A Human-centered Evaluation of a Toxicity Detection API: Testing Transferability and Unpacking Latent Attributes

MEENA DEVII MURALIKUMAR, YUN SHAN YANG, and DAVID W. MCDONALD,
Human-centered Design & Engineering, University of Washington

Perspective is a publicly available, machine learning API that can score text for toxicity. It is available for use in online platforms and communities to limit toxicity and promote civil dialogue. In this work, we adopt a human-centered approach to evaluating Perspective by investigating if human ratings of toxicity align with Perspective's toxicity scores. We also test its transferability by making this comparison for comments from three platforms that have different commenting styles and moderation strategies: news websites, YouTube, and Twitter. Apart from toxicity, the main attribute, we collect participant ratings for three additional attributes: respectfulness, formality, and presence of stereotypes. While disrespect is part of how Perspective defines toxicity, formality and presence of stereotypes were included in the study to explore if they could be hidden/latent attributes that affect toxicity scores from Perspective. We analyzed how participant ratings for these additional attributes vary with respect to Perspective's toxicity score for comments from each platform. We find that for high toxicity scores, Perspective strongly aligns with participant ratings of toxicity and disrespectfulness across all three platforms, providing weak evidence of its transferability. However, our evaluation also surfaced formality and presence of stereotypes as latent attributes that are unrecognized parts of Perspective's scores. We discuss how and why this evaluation is "human-centered," the importance of conducting such evaluations, and implications of these results for content moderation in social platforms.

CCS Concepts: • **Human-centered computing** → **Empirical studies in collaborative and social computing**; • **Computing methodologies** → Machine learning; Artificial intelligence;

Additional Key Words and Phrases: Evaluation, transferability, moderation tools, design, machine learning, Perspective

ACM Reference format:
Meena Devii Muralikumar, Yun Shan Yang, and David W. McDonald. 2023. A Human-centered Evaluation of a Toxicity Detection API: Testing Transferability and Unpacking Latent Attributes. *ACM Trans. Soc. Comput.* 6, 1–2, Article 4 (June 2023), 38 pages.
https://doi.org/10.1145/3582568

1 INTRODUCTION

One promise of social computing is to have meaningful interactions online and connect with diverse people. But as online social interactions have become more common, abusive and hateful speech has also become prevalent. A survey conducted by the Pew Research Center in 2020 on

Authors' addresses: M. D. Muralikumar, Y. S. Yang, and D. W. McDonald, The University of Washington's Department of Human Centered Design & Engineering (HCDE), 3960 Benton Lane NE, 428 Sieg Building, University of Washington Seattle, WA 98195; emails: {mmeena, yunshan, dwmc}@uw.edu.

2469-7818/2023/06-ART4 $15.00
https://doi.org/10.1145/3582568

the state of online harassment shows that 91% of Americans consider online harassment a problem and roughly 41% of Americans have been subject to some form of online harassment, with most of these occurring on social media [22]. The increasing volume of social media interactions poses challenges for timely, comprehensive monitoring and content moderation. To address this challenge and ensure user trust, organizations are building and deploying automated and semi-automated approaches to content moderation that rely on Artificial Intelligence (AI) and Machine Learning (ML) techniques [9].

However, such AI/ML-based techniques are far from perfect, as they can misclassify items and cannot always account for context [16, 24]. More work is required to understand the advantages and pitfalls of applying such techniques. In this work, we evaluate Perspective, a public, ML-based Application Programming Interface (API) that can identify toxic text [17]. We chose Perspective as it is free and open sourced and is already being used in various real-world applications [18]. Perspective defines a toxic comment as, "a rude, disrespectful, unreasonable comment that is likely to make someone leave a discussion" [4] and returns a score that indicates the probability of the comment being considered toxic. Our evaluation of Perspective takes a human-centered stance to understand how people might accept the outputs of the API. Our high-level research question is as follows:

How do human ratings of comment toxicity align with the probability of toxicity generated by Perspective?

Perspective was developed by training on data from different sources [1, 19] and is now available as a web API for application in a wide range of discussion communities. We are interested in understanding how Perspective will perform in different online platforms and communities, which may or may not have contributed to its training data.[1] Therefore our secondary research question is as follows:

Will evaluating data from other online platforms affect how people agree with Perspective's toxicity scores?

To investigate, we evaluated Perspective's performance on comments from three different platform types: news-based websites, YouTube, and Twitter. We chose these platforms because of their distinct commenting styles and different moderation strategies that can lead to different user experiences, frequencies of toxic comments, and possibly even different judgements of what constitutes toxicity. Perspective was trained on data from multiple sources and notably also from Wikipedia Talk Pages and The New York Times (NYT) [1]. The developers of Perspective also conducted a Kaggle competition with data from Civil Comments Platform (a crowdsourced moderation plugin for independent news sites) [19]. We want to note that we cannot be sure if the scores we obtained from Perspective for this study included any model updates from training on the Civil Comments dataset. Nevertheless, we selected news-based websites, because we wanted to include a platform that is similar to one of the source training datasets for Perspective (NYT).

We also aimed to investigate if other textual properties were reflected in Perspective's toxicity score. Thus our participants rated four different attributes for a given comment: formality, respectfulness, presence of stereotypes, and toxicity. We chose respectfulness, as it is reflected in how Perspective defines toxicity. Though formality and presence of stereotypes are not part

[1]We want to note that evaluating Perspective is a moving target, since the model undergoes regular re-training and updates. Data from additional social media sources could be used for further training. The API can also store and use comments submitted to the API for scoring unless specifically mentioned otherwise using input parameters.

of Perspective's output attributes or its definition, we include them to evaluate if they could be hidden or latent attributes with an effect on how Perspective scores toxicity. Therefore, another secondary research question is as follows:

How do human ratings of formality, respectfulness, and presence of stereotypes vary with respect to Perspective's toxicity scores and how do these relationships vary for the platforms—news websites, YouTube, and Twitter?

Our results show that for all three platforms, a high toxicity score from Perspective is a strong predictor of participants rating the comment as toxic and disrespectful. This finding supports Perspective's definition of toxicity and provides weak evidence of its transferability across social media platforms. However, our evaluation also surfaced formality and presence of stereotypes as latent attributes as a high toxicity score from Perspective is also a strong predictor of participants rating the comment as informal and as containing stereotypes. Based on our findings, we discuss our human-centered evaluation of Perspective, implications for automated content moderation, certain limitations, and future directions for research.

Our primary contribution in this work is to conduct a human-centered evaluation of Perspective that is distinctly different from technical evaluations of model accuracy. We consider our evaluation as human centered, because we evaluate if

(1) Perspective's toxicity scores based on a specific definition match human ratings of toxicity based on the same definition
(2) Perspective's scores match with human ratings in different sociotechnical domains (news websites, YouTube, and Twitter)
(3) Additional attributes (formality, respectfulness, presence of stereotypes) that pertain to how online comments are written are construed (or misconstrued) by the model as toxicity

The rest of the article is structured as follows. We review work related to content moderation, evaluation of hate speech detection algorithms, and the issue of transferability in Section 2. In Section 3, we discuss the rationale for our selection of the three platforms: news sites, YouTube, and Twitter. That is followed by details on our process for collecting our comment corpus. In Section 5, we describe the pilot study and the work we did to specifically address the challenges associated with collecting subjective judgments (such as toxicity ratings) from a crowd-sourcing platform. We report on how we conducted the resulting study in Section 6, followed by our findings in Section 7. We close the article with a discussion of our results (Section 8) where we reflect back on Perspective's transferability, our uncovering of potentially latent attributes that may influence Perspective's performance, the design implications of our findings for content moderation, and a brief consideration of some limitations of our findings.

2 RELATED WORK

In this section, we discuss two threads of literature that relate to this type of evaluation. We discuss why detecting and moderating hate speech or toxicity in online spaces is hard and where hate-speech detecting AI/ML algorithms such as Perspective fit. Since our evaluation focuses on Perspective's performance across different online platforms, we also discuss work related to transferability and cross-domain approaches that address hate speech. We also include background information about how Perspective works to provide the reader basic familiarity with the tool.

2.1 Moderation of Hate Speech and AI/ML Approaches to Detection

As online social interactions become more prevalent, so has the problem of online abuse and harassment. Online harassment can manifest in different forms with varying levels of severity

(such as name-calling, threats, trolling, stalking, doxing, hateful or violent speech, cancel culture), resulting in stress, fear, worry, mental trauma, and, in extreme cases, physical violence to its victims [22]. Online abuse and harassment is not a new phenomena, but online platforms still struggle with how to address it. Eight of 10 people think social media companies are doing either a fair or poor job [22]. The coordinated Gamergate hate campaign across different platforms [10], and the existence (and subsequent banning) of hateful subreddits such as r/fatpeoplehate and r/CoonTown [33] show both the prevalence of hateful speech in online spaces and that it can be deliberate, systematic, and large scale, necessitating platforms to take strict action.

Further, content moderation is a moving target that involves constant negotiation of what is problematic content and for who. Hence, it is not always apparent *if, when, and how* to take action. For example, pro-anorexia and pro–eating disorder content faced extensive bans on almost all platforms due to its sensitive nature and the possibility that it encourages self-harm [20]. However, such bans result in loss of social support and resources for information [20] and enforce conformity at the risk of further marginalizing already marginalized groups and invalidating their lived experiences [38]. One-size-fits-all solutions may result in reduced freedoms for both marginalized as well as mainstream topical content. Social media platforms have maintained and updated their community guidelines over the years [47, 57], but it is difficult to consistently and comprehensively decide on a course of action in response to numerous, different cases [57]. The problem of moderating online hate and abuse is by no means straightforward—it was difficult even before throwing AI/ML algorithms in the mix.

However, the sheer volume of online interactions that need to be monitored for potential hate speech and harassment has motivated platforms to implement automated or partially automated approaches, often involving AI/ML algorithms. AI/ML approaches are not without drawbacks, but they are capable of flagging and screening content that is likely problematic based on examples from the past. These technical approaches can protect both the users of the platform as well as the human moderators from being subject to potential abuse. While researchers have acknowledged the potential benefits of using AI/ML approaches, they also raise important questions of transparency [41, 42, 48], justice and fairness [42], and implications for free speech [41]. These valid concerns apply to practices of content moderation in general but are further complicated by the use of AI/ML techniques. Such concerns have warranted and encouraged inspection of content moderation algorithms and analysis of sociotechnical implications of their use. Our human-centered evaluation is one example of how we can begin to unpack some of these complexities as AI/ML techniques are more widely applied to online interactions.

2.2 The Perspective API

Perspective is a machine learning API aimed at improving the quality of online conversations by scoring comments for toxic content [17]. Perspective was developed collaboratively by Jigsaw and Google's Counter Abuse Technology team under a research project called Conversation-AI [4, 6]. Perspective has been used as an underlying mechanism to build tools that give real-time feedback about toxicity to commenters as they type in a comment [15], to filter comments based on a threshold for readers [25], or to prioritise comments that need more attention for moderators [12].

