A. M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. In *12th International AAAI Conference on Web and Social Media*, 2018.

[23] Google Jigsaw. Perspective api. https://www.perspectiveapi.com/#/home.

[24] M. L. Gordon, K. Zhou, K. Patel, T. Hashimoto, and M. S. Bernstein. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *ACM CHI Conferences on Human Factors in Computing Systems*, 2021.

[25] T. Gröndahl, L. Pajola, M. Juuti, M. Conti, and N. Asokan. All you need is "love" evading hate speech detection. In *Proceedings of the ACM Workshop on Artificial Intelligence and Security*, 2018.

[26] A. M. G. Gualdo, S. C. Hunter, K. Durkin, P. Arnaiz, and J. J. Maquilón. The emotional impact of cyberbullying: Differences in perceptions and experiences as a function of role. *Computers & Education*, 2015.

[27] Hatebase. The world's largest structured repository of regionalized, multilingual hate speech. https://hatebase.org/, 2019.

[28] D. Hicks and D. Gasca. A healthier twitter: Progress and more to do. https://blog.twitter.com/en_us/topics/company/2019/health-update.html.

[29] H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran. Deceiving google's perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138*, 2017.

[30] Y. Hua, M. Naaman, and T. Ristenpart. Characterizing twitter users who engage in adversarial interactions against political candidates. In *ACM CHI Conference on Human Factors in Computing Systems*, 2020.

[31] Y. Hua, T. Ristenpart, and M. Naaman. Towards measuring adversarial twitter interactions against candidates in the us midterm elections. In *International Conference on Web and Social Media*, 2020.

[32] Instagram. Our progress on leading the fight against online bullying. https://instagram-press.com/blog/2019/12/16/our-progress-on-leading-the-fight-against-online-bullying/.

[33] E. Jain, S. Brown, J. Chen, E. Neaton, M. Baidas, Z. Dong, H. Gu, and N. S. Artan. Adversarial text generation for google's perspective api. In *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2018.

[34] S. Jhaver, S. Ghoshal, A. Bruckman, and E. Gilbert. Online harassment and content moderation: The case of blocklists. In *Proceedings of the ACM Transactions on Computer-Human Interaction*, 2018.

[35] Jigsaw. Toxic comment classification challenge. https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge, 2017.

[36] H. Kwak, J. Blackburn, and S. Han. Exploring cyberbullying and other toxic behavior in team competition online games. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2015.

[37] J. R. Landis and G. G. Koch. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 1977.

[38] D. Lee. Alphabet-made chrome extension is designed to tune out toxic comments. https://www.theverge.com/2019/3/14/18265851/alphabet-google-jigsaw-tune-chrome-extension.

[39] K. Mahar, D. Karger, and A. X. Zhang. Squadbox: A tool to combat online harassment using friendsourced moderation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018.

[40] B. Maher. Can a video game company tame toxic behaviour? https://www.nature.com/news/can-a-video-game-company-tame-toxic-behaviour-1.19647.

[41] S. Melendez. Twitter automatically flags more than half of all tweets that violate its rules. https://www.fastcompany.com/90528941/twitter-automatically-flags-more-than-half-of-all-tweets-that-violate-its-rules?utm_source=morning_brew.

[42] C. Newton. The trauma floor. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona.

[43] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. Abusive language detection in online user content. In *The Web Conference*, 2016.

[44] J. A. Pater, M. K. Kim, E. D. Mynatt, and C. Fiesler. Characterizations of online harassment: Comparing policies across social media platforms. In *Proceedings of the 19th International Conference on Supporting Group Work*, 2016.

[45] Pew Research Center. Online harassment 2017. https://www.pewinternet.org/2017/07/11/online-harassment-2017/, 2017.

[46] D. Razo and S. Kübler. Investigating sampling bias in abusive language detection. In *4th Workshop on Online Abuse and Harms*, 2020.

[47] E. M. Redmiles, J. Bodford, and L. Blackwell. "i just want to feel safe": A diary study of safety perceptions on social media. In *International AAAI Conference on Web and Social Media*, 2019.