For a given comment, Perspective returns a numeric score indicating the probability of that comment being toxic. If two different comments X and Y are scored as 0.75 and 0.99, respectively, then it does not mean that Y is more toxic than X. These Perspective scores reflect the likelihood that each comment is toxic. The Perspective API guidelines [8] suggest the following thresholds for labeling a comment as toxic: Comments with <= 0.30 probability would be labeled as "not toxic,"

comments with >0.30 and <0.70 probability would be labeled as "hard to say if toxic," and anything with >= 0.7 probability would be "toxic."

Beyond official documenation and guidelines, Rieder and Skop [59] have offered a critical examination of how Perspective was developed. They juxtapose how both openness and platformization are apparent in developing Perspective as a content moderation system.

2.3 Evaluating AI/ML Approaches to Hate Speech Detection

Evaluating AI/ML efficacy for hate speech detection is difficult. This section covers a wide range of evaluation approaches that are based on different techniques. We find that the most common approaches do not rely on new, independent human judgements of performance.

Evaluations of AI/ML algorithms are carried out with different methods and goals. Different methods and approaches can uncover varied strengths and weaknesses. The most common AI/ML evaluation approaches include testing the model performance on a specific sample resulting in metrics such as precision, recall, area under the receiver operating characteristic curve (AUC), and so on. These are standard measures that can give a sense of how well an algorithm performs against data similar to what it was trained on. Perspective has a high AUC, 0.96857 [3], which is a measure of how well it performs on this classification task. Documentation for this tool cautions that Perspective is not meant to replace human moderation [54].

AI/ML algorithms are sometimes also tested against adversarial examples as a means of understanding risks or security vulnerabilities. Adversarial evaluations rely on intentionally created or modified data, designed to cheat or bypass detection. For example, Hosseine et al. show how an early version of Perspective performs when words have been modified to make toxicity detection difficult, while its toxicity can still be understood by humans (e.g., using "idiiot" and "stu.pid") [45]. In a different adversarial evaluation conducted by Gröndahl et al., seven state-of-the-art hate speech detection models broke when a text comment had spaces between words removed or when the one word "love" was added to the text. With the same changes to the text, the original meaning was still understood by humans [43]. As a type of real-world example, the use of lexical variations such as "thynsporation" and "anarexic" by the pro-ED community helped evade detection once the equivalent, unmodified tags were suppressed by Instagram moderators and moderation tools [31]. Calabrese et al. [30] developed a new evaluation technique called "Adversarial Attacks against Abuse" to measure how well a classifier works on dynamically generated examples that cover more categories of toxicity, while Röttger et al. [60] developed a suite of functional tests specifically for hate speech detection to improve diagnosis of the results.

Understanding whether an AI/ML algorithm is enacting a bias requires a special form of evaluation as well. AI/ML algorithms may be enacting a hidden bias that can be hard for the designers to see, because it may be systemic in the data or a function of the algorithmic technique. A systemic bias in this context is a systematic error or assumption that creates unfair and discriminatory predictions [39]. Garg et al. [40] survey and categorise different biases and corresponding evaluation techniques in toxic speech detection. Early versions of Perspective rated comments referencing certain demographic groups such as lesbian, gay, and homosexual as highly toxic because of an imbalanced training set [37]. Post hoc evaluations on how AI/ML models perform across different racial, cultural, demographic, and/or intersectional groups (e.g., References [36, 37, 56]) and developing more nuanced metrics to measure unintended biases [29] are important efforts in support of model fairness and transparency [54].

Arango et al. point out that despite studies reporting state-of-the-art performance in detecting hate speech, the problem is still persistent in online platforms [26], indicating a research-practice gap. A reason for this gap could be that these algorithms do not transfer easily across different datasets. Their evaluation and Gröndahl et al.'s evaluation of seven state-of-the-art algorithms in

detecting hate speech [43] show evidence of poor generalizability and transferability. We explore this aspect of algorithms further, aiming to complement such technical evaluations with human-centered evaluations that embed a sociotechnical perspective.

While we consider Perspective to be a state-of-the-art ML model to detect toxicity in text, we are interested from a human-centered perspective in evaluating how the model behaves across different social media platforms. That is, given different types of online platforms and their differing interactions, how well do human ratings of toxicity for a comment correspond to the probability of toxicity for that comment generated by Perspective? This approach to understanding an AI/ML tool is distinctly different from the common evaluation of these tools against previously labeled data—often the very data that was used to train the model.

While the work by Hoffman et al. [44] was not explicitly focused on hate speech, it helps motivate two aspects of our problem. In their work, Hoffman et al. [44] conducted a human-centered evaluation of an AI/ML tool that focused on labeling "politeness." They found that the tool generated somewhat confused scores of "impolite" and "neutral" text samples, and that "polite" scores agreed with human ratings only at a higher cutoff threshold than recommended by the developers. This work motivated our "human-centered" approach that is focused more on the way humans might see and interpret toxicity. Further, the findings from Hoffman et al. [44] helped us understand that such evaluations can uncover key aspects of how these tools might be more effectively used in practice, *in situ*, rather than as pure technical artifacts that are capable of meeting high technical performance criteria. With our evaluation of Perspective, we intend to understand the challenges of using Perspective and inform its potential usage.

2.4 Transferability—Application of AI/ML across Platforms

The field of AI/ML research recognizes that the models, when incorporated into the real world, encounter messy, complex data. The concept of transferability for AI/ML expresses the challenge of learning in one domain and applying that learning successfully in a different domain.

Perspective was trained using data from different sources that include but are not limited to the Wikipedia Talk pages, NYT comments, and the Civil Comments dataset [1, 19]. The question of how well Perspective can perform on different online platforms that may or may not have contributed training data is both an important and interesting question. Selbst et al. discuss the portability trap that is a "failure to understand how re-purposing algorithmic solutions designed for one social context may be misleading, inaccurate, or otherwise do harm when applied to a different context" [61]. Even within the same domain, fairness concerns and social context may differ. For example, using ML for recidivism prediction when courts in different jurisdictions are likely to have different populations and differing frequencies of types of crimes. Though transfer learning assures some degree of portability, Selbst et al. warn that it may not be enough to capture the variety of differences in the social context between different domains [61].

In the case of detecting hate speech, *platform policies, moderation strategies, and community norms could constitute differences in social context.* Through a content analysis of moderation policies from 15 social media platforms in 2016, Pater et al. found that definitions of harassment and response to harassment were inconsistent across all the platforms [57]. A more recent (2019) content analysis of policies from 11 social media platforms by Jiang et al. also shows similar variability even though the policies have evolved since 2016 [47]. The authors note that decisions to label something as problematic might be based on what affects each platform the most. Variability could also result from bottom-up influence with community-constructed norms having more effect on user behavior [57] and different communities following or prioritizing different norms [34].

Some prior work illustrates the usefulness of cross-domain approaches to addressing moderation challenges [32, 35]. As Chandrasekharan et al. [35] point out, lack of sufficient training data is

a problem when building new classifiers, resulting in the use of data from other sources or use of off-the-shelf classifiers. Chandrasekharan et al. developed a Bag of Communities (BoC) approach, where they leverage existing data from other communities (4chan, Voat, Reddit, MetaFilter) to classify hate speech in a target community [35]. However, since the target community's comments were similar to the comments in abusive communities (such as 4chan's), using an "only abuse BoC" model generated better results. The usefulness of such cross-domain approaches could thus depend on how well the norms of source and target communities align, which in turn requires community-level insight [35]. Chandrasekharan et al. [32] leverage this "cross-community learning" approach to build Crossmod—a moderation system for Reddit, using an ensemble of classifiers trained on data from different source communities.

We have not found prior work that evaluates the transferability of an AI/ML model from the users' perspective. Our evaluation considers whether the Perspective algorithm can predict toxic/hate speech across different social computing platforms. In short, how well does Perspective perform across different social computing platforms? Empirically investigation can help us understand how a given AI/ML algorithm performs in different social contexts, its strengths and weaknesses, and the possible presence of hidden biases. Without this insight, we could fall into "portability traps" [61], violate user trust [24], cause user migration to other spaces and polarization [31], and disproportionately affect marginalized users [28].

3 PLATFORM CHARACTERISTICS AND COMMENTING STYLES

We considered human and Perspective performance across three different social media platforms with hypothetically different styles of commenting. Our premise is that platform characteristics such as different user experience, layout, and structure of threading would contribute to distinct commenting styles. The layout and visual hierarchy of a website shape how users are guided through and how they experience the content. The design reflects an intentional user experience and influences user behaviors on the website, such as posting and commenting. The differing experiences and commenting styles may therefore impact how comments are perceived by users on those platforms. The three social media platforms that we focused on, news websites, YouTube, and Twitter, all leveraged the inverse-pyramid and message board design patterns [62]. However, each platform type instantiated these patterns slightly differently in their user experience design.

News websites such as ABC News, for example, follow the inverse-pyramid writing style across their news article pages. At the top of the page, there is the news article, which is the most important content and it also initiates the subsequent commenting. The article content is followed by supporting details and the commenting section (message board design pattern) placed at the bottom of the page (see Figure 1). Viewing or contributing to news comments often requires scrolling through the higher priority content. The structure of threading within the comment section is also designed to have a clear hierarchy between the comments and any subsequent replies through the use of indentation and other visual cues , thus shaping the user's commenting behavior (see Figure 2). On a news website, the user journey of reading, commenting, and replying is structured around a specific, focal, news item, with an idealized point/counterpoint discussion.

However, a platform like Twitter reflects a different user experience designed to harness more open-ended discussion. The entire Twitter platform is based on the message board design pattern [62]. Users can easily access and engage with public content and any public comments. Within each content item, the initiating post and commenting threads carry similar visual weights and are easily seen at a glance. Compared to news websites, where article content has a distinctly higher visual priority than comment threads, Twitter's relationship between initiating posts and comment threads is relatively flat. The inverse-pyramid is still in the content presentation on Twitter, but the instantiation of this pattern is much less obvious (see Figure 3). Another key user experience

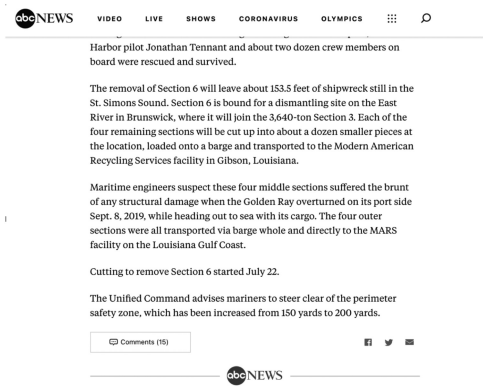

Fig. 1. The comment section is at the bottom of the page and requires additional scrolling to access.

Fig. 2. The comment section of ABC news uses indentation to reflect the hierarchy of the threads.

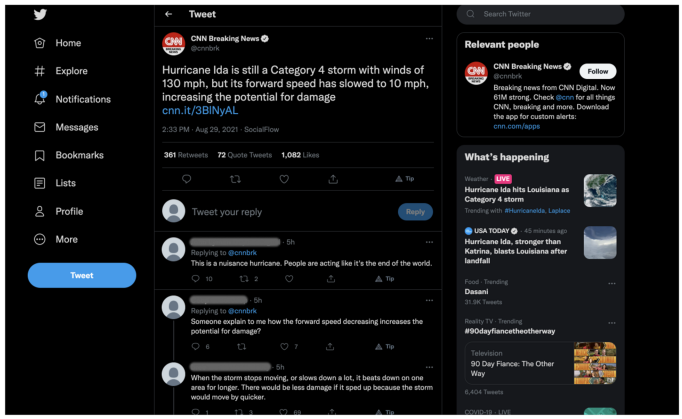

Fig. 3. The message board pattern is clear in the middle pane, with some decorations on the sides. The inverse-pyramid is still present, where the main post is located on the top and the comment section is below. However, the instantiation of this inverse-pyramid is, perhaps, a little less obvious because of the similarity between the initiating post and comments.

difference for Twitter is the way that content is prioritized and the way comments are ordered may be unique to each user based on their known social connections and prior content interactions. On Twitter, the user journey of reading, commenting, and replying is very loosely structured with a discussion experience more like a cocktail party.

We consider YouTube to have a user experience somewhere between that of news websites and Twitter. Similarly to news websites, YouTube also has a clear inverted-pyramid content hierarchy with the videos having the most visual weight. In YouTube, the video provides a clear initiating post that can serve to focus a discussion, which makes its commenting somewhat similar to that of news websites. In contrast to news websites, YouTube users can quickly scroll past the video player to view the commenting section, making comments and discussions more readily accessible. In this way, YouTube comment sections resemble the Twitter experience where users can easily jump to the commenting section and start reading and replying to messages.

In our evaluation, we varied the platform but focused only on comments related to news as reported by recognizable news outlets. The reasoning behind maintaining the news genre is to be systematic in how we test the transferability of Perspective. We specifically focused on news content from recognized news outlets on news websites, YouTube, and Twitter. We have argued above that the user experience of commenting is different across the three platform types. This differing experience could yield different frequencies of toxicity of comments and differing judgments of what constitutes toxicity. The transferability of an AI/ML model across these different contexts is a distinct challenge for any AI/ML model. But there is another factor to the user experience across the three platforms. The behavior of the recognized news agency is also shaped by the platform where they are posting. On a news website, hosted and managed by the news agency, the agency has complete control. On Twitter and YouTube, a news agency posting content is constrained by the design decisions of the specific platform. This point again argues that commenting behaviors on news topics might vary across the three types of platforms, thus challenging the transferability of Perspective's model of toxicity for discussion contributions across these platforms.

Next, we describe how we collected our comment corpus across the three types of platforms.

4 COLLECTION OF COMMENT CORPUS

Different types of content elicit different comments and discussions from viewers. Posts about experiences with pets, movies, social events, sports, or political actions can generate very different commenting responses. We sought to avoid a bias resulting from collecting different types of content from the different platforms by focusing on comments related to news posted by recognizable news agencies. We sourced comments that were responses to news posts and covered a wide range of news categories including politics, entertainment, health, sports, and business. All categories were present across the different platforms. Across all three types of social media platforms we collected from recognizable news agencies. Not every news agency was posting content on every type of platform, although several were present on all. News website comments were collected from sites with active public comment sections. ABC News, FOX News, and *The Washington Post* were our main source for this category. Twitter comments were collected from news posts by official media accounts for ABC, FOX, MSNBC, CNN, ESPN, CBS, BuzzFeed News, Vox, *New York Times*, *Wall Street Journal*, and *The Washington Post*. Similarly to Twitter, YouTube comments were also collected from videos from official media accounts listed above.

The data collection was spread out over 6 months, with approximately 40 comments each week from all three types of platforms (news websites, Twitter, and YouTube). The collection of comments was performed by hand[2] on a convenience basis with a focus on covering a diversity of potential news topics and having a wide selection of comments. We acknowledge that each platform has its own moderation policies, and once a toxic comment has been posted, the chances of it being moderated likely increase over time. Hence, we specifically looked for current news, sorted comments by recency, and collected those that were added most recently. Our current dataset has over 1,000 comments—a little over 300 comments per platform. It is important to note that our approach may not be able to collect the most toxic, offensive, or harassing comments. In fact, Figure 4 illustrates that across all three platforms our comment dataset is skewed toward the non-toxic end of the Perspective probability scores.[3] We attempt to mitigate the skew by sampling from our comments as we describe below in Section 6.

In the next section, we explain our efforts to design a survey instrument that encourages deliberate and consistent judgments from study participants rather than simply conformity with our expectations.

[2]Automated collection of comments was not always possible, because some platforms did not have the APIs to support it.
[3]The Perspective scores used here and for all our analyses are raw API outputs and not calibrated to the target domain.

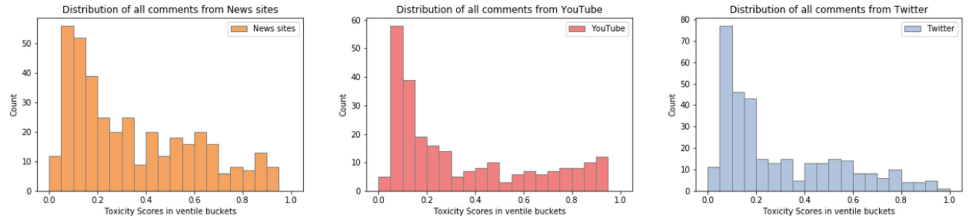

Fig. 4. Distribution of all comments collected, across Perspective's score range, by platform. The left histogram represents scores for comments from News websites, middle comments from YouTube, and right comment scores from Twitter.

5 ITERATIVE STUDY DESIGN

We used Amazon Mechanical Turk (AMT), a crowd-sourcing platform, to recruit participants and collect toxicity ratings. AMT is widely used to label data when creating an AI/ML tool. We had to ensure that the workers on AMT are paying attention to the task, its instructions, and not behaving randomly. Further, since we recognized that judging toxicity is a subjective and difficult task, we needed a carefully designed survey instrument to ensure that participants are paying attention to facilitate accurate and consistent judgments. We iterated through a number of scoring methods, priming questions, survey layouts, and participant qualifications. Below, we outline some key steps in our iterative design process. We first discuss how different attributes used in the study are defined and then describe the different iterations of scoring methods, participant selection, and layout. We obtained Institutional Review Board approval before conducting the study.

5.1 Definitions and Priming Questions

Participants rating a comment for toxicity might have very different opinions of what constitutes toxicity. We leveraged the definition provided by Perspective for their rating task as the one for our task (i.e., "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion") [4]. We recognize that this is not a universal or comprehensive definition but operationalizes toxicity in a way that should make our user ratings comparable to those from the Perspective API.

One challenge of evaluating or rating toxicity is that the underlying thought process may vary for each individual. For some people, it may be a snap judgment, being somewhat instantaneous, while for others the decision may be more deliberative and reflective. We wanted to shift the likelihood of the decision making more toward the deliberative side of the spectrum to reduce unconscious biases. To aid in this shift, we asked participants to first rate three additional attributes for each comment: *respectfulness*, *presence of stereotype*, and *formality*. We use these three rating dimensions to prime the participants' thinking about each comment. Below we discuss our rationale for selecting these particular attributes.

5.1.1 Respectfulness. We chose the attribute of *respectfulness*, because it is specifically mentioned in Perspective's definition of toxicity. The hierarchical relationship suggests respectfulness is a critical component of toxicity. By explicitly asking about respectfulness, we can analyze whether it is a reliable component of toxicity or if it could potentially be an independent attribute. We defined respectfulness as *"a range that indicates whether the written comment illustrates that the author of the comment shows deference, care, or understands the potential feelings of the reader of the comment."* We expect that ratings of respectfulness should contrast with toxicity. That is, the more respectful a comment is, the less toxic it is and vice versa.

5.1.2 Presence of Stereotypes. As we collected our news-related comment data and worked with some of the data that had been used to train Perspective, we noted that some toxic comments seemed to activate stereotypes. We noted that comments that mentioned stereotypes of women, homosexuals, Asians, and Blacks were more likely to have higher probabilities of being toxic. However, there are also many comments that had stereotypes but were not considered toxic. Stereotypes are *not* part of the Perspective toxicity definition, so we picked this attribute as an example of a potentially hidden or latent attribute that is part of how Perspective scores comments. We defined a stereotype as *"a widely held, but a fixed and oversimplified image or idea about a particular type of person, group of people, or thing."* If stereotypes are part of what activates a toxicity judgment, then we should see a higher likelihood of stereotypes being judged as present in the comment when the comment is also judged to be more toxic.

5.1.3 Formality. Last, since our evaluation is considering the transferability of Perspective, and since we operationalized the cross-platform condition as the style of commenting, we primed our participants to consider the formality of the comment as a textual expression. The formality of a text expression may have a direct impact on the way a reader interprets what the comment means. We defined formality as *"a range from 'very formal' written English with proper grammar and punctuation to somewhat formal written statements that might be similar to spoken language to the 'very informal' texting representations that might include 'LOL,' 'brb,' and similar terms, etc."* Given our descriptions of the platforms and the likely interaction styles (cf. Section 3), we expect that comments being judged as more formal would most likely come from news websites while comments judged as least formal would likely come from Twitter. The judged formality of comments from YouTube we expect to fall somewhere in between. Ratings of this attribute allow us to understand whether we have reasonable construct validity for the design differences across the three social media platform types. This allows us to understand whether Perspective scores across content from the three different platform contexts represent how Perspective might operate in different social media contexts or whether all of the comments and the three social media contexts are largely the same.

5.2 Survey Development

Our first survey prototype was quite straightforward, working under the assumption of "try the simplest thing first." The survey asked participants to rate the toxicity of a comment by selecting "toxic," "hard to say if toxic," or "not toxic" from a dropdown choice menu. We tested this survey by gathering ratings for 99 comments, with each comment rated by five different participants. There were 71 comments with at least 3/5 agreement. However, more than half of the comments, 41, were rated as "hard to say if toxic." While it is possible that this reflects an accurate judgment, these scores seemed to reflect "Goldilocks" ratings [11] where people tend to gravitate toward moderate options rather than extremes when faced with difficult choices. As an attempt to disrupt the behavior of mostly picking the middle "hard to say" choice, we changed the toxicity rating to a 10-point scale. However, again, most scores fell in the middle range of 4–7. It is possible that there was genuine ambiguity in the data that caused participants to rate it this way. It was also possible that the labeling scheme did not address the challenge of making these subjective judgements. Since we could not evaluate the validity of these judgements, because of their subjective nature, we decided to update the way in which we elicit these judgements.