[48] E. M. Redmiles, S. Kross, and M. L. Mazurek. How well do my results generalize? comparing security and privacy survey results from mturk, web, and telephone samples. In *Proceedings of the IEEE Symposium on Security and Privacy*, 2019.

[49] E. M. Redmiles, Z. Zhu, S. Kross, D. Kuchhal, T. Dumitras, and M. L. Mazurek. Asking for a friend: Evaluating response biases in security user studies. In *25th ACM SIGSAC Conference on Computer and Communications Security*, 2018.

[50] N. Sambasivan, A. Batool, N. Ahmed, T. Matthews, K. Thomas, L. S. Gaytán-Lugo, D. Nemer, E. Bursztein, E. Churchill, and S. Consolvo. "they don't leave us alone anywhere we go": Gender and digital abuse in south asia. In *Proceedings of the Conference on Human Factors in Computing Systems*, 2019.

[51] A. Saravanaraj, J. Sheeba, and S. P. Devaneyan. Automatic detection of Cyberbullying from Twitter. *International Journal of Computer Science and Information Technology & Security (IJCSITS)*, 2016.

[52] K. Thomas, D. Akhawe, M. Bailey, D. Boneh, E. Bursztein, S. Consolvo, N. Dell, Z. Durumeric, P. G. Kelley, D. Kumar, D. McCoy, S. Meiklejohn, T. Ristenpart, and G. Stringhini. Sok: Hate, harassment, and the changing landscape of online abuse. In *Proceedings of the IEEE Symposium on Security and Privacy*, 2021.

[53] US Census. United states census bureau. https://www.census.gov/data.html, 2021.

[54] C. Van Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. De Pauw, W. Daelemans, and V. Hoste. Automatic detection and prevention of cyberbullying. In *International Conference on Human and Social Analytics*, 2015.

[55] W. Warner and J. Hirschberg. Detecting hate speech on the World Wide Web. In *Proceedings of the second workshop on language in social media*, 2012.

[56] M. Wich, H. Al Kuwatly, and G. Groh. Investigating annotator bias with a graph-based approach. In *4th Workshop on Online Abuse and Harms*, 2020.

[57] E. Wulczyn, N. Thain, and L. Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, 2017.

[58] S. Zannettou, T. Caulfield, J. Blackburn, E. De Cristofaro, M. Sirivianos, G. Stringhini, and G. Suarez-Tangil. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*. ACM, 2018.

Appendix—Survey instrument

Initial consent form and university contact information
I understand the consent form. I certify that I am 18 years old or older. By clicking the "Yes" button to enter the survey, I indicate my willingness to voluntarily take part in this study.
o Yes
o No

Pre-exercise questions
Today we would like you to help us with this short survey and to review some comments that have been posted online.

What types of sites do you use? [Checkbox]
o Social Networking (Facebook, Twitter)
o Video (YouTube, Twitch)
o News (CNN, Fox, NYT, WSJ)
o Community Forums (Reddit, Craigslist, 4chan)
o Email or messaging (Gmail, WhatsApp, Facebook Chat)

In general, how would you rate the impact of technology on people's lives? [Radio]
o Very positive
o Somewhat positive
o Neutral – neither positive nor negative
o Somewhat negative
o Very negative

Based on your experience, toxic posts or comments are ... [Radio]
o Not a problem
o Rarely a problem
o Occasionally a problem
o Frequently a problem
o Very frequently a problem

Who is the most responsible for addressing toxic posts or comments? [Rank, 1-5]
o Law enforcement
o Users who observe the comment
o The user who receives the comment
o The website hosting the comment
o The user who sent the comment

What does the term "toxic comment" mean to you?
o Open Ended

Labeling exercise
In this section, you will be shown twenty comments taken from various online platforms. Some of these comments may contain toxic content. For each comment, you will be asked a series of questions regarding your perspective on the comment.

[Random comment displayed from Reddit, Twitter, or 4chan. This comment remains visible throughout the subquestions below.]