Our first two iterations convinced us of the need to design a better way to elicit and verify deliberate toxicity judgments. Our next iteration was inspired by Huffaker et al.'s work [46]. They used anchor comparison as a strategy to elicit judgment on whether some text is intrinsically emotional or emotionally manipulative. We extended this approach by providing two anchors for

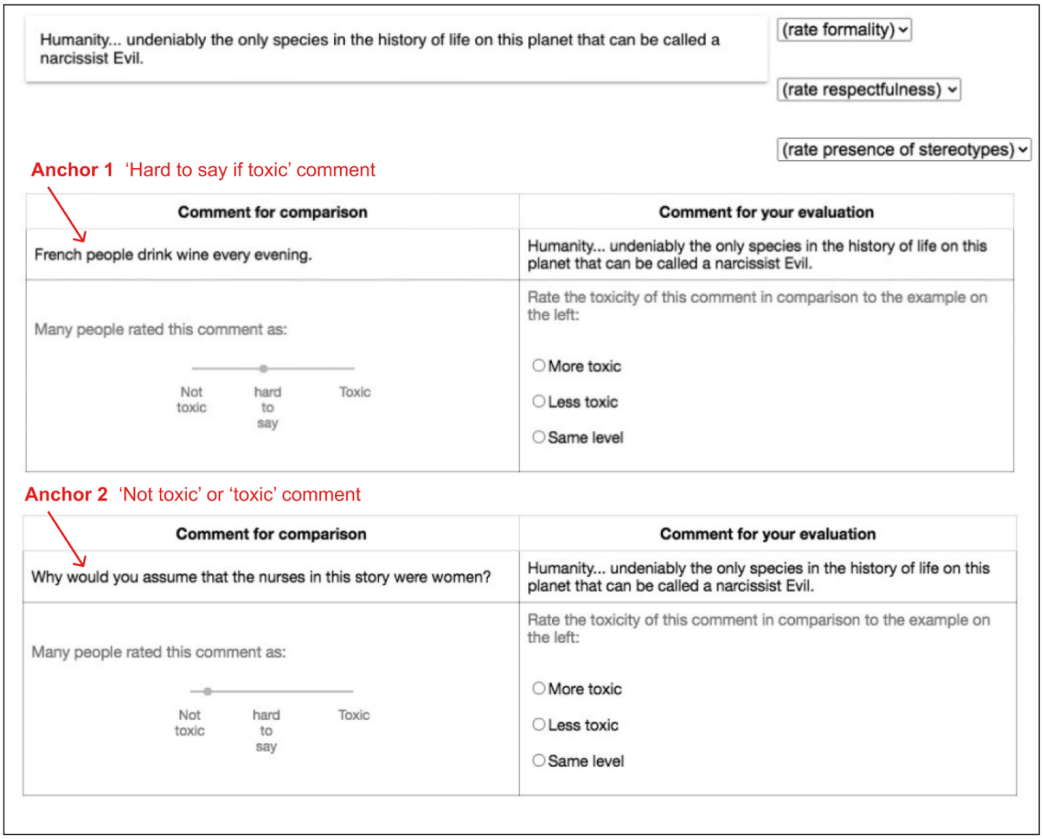

Fig. 5. The same comment is asked to be evaluated twice with respect to two different anchor comments.

every comment. One anchor (A1) would always be a comment with a known middle rating (i.e., "hard to say if toxic") while a second anchor (A2) would be one of the extreme ratings, either "toxic" or "not toxic." In Figure 5, the comment has to be evaluated with respect to a "hard to say if toxic" anchor A1 (card on top) and non-toxic anchor A2 (card below). Given an anchor, the participants must then rate whether the target comment is "more toxic," "less toxic," or "at the same level of toxicity."

Having a participant rate the same comment against two different anchors results in an important side effect. Two ratings allow us to check for consistent and attentive responses. This is a well-known psychometric survey design technique. The benefit can be illustrated with a simple example. Given a comment rated twice, it should not be rated less toxic than a known "non-toxic" comment and more toxic than a known "hard to say" or known "toxic" comment. A participant's ratings with respect to A1 and A2 need to be consistent (see Table 1), or they have had some lapse in attention. We used six of Perspective's publicly available example comments and their corresponding ratings as our initial anchors [7]. As we collected ratings we were able to use some of our previously rated comments that had at least 80% agreement as anchors. In the final survey, the attribute ratings (described in Section 5.1) and anchor comparisons are presented together.

Like most research that relies on crowdworkers, our tasks also included gold standard comments. These are comments where the comments' ratings compared to any given pair of anchors are clear and well defined. We used both gold standard comments and the dual ratings from our anchors to

Table 1. This Table Lists Valid, Consistent Pairs, and the Corresponding Toxicity Rating for Different A1 and A2 Ratings

Responses to A1 (hard to say)	Participant Response to A2 (not toxic)			Participant Response to A2 (toxic)		
	more	**less**	**same**	**more**	**less**	**same**
more	toxic	*NA*	*NA*	toxic	hard to say/ toxic	toxic
less	hard to say/ non-toxic	non-toxic	non-toxic	*NA*	non-toxic	*NA*
same	hard to say	*NA*	*NA*	*NA*	hard to say	*NA*

For example, if the participant responded to A1 with "more," then we find all possible responses with respect to A2 in Row 1. Further, if the participant responded to a non-toxic A2 with "less," then this is an invalid response.

inform our decisions about the quality of a participant's response. We developed a specific rubric to decide whether any given set of responses would be used in our analysis (see Appendix A.2 for detailed rubric).

5.3 Crowdworker Participant Criteria

AMT has many different criteria that can be set to select the types of workers who will be eligible for a given task. Some criteria are based on user-declared categories, like age, while other criteria are set as a function of worker performance. Mitra et al. recommend several person-centric strategies for recruiting and selecting crowdworkers, such as screening, training, and financial incentives to get quality data [55]. Our approach relied on criteria available from AMT and therefore reflects a "screening" approach to crowdworker recruitment and selection. We restricted the worker geographic location to the U.S. and required an initial approval rate of 80%. For each Turker, this rate indicates the percentage of tasks that were completed by them and approved. The 80% approval threshold should be considered quite generous for this type of task. Through several trials, we raised this approval rate incrementally and evaluated how different Turkers performed. In one of our trials, we used the "Masters" qualification. This qualification is given to Turkers for consistent, quality work completion, over thousands of tasks completed. Amazon does not explicitly state how it selects or awards the "Masters" qualification, but these workers are considered the highest-quality workers on AMT. Based on our trials, it was clear that "Masters" workers were highly consistent and highly attentive to our task. All of the data we report on here was completed by Masters workers. We rewarded our participants $2.00 to complete each survey task that averaged 10 minutes of work.

6 CONDUCTING THE STUDY

We had participants rate a total of 300 comments. We selected 100 comments from each platform, spread over Perspective's probability score range. We selected comments to roughly balance the number of comments in each decile probability bucket (e.g., 0.0 to 0.1, 0.1 to 0.2, and so on; see Figure 6). We avoided excessively long and discursive comments, because they are more likely to introduce several different ideas, any one of which might activate one of our judgment attributes. The average length of a comment in words was 21.13 and the standard deviation, 10.8.

The final survey task consists of several parts: instructions and task description, demographic collection, and the anchor comparison task. Our task description explicitly stated our use of attention checks, approximate time required to complete the task, and a "trigger" warning about the potential impact of reading hateful and disturbing comments (see Figure 15 in Appendix A.1). We collected basic demographic data in the second step (see Appendix A.4 for an overview of the data).

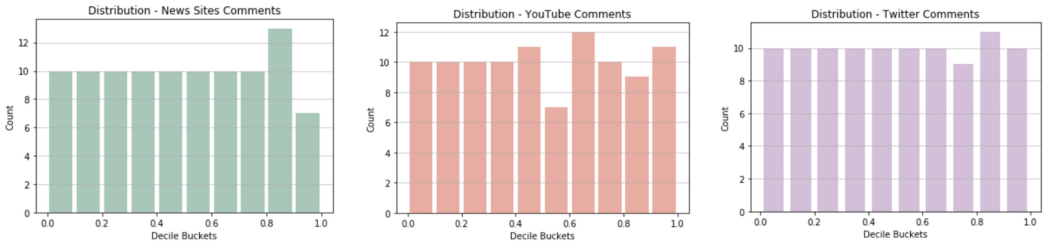

Fig. 6. Distribution of comment scores across Perspective's score range for comments rated by our participants. The goal was to have roughly equal numbers of comments in each decile bucket for each platform.

The analysis below relies on platform usage information that was collected as part of the demographic information. In steps 3 and 4, a total of 15 comments had to be rated. For each comment, we ask the participant to first rate the formality, respectfulness, and presence of stereotypes. The participant then rates the toxicity of the comment in comparison to the two anchors.

Among the 15 comments were three gold standard comments. We used the participants' ratings of our gold standard comments and any inconsistent anchor comparisons to evaluate the quality of the participant's response. Our acceptance or rejection criteria, allowing some room for mistakes, are detailed in the Appendix (see Table 11). We used these criteria to decide on when to approve their work and use the data, approve their work and not use the data, or reject the participant's response. Our analysis only includes data that was generated by compensated workers. There is no condition where we would include any data generated by a worker who was not compensated.

7 ANALYSIS AND FINDINGS

We collected ratings for 300 comments with each comment being rated by 5 different participants. Each participant rated at least 12 comments from the set and there were a total of 55 unique participants who provided ratings. First, we provide information about which platforms our participants commonly used. We describe how we aggregated ratings to consider the inter-rater reliability scores. After describing these preliminary steps of data aggregation and analyses, we examine the relationship between each attribute that participants rated and Perspective's toxicity attribute by fitting ordinal logistic models. We generate probability curves to interpret the model and depict these relationships. Finally, we analyze how the attributes toxicity, formality, respectfulness, and presence of stereotypes relate to each other, from the participant's point of view.

7.1 Participant Platform Familiarity

Our survey instrument collected information about how frequently our participants used different social media platforms, including Facebook, Reddit, Twitter, YouTube, and news websites. The frequency of use for news websites, YouTube, and Twitter are shown in Table 2. The majority of our participants (at least 40 of 55 total) use these three types of platforms at least once per week. This indicates that many of our participants are quite familiar with these platforms and their characteristics. Further, there are participants who use two or more of our target social media platforms.

When our participants were making judgments, our survey instrument did not expose from which type of platform any comment was collected. This may make the judgment task harder as a participant in this condition is not able to leverage their knowledge of any normative behavior or normative linguistic usage when evaluating a comment. This condition is similar to the way data was initially scored when creating Perspective's models and is very similar to the context in which an AI/ML tool executes. That is, Perspective uses only the comment text to produce a score.

Table 2. Participants' Self-reported Information about How Frequently They Used News Websites, Twitter, and YouTube

	News websites	YouTube	Twitter
Several times a day	21	28	11
About once a day	13	14	18
About once a week	12	9	9
About once a month	3	3	3
Less than once a month	4	0	4
Never	2	1	10

Table 3. Results of Mann–Whitney U Test Comparing Toxicity Ratings between Two Groups, Users and Non-users, for Each Platform

	News Websites	YouTube	Twitter
Mann–Whitney U Test (p value)	$p > 0.05$	$p > 0.05$	$p < 0.05$

A study could certainly select participants based on their usage of a platform and have them rate comments exclusively from that platform. That type of study has a stance that normative notions derived from a person's experience on a platform are unlikely to be captured by an AI/ML model and that those, likely subtle, differences will severely impact the transferability of an AI/ML model. Our study takes a slightly different stance by first asking whether there is any agreement between Perspective's probability scores and human judgments. If there is no agreement or very weak agreement, then the subtlety of normative experience within a platform might be the explanation and a more methodologically sophisticated study could be a way to demonstrate that effect through a follow-on study. We conducted this simpler study first.

We wanted to see if our participants' ratings had a bias that was a function of their self-declared social media use for a given platform. Based on their self-reported frequency of social media use (see Table 2), we ran a Mann–Whitney U Test comparing toxicity ratings for comments from a particular platform for "frequent" and "infrequent" users of that platform. We considered "frequent" users of a platform to be those who selected "about once a week" or more frequent. The "infrequent" users were those who selected "about once a month" or less frequent. The results of running the Mann–Whitney U Test for both groups, platformwise, are shown in Table 3.

The results show that for frequent and infrequent users of news websites and YouTube there is no significant difference when they are only rating comments from those respective platforms. This is a good start. However, for frequent and infrequent users of Twitter, when rating only comments coming from Twitter, there is a small difference. We then examined whether this "Twitter Bias" would impact comments from the other platforms. A Mann–Whitney U test comparing all ratings of all comments for frequent and infrequent users of Twitter showed no significant difference (p value > 0.05). This provided us some confidence that the "Twitter bias" would not be a pervasive influence across all of our ratings. Additionally, it should be pointed out that many of the "frequent" Twitter users in our study were also "frequent" users of news websites and YouTube. Therefore, our subsequent analysis will aggregate our participant judgments.

We want to point out that finding a slight "Twitter bias" provides some support that evaluating concepts like toxicity with participants in a more narrow context may result in higher-quality annotations. If our results below were to find that the Perspective model failed to function well across the three platforms (i.e., did not demonstrate transferability), then the "Twitter bias" that we both demonstrate and then subsequently discount may actually be portion of the cause and would warrant a deeper study.

Table 4. Krippendorf's Alpha and Level of Agreement for Participants' Scores for All Rated Attributes

Attribute	Comments, Raters	Krippendorff's alpha	Level of Agreement
Formality	300,55	0.301	Fair
Respectfulness	300,55	0.565	Moderate
Presence of stereotypes	300,55	0.396	Fair
Toxicity	300,55	0.444	Moderate

7.2 Rating Reliability and Aggregation

We intended to understand how well our participants performed on the rating task. We noted that this type of judgment task is quite difficult, and our early attempts to reliably rate comments showed high variances. We described above (cf. Section 5.2) our efforts to manage this difficulty by creating a survey that guides the participant through the judgments and by devising ways to figure out which participant groups were more consistent in their scores.

We analyzed the inter-rater reliability of our data using Krippendorff's alpha. We chose Krippendorff's alpha, because it works for ordinal variables and arbitrary number of items (comments) and raters and is flexible about missing data. Table 4 shows Krippendorff's alpha and level of agreement for each attribute scored in our survey. Our primary motivation for checking inter-rater reliability is to get guidance on how to aggregate and analyze ratings from different participants. Given the subjective nature of this type of scoring, we expect differences in opinion on what is toxic.

Our interpretation follows the benchmarks by Landis and Koch [13, 52] with 0.21–0.40 considered "fair," 0.41–0.60 considered "moderate," 0.61–0.80 considered "substantial," and >0.81 considered "near perfect." Our alpha measures indicate fair to moderate levels of agreement. The judgments are properly ordinal measures, since the relative "distances" between the textually described labels are not well defined. Calculating means for these types of scores is not strictly a fair treatment of the data. Therefore our primary analysis is to combine five ratings for each comment by taking the mode or majority rating.

Toxicity is our focal attribute. As we described we had two different ways to validate the ratings of our participants. We worked to get at least five valid ratings for every comment, but even with attempts to collect additional ratings, we were not able to always have five. Excluding inconsistent toxicity ratings resulted in 1,426 (of a best case of 1,500) comment ratings. For each comment, the set of valid ratings were collapsed into three ordered categories: Not toxic, Hard to say, and Toxic.

A majority rating is sometimes from only four or three valid scores. A majority rating of 2/4 is not necessarily weak or poor data, as the tie could be between two judgments of "Hard to say" and two judgments of "Not toxic." These types of tied scores reflect some of the difficulty in this specific judgment task. Our acceptance of scores for our other attributes, Formality, Respectfulness, and Presence of stereotypes, was based on the acceptance or rejection of the toxicity scores for the respective comment. These scores are also ordinal data and were also collapsed using the mode (majority rating). In the Appendix, we provide detailed information about observed agreement for the different attributes (Tables 12 and 14) and how we resolved cases where responses were split across two different ratings (Tables 13 and 15).

7.3 How Do Participants Agree with Perspective Toxicity Score?

Our primary hypothesis is based on the relationship between how our participants rate the toxicity of a comment and the Perspective API probability of a comment being toxic. This is quite different than the way an AI/ML tool is commonly evaluated. In essence, we are asking How well do users agree with the prediction?, whereas the common AI/ML question is How well does an AI/ML tool agrees with prior scores given by users? This may seem esoteric, and even pedantic, but one

Table 5. The Distribution of Comments in Categories Not Toxic, Hard to Say, and Toxic for Perspective and Participant Scores

	Not toxic	Hard to say	Toxic
Participants	48	96	156

	Not toxic (<= 0.30)	Hard to say (0.31–0.70)	Toxic (>0.70)
Perspective	92	121	87

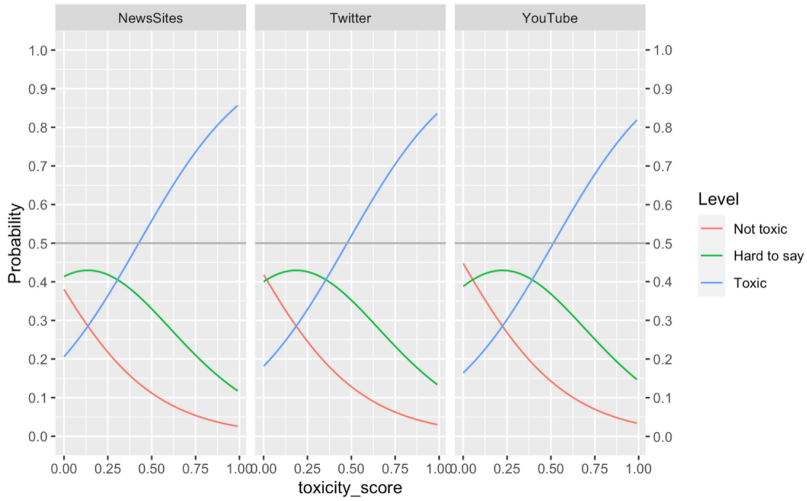

Fig. 7. Probability of how human toxicity ratings correspond to Perspective predictions for news websites, Twitter, and YouTube comments related to news topics. For example, as the toxicity score increases (x axis), the probability that a participant would have rated the comment as toxic also increases (y axis).

direction of the question is clearly valuable during the creation of the AI/ML tool, whereas the direction of our question is about whether the tool might satisfy user expectations once deployed.

For our first analysis, we considered whether the Perspective toxicity score is a predictor of the participants' ratings. The distribution of the comments across these categories in Table 5 shows differences between Perspective and our study participants.

We ran an ordered logistic regression in R with our participants' toxicity rating as the outcome variable and Perspective API toxicity score and platform (i.e., news website, YouTube, Twitter) as predictor variables. The coefficient for toxicity score, 3.1672, was significant with a p value of 7.57e-13,[4] while the platform was not significant. These results indicate that a toxicity score is a strong predictor of how participants will rate toxicity but there are no discernible platform effects. The resulting probability curves from the ordered logit model are shown in Figure 7. The probability curves are shown in different panels based on the platform. We note that these curves are all from the same model based on toxicity ratings by participants. Separation by the platform is to make the curves easier to read. These curves shows the probability (y axis) that user will rate a comment as Not toxic, Hard to say, or Toxic as a function of the Perspective's toxicity scores (x axis).

[4]We report the exact p value, since this is a Maximum Likelihood Estimation method and we deal with likelihood, not probability.

Table 6. The Distribution of Comments for Participant Ratings of Text Formality

Formality rating	Percentage of comments labelled with this formality rating
Very formal	1.08%
Formal	29.96%
Informal	60.65%
Very informal	8.30%

These types of probability graphs can show us several important things. These graphs show us where the user is largely in agreement with Perspective and where the underlying AI/ML model may be confusing one category for another as a function of the user judgments. The crossing points of these curves tell us which category (Not toxic, Hard to say, and Toxic) the user is most likely to agree with at which Perspective API score. This is important, because it can tell us whether the Perspective cutoff values are valid across all platforms or whether they might vary. An important aspect of reading these types of graphs is to note the 0.50 probability line, which is roughly halfway up the y axis. When the curve is above this line, it means there is greater than 50% probability that the user agrees with the AI/ML prediction; anything below this horizontal line is somewhat worse than flipping a coin.

The curves in Figure 7 demonstrate that scores of "Toxic" are quite accurate. The curves suggest that the Perspective API cutoff for Toxic at 0.70 or greater is quite high. In fact, the probability that users would agree that a comment was Toxic is better than 50% for Perspective scores greater than 0.55 for all of the three platforms. This finding illustrates that Perspective shows transferability of toxicity scoring across different platforms with different styles and norms of interaction in the context of discussions of news-related items. However, for all platforms predictions of "Not toxic" and "Hard to say" are largely guessing. That is, the probability of the users agreeing with these predictions is quite low. This illustrates how hard it is to get these predictions correct in the view of a user.