How toxic do you personally find this comment? [Radio]
o Extremely toxic
o Very toxic
o Moderately toxic
o Slightly toxic
o Not at all toxic

[If any but not at all] Please select all the reasons why below: [Checkmark]
o It is threatening (ex. "If you comment on my feed again, I will personally beat you up")
o It is insulting (ex. "You are an absolutely disgusting hypocrite")
o It is profane (ex. "Why the f*** not you heathen?")
o It is an attack on a person or group based on their identity (ex. "Women are stupid and belong in the kitchen")
o It is sexual harassment (ex. "All I want from you are nudes")
o Other, please specify

Is this a type of comment you would want to see? [Radio]
o I would never want to see it
o It depends on the context

o This is fine for me to see

Should this comment be allowed online? [Radio]
o This comment should be removed
o It depends on the context
o This comment should be allowed

Post-exercise questions
For the sites you use, have you ever seen comments similar to the ones we showed you? [Radio]
o Yes
o No

Have you ever personally been the target of comments similar to the ones you reviewed? [Radio]
o Yes
o No

Is there anything else you would like to tell us about toxic comments?
o [Open ended]

Demographic questions
Which category below includes your age? [Radio]
o Under 18
o 18 - 24
o 25 - 34
o 35 - 44
o 45 - 54
o 55 - 64
o 65 or older
o Prefer not to say

Race [Checkbox]
o White
o Hispanic or Latino
o Black or African American
o Native American or American Indian
o Asian / Pacific Islander
o Other [open ended]
o Prefer not to say

What is your gender? [Radio]
o Female
o Male
o Nonbinary
o Prefer not to say
o Other [Open ended]

Would you describe yourself as transgender? [Radio]
o Yes
o No
o Prefer not to say

What is the highest degree or level of school that you have completed? [Radio]
o Less than high school degree
o High school graduate (high school diploma or equivalent including GED)
o Some college but no degree
o Associate degree in college (2-year)
o Bachelor's degree in college (4-year)
o Master's degree
o Doctoral degree
o Professional degree (JD, MD)
o Prefer not to say
o Other [Open ended]

Do you consider yourself to be: [Radio]
o Heterosexual or straight
o Homosexual
o Bisexual

o Prefer not to say
o Other [Open ended]

How important is religion in your life? [Radio]
o Not important
o Not too important
o Somewhat important
o Very important
o Prefer not to say

Generally speaking, do you consider yourself politically conservative, liberal, independent, or something else? [Radio]
o Conservative
o Liberal
o Independent
o Other [Open ended]
o Prefer not to say

Are you a parent? [Radio]
o Yes
o No
o Prefer not to say

What term have we been using to describe comments in this survey? [Radio]
o Mean
o Abusive
o Bullying
o Toxic

Appendix—Divergent Comments

We show examples of divergence between participant scores and the Perspective API score in Table 8 and Table 9.

Category	Sample	% Comments
Profanity	"Damn cultural colloquialisms being all complicated and stuff"	82%
Negative sentiment	"Song sucks. But I vibe with how he mixed 3 of Lil Wayne videos into one."	6%
Inverted sentiment	"The stupid tongue! :D Cutest thing"	4%
Adult	"penis size shrinking in males exposed to contaminated environments."	2%
Other		6%

Table 8: **High Perspective API Score, Low Groundtruth Score**–The Perspective API over-emphasizes profanity, which participants rarely found toxic.

Category	Sample	% Comments
Identity attack	"We need to actively fight against racism by only targeting white people and letting coloured people get away with crimes"	39%
Political	"And you're getting closer to being locked up behind bars #pediphile #RapistTrump"	17%
Adult	"Hello World! Wanna Hookup; Trade Nude Pics? Visit ...	11%
Threats	"Store them in an unventilated room with hoses that run between the room and your car's exhaust pipe. That'll solve your problem."	9%
Profanity	"the same person that invited you? nofuckingbody	3%
Negative sentiment"	"Sounds like you're a no one who's gonna die bitter and alone and forgotten"	3%
Other		18%

Table 9: **Low Perspective API Score, High Groundtruth Score**–The Perspective API currently misses nuanced identity-based attacks and political attacks that participants deemed toxic.