There is one more subtle aspect of the curves in Figure 7. The 50% crossing point is slightly different for news website compared to Twitter and YouTube. That is, for comments from a news website a toxicity probability score of 0.45 is roughly where humans are more likely to agree that a comment is toxic. In the case of YouTube and Twitter the curve shifts toward the right, meaning humans are only more likely to agree when the probability of a comment being toxic is higher. While the difference is not statistically significant, that humans can "see" that news websites have a lower threshold for what constitutes a toxic comment is some support that our users are detecting differences in the platform without being explicitly told.

7.4 How Do Formality, Respectfulness, and Stereotypes Relate to Toxicity Score?

Our participants rated three additional attributes for each comment. These attributes reflected different aspects of what might be happening in the way users consider the toxicity of a particular comment. Respectfulness was a type of check on a dimension that was specifically mentioned in the definition of toxicity. Formality and presence of stereotypes reflect possible latent attributes that may be part of the way people, or Perspective, judge toxicity. We cover our findings for each of these attributes in the subsections below.

7.4.1 Formality. The style and form of a comment may be one aspect that influences how users see a contribution as either toxic and offensive or somewhat less so. We examined how our collected ratings of formality relate to the Perspective toxicity scores. Table 6 shows the percentage

Table 7. The Distribution of Comments in Categories, Not Toxic, Hard to Say, and Toxic for Participant Ratings of Text Formality

	Not toxic (<= 0.30)	Hard to say (0.31–0.70)	Toxic (>0.70)
Very formal	2	1	0
Formal	34	38	11
Informal	46	70	52
Very informal	4	5	14

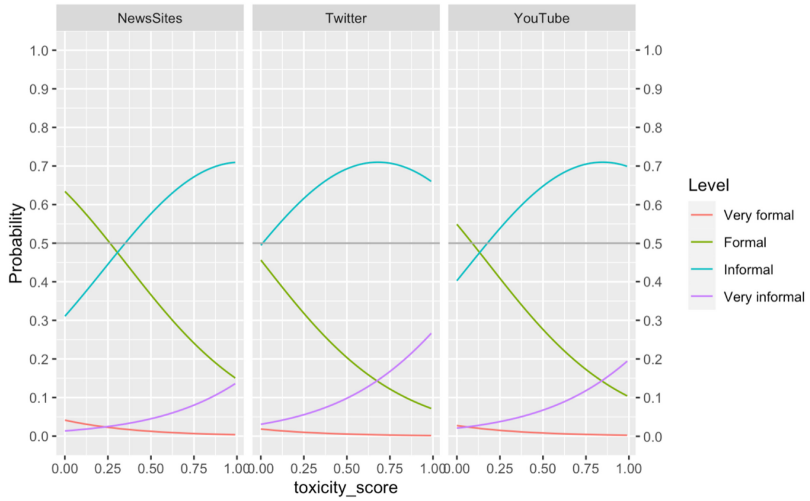

Fig. 8. Probability of text formality rating corresponding to Perspective predictions for news sites, Twitter, and YouTube comments related to news topics. The probability curves are all from the same model based on formality ratings by participants. Separation by platform is to make the curves easier to read.

of comments labelled with each of the formality ratings and Table 7 shows. The table illustrates that formality extremes, Very Formal and Very Informal, are much less common.

We ran an ordered logistic regression in R with formality rating as the outcome variable and toxicity score and platform as predictor variables. Both toxicity score and platform effects were significant in this model (respective p values are 0.0000001468 and 0.02).[5] We also tested for interaction effects but if they were present, they were not discernible (statistically insignificant). The resulting probability curves from the ordered logit model are shown in Figure 8.

The curves for formality demonstrate a relatively high probability that a toxic comment is expressed with informal text *across all three platforms*. This is much easier to see if we condense our categories of formality to just two, something more informal and something more formal. Figure 9 shows the probability curves with categories collapsed into lower and higher formality.

The curves for news websites look almost ideal in response to toxicity scores. Up until a certain point (toxicity score ≈ 0.27), text is clearly distinguished as formal, beyond which as toxicity increases, chances of text being informal also increases. News websites also have the highest probability of a text being formal for lower toxicity scores. This indicates that the most formal comments

[5]We report the exact p value, since this is a Maximum Likelihood Estimation method and we deal with likelihood, not probability.

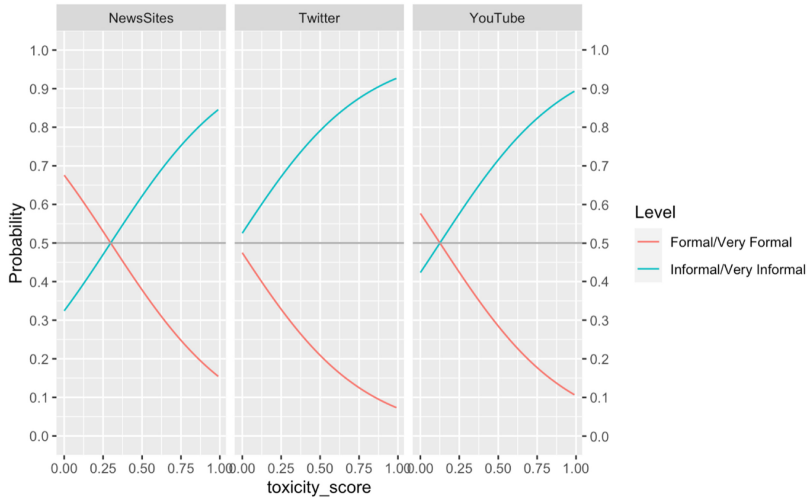

Fig. 9. Probability of collapsed formality rating corresponding to Perspective predictions for news websites, Twitter, and YouTube comments related to news topics. For example, as the toxicity score increases (x axis), the probability that a participant would have rated the comment as formal decreases (y axis).

in our data come from news websites. The curves for YouTube have less overlap than news websites but also intersect, unlike Twitter, indicating that comments from YouTube fall somewhere in the middle of the formality spectrum.

It is the Twitter curves that show, perhaps, the most interesting property. There is no crossover point where the formality or informality of the text corresponds to a low toxicity score. This illustrates that many of our Twitter content samples were rated as informal, a common stylistic property of that platform. But this also illustrates that some aspect of textual formality is tightly coupled with toxicity. We discuss this a little later as well in our analysis in Section 7.5.

These probability curves illustrate that when Perspective scores are high, participants' ratings of text are more likely to be informal. This suggests that the Perspective's understanding of toxicity is somehow related to the formality of the textual expression. This correlation could be a result of human judgment or biases during the labeling process, or it could be inherent to how Perspective works.

The graphs in Figure 9 and the fact that they reflect statistically significant differences help to validate our construct that the platform styles we hypothesized are, in fact, different. That is, the curves reflect more formality for news websites, least formality for Twitter, and something somewhere in between for YouTube. Without telling our human participants the origin of the specific comments, they are able to detect clear differences over our sample set of comments. This does not mean that they can detect the originating platform for a single comment, just that they detect it in aggregate.

7.4.2 Respectfulness. Our definition of toxicity, and that used during the development of Perspective, specifically invokes the respectfulness of a comment as an attribute of toxicity. We had our participants rate this attribute as a type of supporting validity check. We would expect that there should be a rather strong negative relationship between respectfulness ratings and toxicity ratings. That is, as toxicity increases the respectfulness should decrease.

We ran an ordered logistic regression in R with respectfulness rating as the outcome variable and toxicity score and platform as predictor variables. Toxicity score had a statistically significant effect

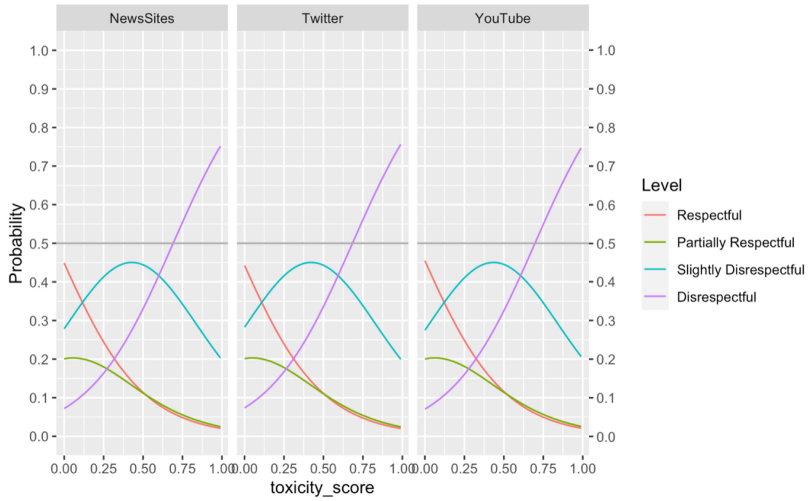

Fig. 10. Probability of respectfulness rating corresponding to Perspective predictions for news websites, Twitter, and YouTube comments related to news topics. For example, as the toxicity score increases (x axis), the probability that a participant would have rated the comment as disrespectful also increases (y axis). The probability curves are all from the same model based on respectfulness ratings by participants. Separation by platform is to make the curves easier to read.

in this model (p value = 1.401e-16[6]). The platform had no discernible effect on the probabilities of different respectfulness ratings. The resulting probability curves from this model (Figure 10) are nearly identical for all three platforms. One can fairly easily see how the platform variable was insignificant in this model.

Figure 10 also shows that our expectation regarding the relationship between toxicity and respectfulness holds. In general, as the toxicity score increases, the likely respectfulness decreases. That is, as the toxicity score increases, the probability of disrespectfulness increases. These probability curves illustrate that respectfulness is clearly an attribute of the way people evaluate the toxicity of a comment. Our use of respectfulness as a validity check on toxicity ratings is at least consistent with our expectations. This seems to show that our participants were likely rating toxicity using a similar mental construct as the raters whose data were used in the creation of the Perspective model.

7.4.3 Stereotypes. We analyzed if the toxicity score is a predictor of how participants rated the presence of stereotypes. As we worked with the comment data and Perspective, we became aware that stereotypes were often expressed in the comment data. The different ways that Perspective scored these comments made us wonder if stereotype was a potentially latent attribute for toxicity. In this case, a latent attribute is a type of construct that human evaluators might be using when making a toxicity judgment but that was not explicitly called out by the original ratings collection. Latent attributes are interesting in AI/ML models, because they are something that a model may have learned, may detect, and may score but about which it has no explicit features. Latent attributes in an AI/ML model may be one source of bias in an AI/ML score that human-centered evaluations should seek to uncover. Descriptive statistics on how ratings of stereotypes varied with toxicity scores is listed in Table 9.

[6]We report the exact p value, since this is a Maximum Likelihood Estimation method and we deal with likelihood, not probability.

Table 8. The Distribution of Comments for Participant Ratings of Stereotypes in Text

Stereotype Rating	Percentage of comments labelled with this stereotype rating
Not present	55.07%
Possibly present	26.81%
Present	11.6%
Heavily present	6.52%

Table 9. The Distribution of Comments in Categories Not Toxic, Hard to Say, and Toxic for Participant Ratings of Stereotypes in Text

	Not toxic (<= 0.30)	Hard to say (0.31–0.70)	Toxic (>0.70)
Not present	57	58	37
Possibly present	24	26	24
Present	8	15	9
Heavily present	2	9	7

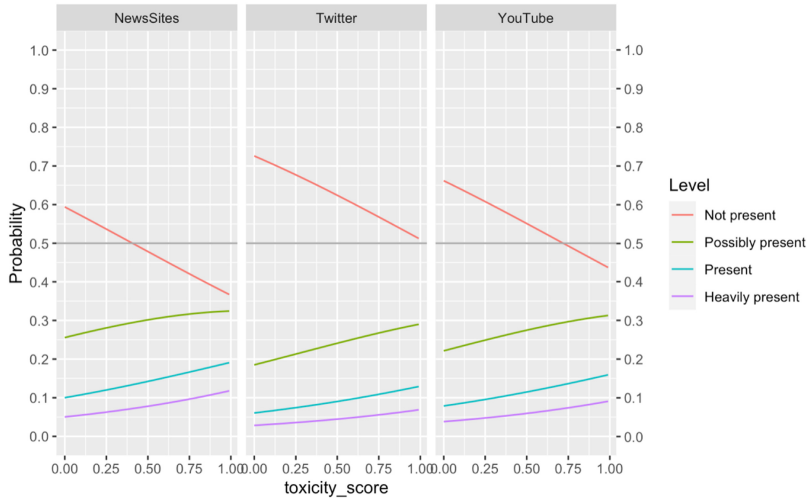

Fig. 11. Probability of stereotype rating corresponding to Perspective predictions for news websites, Twitter, and YouTube comments related to news topics. The probability curves are all from the same model based on the stereotype attribute rated by participants. Separation by platform is to make the curves easier to read.

From Tables 8 and 9, we see that participants primarily used the "not present" and "possibly present" rating compared to the "present" and "heavily present" rating. Roughly half of the rated comments did not have stereotypes according to our participants. Using toxicity score and platform as the predictor variables shows that toxicity score has a significant effect on this rating (p value = 0.0215).

The probability curves in Figure 11 explain the behavior better. Across all platforms, the probability of "not present" decreases with increasing toxicity score. As well, for all platforms, all of the other ratings of presence (i.e., "possibly present," "present," and "heavily present") increase with

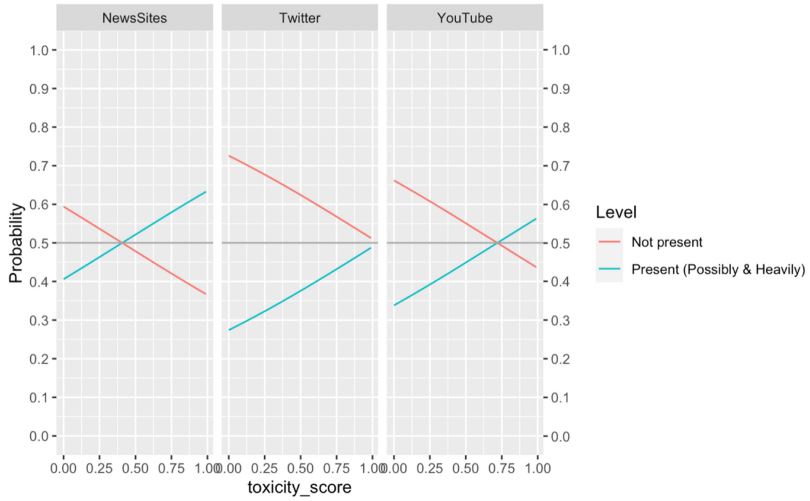

Fig. 12. Probability of combined stereotype ratings corresponding to Perspective predictions for news websites, Twitter, and YouTube comments related to news topics. For example, as the toxicity score increases (x axis), the probability that a participant would have rated the comment as having stereotypes also increases (y axis).

increasing toxicity. Visually, the increasing likelihood of stereotype ratings are nearly parallel, suggesting that all three forms of potentially present stereotypes are somehow similar.

Building off of that insight, we added the probabilities of "possibly present," "present," and "heavily present" and generated a single curve (see Figure 12). Using the "not present" compared with all forms of some how present stereotypes, we see that high toxic scores correspond to a substantial probability of the text having stereotypes in both news websites and YouTube. Low toxic scores correspond to a substantial probability of the text not having stereotypes (comments from Twitter have the highest probability). This evidence suggests that the presence of stereotypes is a latent attribute of how individuals understand the toxicity of a comment. Figure 12 again shows that without explicitly signaling to our human raters from which platform a given comment originated, they can "see" differences in aggregate.

7.5 How Do Formality, Respectfulness, and Stereotype Ratings Relate to Toxicity Ratings?

There is at least one problem with analyzing each of the attributes as if they were separate and individually identifiable by our participants. The major problem is that the concept of "toxicity" is multi-faceted. The very definition we presented above has at least three potential components, and there may be more. We sought to understand how our three attributes related to each other and to the toxicity ratings. For example, did participants' toxicity ratings vary based on how they rated formality, respectfulness, and presence of stereotypes? Such an analysis helps us understand how participants perceived toxicity and if they considered formality, respectfulness, and the presence of stereotypes as indicators of toxicity.

We ran an ordered probit with toxicity rating as the outcome variable and formality, respectfulness, and presence of stereotypes as predictor variables. We used individual observations ($n = 1426$), since we wanted to analyze how participants rated the attributes at a granular level. Our analysis accounted for the fact that comments and participants were repeated across this dataset to avoid overestimation. Both respectfulness and the presence of stereotypes had a

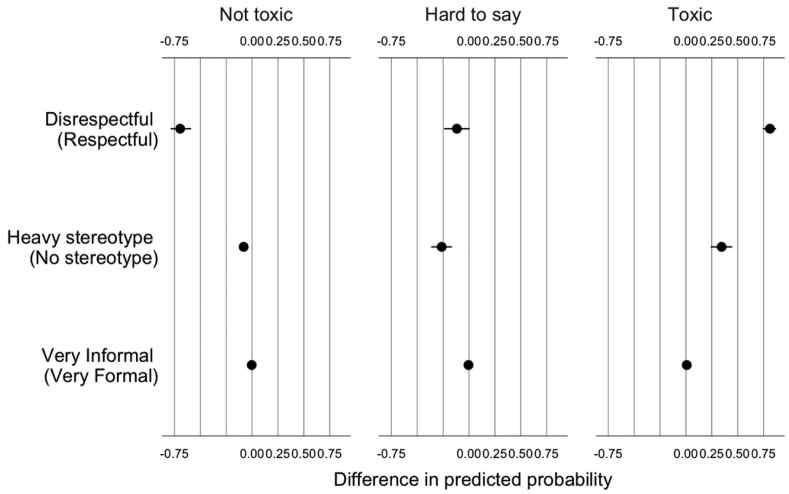

Fig. 13. Difference in the probability of predicting toxicity simulated by changing each predictor variable from its lowest value to highest value.

significant effect on how toxicity is rated (respective p values <2e-16 and 1.04e-09). However, in this model *formality ratings do not have a significant effect on how toxicity is rated (p value = 0.6781).*

For this model, we calculated first differences in predicted probability. This is the difference in probability of a comment being toxic, caused by changing a predictor variable from its lowest value to the highest value [50]. In Figure 13, we see that if we change the respectfulness rating from Respectful to Disrespectful, then there is a high positive difference in the probability of a comment being toxic and a high negative difference in the probability of a comment being rated not toxic.

However, changing the formality rating from Very Formal to Very Informal does not affect the probability of a comment being Toxic, Not toxic, or Hard to say—the difference is essentially zero. Formality ratings did not have a significant effect on toxicity ratings as per the probit model as well. One plausible explanation is that this illustrates a key difference between the way that humans evaluate comments for toxicity and the way that Perspective evaluates comments. This suggests that when humans rate comments for toxicity they can separate formality from toxicity. Comments written in a formal style could still be offensive, disrespectful, or toxic. As well, comments written in an informal style with txt language, emoticons, and shorthand could possibly be non-toxic. We list examples of such comments from our dataset in the Appendix section (Table 16). Our study participants are able to comprehend and distinguish these different aspects of a comment. But since participants' formality ratings show a strong correlation to Perspective's toxicity scores (Section 7.4.1), Perspective *potentially conflates the toxicity of a comment with the formality of its textual expression.*

8 DISCUSSION

8.1 Human-centered Evaluation of Perspective

One objective was to evaluate the performance of Perspective's model across different social platforms with hypothetically different styles and norms of interaction. We found that Perspective's toxicity detection scores align with how potential users are likely to rate toxicity. While this alignment is consistent across news websites, YouTube, and Twitter, it only applies to relatively high

toxicity scores. Since most moderation use cases involve flagging comments that are likely or highly toxic, we believe Perspective demonstrates potential utility for moderation in news-related discussions that occur on different social platforms. Further, Perspective's motivating definition of toxicity (i.e., "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion") aligns with participants' ratings of disrespectfulness across all three platforms, indicating that the attribute of "disrespect" is more present in more toxic scored comments.

The second objective of this evaluation is more subtle and is operationalized by our choice of attributes for which participants provided ratings, especially "presence of stereotypes" and "formality." Neither formality or stereotypes are part of Perspective's toxicity definition and neither are they subattributes [2]. But our work with sample data and our growing comment corpus led us to suspect that formality and stereotypes could be aspects of the text that influence humans and were potentially reflected in Perspective's scores as latent attributes.

Our evaluation provides evidence strongly suggesting that Perspective has learned something about the existence of stereotypes. In particular, it seems that if Perspective scores some text with a high probability of being toxic, then it is more likely that the text expresses some type of stereotype. This is an important finding, because traditional evaluations of AI/ML tools are not designed to uncover latent attributes of the model. This result indicates that careful and systematic human-centered evaluations of AI/ML tools can provide deeper insights into the way that these tools function—insights that are unlikely to be recognized by the developers. The claim that we are making here is not a one-off. Our evaluation demonstrates support for these same claims through the formality of text attribute. It also provides a deeper insight into the ways that participants of our study made judgments differently than Perspective did for the dataset that we evaluated.

We asked participants to explicitly consider the formality of comments as an operationalization of commenting style. We found that the style (i.e., formality) of a comment did not influence participants' toxicity judgments. However, that is not the case with Perspective. Again, text formality is not an explicit subattribute in Perspective's definition of toxicity. As such, this attribute would not seem to be obviously included in the way humans labeled the original training data. This suggests, that the formality of the text expression was learned as a latent attribute during the training process. This also points out that humans participating in carefully structured evaluations can and do observe the dimensionality of separate attributes in situations where an AI/ML model may be correlating or conflating what are logically separate attributes.