Appendix—Regression Tables

We present the full parameters and outputs for our quasi-binomial model that predicts the likelihood of rating a random comment as toxic (1 or 0). We show the independent treatment group (either binary or categorical depending on the variable), the reference group, the model coefficient (β), error (*SE*), z-score (z), p-value, and the resultant odds ratio (*OR*). Due to space restraints, we do not present full model results for each individual sublabel model (i.e., whether participant would rate a random comment as an insult, an identity attack, a threat, as profane, or as sexual harassment), and instead direct the reader to the extended version of the paper available at https://arxiv.org/abs/2106.04511.

Demographic	**Treatment**	**Reference**	β	*SE*	z	$Pr(>\|z\|)$	*OR*
Gender	Female	Male	-0.049	0.015	-3.250	0.001	0.952
Gender	Nonbinary	Male	-0.347	0.116	-2.986	0.003	0.707
Age	65 or older	35 - 44	-0.024	0.042	-0.562	0.574	0.977
Age	18 - 24	35 - 44	0.213	0.028	7.488	0.000	1.238
Age	25 - 34	35 - 44	0.204	0.019	10.817	0.000	1.227
Age	55 - 64	35 - 44	-0.020	0.030	-0.665	0.506	0.980
Age	45 - 54	35 - 44	-0.029	0.025	-1.167	0.243	0.972
Race	Minority	Non-minority	0.119	0.016	7.277	0.000	1.126
LGBTQ+	LGBTQ+	Not LGBTQ+	0.497	0.020	25.225	0.000	1.644
Political affiliation	Independent	Liberal	-0.104	0.018	-5.758	0.000	0.901
Political affiliation	Conservative	Liberal	0.024	0.018	1.308	0.191	1.024
Religion	Not too important	Not Important	0.195	0.026	7.617	0.000	1.216
Religion	Somewhat important	Not Important	0.453	0.021	21.947	0.000	1.572
Religion	Very important	Not Important	0.610	0.020	30.177	0.000	1.840
Parent	Yes	No	0.285	0.016	17.360	0.000	1.330
Education	College	High school	0.130	0.026	4.945	0.000	1.139
Education	Advanced degree	High school	0.311	0.030	10.325	0.000	1.365
Impact of Technology	Very negative	Neutral	-0.220	0.080	-2.752	0.006	0.803
Impact of Technology	Somewhat negative	Neutral	-0.140	0.032	-4.357	0.000	0.870
Impact of Technology	Somewhat positive	Neutral	-0.032	0.023	-1.402	0.161	0.968
Impact of Technology	Very positive	Neutral	0.133	0.025	5.318	0.000	1.142
Toxic Content a Problem?	Rarely a problem	Not a problem	0.029	0.034	0.863	0.388	1.030
Toxic Content a Problem?	Occasionally a problem	Not a problem	-0.043	0.032	-1.314	0.189	0.958
Toxic Content a Problem?	Frequently a problem	Not a problem	0.028	0.033	0.848	0.397	1.029
Toxic Content a Problem?	Very frequently a problem	Not a problem	0.117	0.037	3.188	0.001	1.125
Party most responsible	Law Enforcement	Bystander	0.248	0.035	7.093	0.000	1.282
Party most responsible	User who Receives	Bystander	-0.334	0.032	-10.427	0.000	0.716
Party most responsible	Hosting Platform	Bystander	-0.348	0.028	-12.481	0.000	0.706
Party most responsible	User who sent the comment	Bystander	-0.480	0.027	-17.973	0.000	0.619
Witnessed Toxic Content	True	False	-0.249	0.018	-14.208	0.000	0.779
Experienced Toxic Content	True	False	0.394	0.017	23.547	0.000	1.482

Table 10: **Toxicity Model**—Logistic regression showing the likelihood a participant will flag a random comment as toxic.