That an AI/ML model may learn latent attributes that are not explicitly considered by the designers of the tool is understandable. One key part of AI/ML model development is feature engineering. In this part of development, details of the potential inputs to the tool are dissected to understand which details distinguish one input from another. These details, or features, form a hypothesis about what might be important to differentiate the potential goals for predicting a label or a probability distribution over a set of labels. As a simple example, a contribution to a discussion might have features such as character length, number of words, sets of words or *n*-grams, the number of capital letters, among many possible other features. At times a feature engineer will pick or create features that are known to be related to some rated attribute that the human raters might provide. For example, there might be a specific subset of words that are swear words, or specifically vulgar words that may be a direct representation of rudeness. The underlying algorithm takes the human ratings, the input objects, and tries to understand how the existence of the features and potential relationships among features can be used to group inputs reliably into the desired categories that result in the labels. Our main point is that feature engineering makes explicit attempts to represent a given attribute with underlying features. However, when there are large numbers of features the relationships among them may come to represent attributes of the inputs that were not envisioned when the AI/ML tool was being

Table 10. A Summary Table Showing which Variables Were Significant in Predicting Human Ratings

Outcome variable	Predictor variable 1: Toxicity Score from Perspective	Predictor variable 2: Platform (News websites, YouTube or Twitter)
Toxicity	p **<0.05**	p >0.05
Respectfulness	p **<0.05**	p >0.05
Formality	p **<0.05**	p **<0.05**
Presence of stereotypes	p **<0.05**	p >0.05

As the table shows, toxicity score from Perspective is significant across all outcome variables, and platform effects are significant only for the formality rating.

designed and trained. When unintended relationships among potential features are learned, they may result in the tool having latent attributes that it may or may not label correctly.

We want to be fair and note that the developers of Perspective are aware of this issue. Perspective can provide a set of additional attributes related to toxicity. Perspective can score several additional attributes including the following: Severe Toxicity, Insult, Profanity, Identity attack, Threat, and Sexually explicit. As well, there are a set of additional "experimental" attributes that can be scored, with a clear statement of potential limitations for those attributes. Some of these attributes may be well aligned with specific features that were part of feature engineering in the design phase of this tool. However, we need to point out that the two latent attributes uncovered by our evaluation are not in the current list of Perspective's potential attributes.

Uncovering such latent attributes supports further characterization of how Perspective behaves and illustrates a way that research might also uncover latent biases. Much of the Fairness, Accountability, Transparency, Ethics (FATE) research seeks to balance complex tradeoffs with difficult social implications. However, latent attributes in an AI/ML model are much like latent bugs in software; they are difficult to discover, can have severe impacts, are revealed in rare cases, and are difficult to prevent. Training an AI/ML model developer to have more awareness of FATE issues can only go so far when the specific methods of feature engineering result in possibly millions of potential connections among features any of which might result in a latent feature with detrimental consequences. For example, if we did not uncover the presence of stereotypes as a latent attribute, then we would not be able to explore questions about whether Perspective considers some stereotypes more toxic than others. Further, methods and techniques that help us to uncover latent attributes allow us to consider whether those attributes are desirable behavior from the AI/ML in the first place.

8.2 Detecting Platform Differences

In testing Perspective's transferability, we find that it performs consistently and similarly in scoring toxicity for news-related comments across all three platforms. However, as we premised in Section 3, we believe that these platforms offer different user experiences that directly influence how comments are perceived. When we aggregate ratings from our participants, our results exhibit a ranking or ordering effect for the platforms, especially for the attributes of formality and presence of stereotypes. Participants have detected platform-based differences related to how formal a comment is or if it exhibits a stereotype (without knowing the source platform) and these differences are reflected in their ratings. The source platform had significant effects as a predictor variable only for formality ratings however(see Table 10).

As the plots in Figure 14 depict, the trends are similar for both formality (left) and presence of stereotypes (right) across news websites, YouTube, and Twitter. But the cutoff points (where the two curves intersect and/or cross the x axis for 50% probability) differ. For example, higher toxicity scores correspond to a high probability of participants rating the text as informal in the following

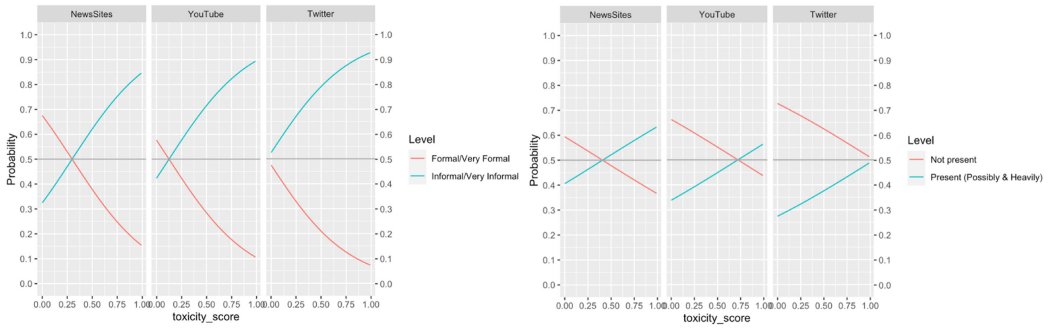

Fig. 14. Probability of combined formality (from Figure 9) and stereotype ratings (from Figure 12) corresponding to Perspective predictions for news websites, Twitter, and YouTube comments, related to news topics.

order: Twitter, YouTube, and news websites. Conversely, lower toxicity scores correspond to a high probability of participants rating the text as formal in the following order: news websites, YouTube, and Twitter (less than 0.5). Uncovering latent attributes can help us further evaluate and understand how these attributes play out differently in each platform. If we were to evaluate the model by this latent attribute, then it clearly performs better on news websites than on Twitter or YouTube.

8.3 Implications for Content Moderation and Social Platforms

Lipton [53] points out that having an AI/ML tool is not the same as having an automated or even assistive system. One needs to create "decision rules" based on the algorithm outputs to put the AI/ML into use. That is, the outputs of the AI/ML tool may be useful, and potentially necessary, but they are not sufficient. Using Perspective as part of an automated or assistive tool requires us to specify appropriate thresholds for the probability scores. If we were using Perspective simply to flag potentially toxic comments for the consideration of a human moderator to evaluate, then we might pick a slightly lower threshold. Likewise, if we were providing feedback to a potential comment contributor (Reference [15], then we might want a lower value simply to provoke thoughtful consideration before the contribution was posted. However, if we were trying to fully automate some aspect of moderation, then we might want to set a very high threshold.

While our evaluation found that Perspective's toxicity scores align with how our participants rated toxicity and disrespectfulness, this transferability creates a type of design challenge. The challenge is that subtle aspects of the discussion topic and social norms of interaction likely influence the selection of an appropriate threshold. Our argument that the ratings produce a type of ordering effect (Section 8.2, immediately above) illustrates one of the potential issues. Any platform looking to adopt Perspective should likely conduct its own evaluation with its own data. Each platform or community should analyze content from their users and norms around user interactions to understand how these factors relate to Perspective's attribute scores. Kumar et al. [51], for instance, explore how ML classifiers such as Perspective can overcome the limitations of one-size-fits-all approach by "personalized tuning" where model thresholds are set to accommodate diverse perspectives of users. Some platforms will more aggressively moderate abuse and harassment, likely because these aspects are more frequently encountered or are more severe [47].

An appropriate evaluation study may reveal different thresholds that will be more suited to the community's specific use cases. These specialized thresholds can then be used to make potential content moderation decisions. We support guidelines from Perspective about not using it for a fully automated moderation system and building human-in-the-loop systems instead [5]. Even if some actions are automated based on Perspective scores (such as adding a warning message to

the comment or hiding it and displaying on demand), having checks in place that are supervised by human moderators can give users an opportunity to appeal and contest decisions. Emulating prior work from Perspective [37], there should also be deliberate efforts to ensure that content moderation features designed around Perspective do not disproportionately marginalize and exclude specific groups.

The character of an online community is shaped by its users and their norms and values, as well as the community's moderation strategies. Though platform stakeholders can initially set down codes of conduct, users can also collaboratively and implicitly construct norms that might take precedence [34]. Community policies are not set in stone. They can, and probably should, change with time and respond to unanticipated issues that further problematize online interactions [47, 57]. Though we are now dependent upon automated approaches to content moderation to address the scale of the problem, a one-size-fits-all approach is unlikely to work [51]. We need to carefully and constantly reevaluate an AI/ML model's predictions along with its social implications.

8.4 Limitations

There are some limitations present in this work, including methodological ones. While 300 is a reasonable comment sample, testing with more data could reveal additional, possibly different, insights. Our dataset was selected by sampling for Perspective scores and using a random sample can also reveal other insights. In this work, we use news-related comments in the English language rated by participants from the U.S. Further analysis with a different population and/or for online comments in different languages that Perspective supports can provide more insights and is important future work.

The difficulty of making "toxicity" judgments creates variance in the data that is not trivially resolved. We addressed this in our analyses by taking the mode or majority rating for each attribute, resulting in loss of information and dismissal of some participants' judgments. An approach to resolve these discrepancies in judgment might leverage open-ended responses along with the ratings to understand how disparate judgments might be resolved.

Our use of anchor comparisons as a way of eliciting such subjective judgments has both benefits and tradeoffs. Absolute ratings can result in the same toxicity rating for both a mildly toxic and a severely toxic comment [27]. Relative ratings or anchor-based ratings can address this drawback and capture degrees of toxicity [27]. However, the use of anchor-based ratings could have also introduced an *anchoring bias* [49]. While we believe that using anchor comparisons generated higher-quality data that are also reflective of participants' understanding of the content and experiences, we do not have empirical evidence for it.

There is another methodological issue in the way these judgments are collected. We believe that what constitutes a "toxic" contribution is also a function of the context [58, 63]. Our method, as in many studies, did not include potentially contextualizing information that might influence human judgments. Potential context that could be included in a future rating task might be, the title or topic of the news story generating the conversation, a set of previous or subsequent comment contributions, and associated imagery like GIFs or JPGs. While we believe that context is important, we are unsure about how it changes the difficulty of the task. Prior work suggests that providing context does affect how people perceive toxicity but it can either increase or decrease the level of perceived toxicity [58].

8.5 Future Work

We believe there is a lot of scope for future work that either builds upon this human-centered evaluation or focuses on integrating Perspective into content moderation systems. Qualitative methods can be used to supplement these findings and provide a first-hand account of how users

feel about interacting with such systems (for example, see comments section in Reference [23]). The moderators are also an important stakeholder and these evaluations can be repeated or adapted to examine if they agree with Perspective. Since they often act as an intermediary between the platform and user, their perspectives are crucial as well. This work illustrates one way to test the transferability of Perspective. It is possible that there are methods that can be used and other platforms that can be tested to investigate transferability more extensively.

Another direction for future work is to compare our annotation techniques to the one that was used to label training data for Perspective. Based on the data in the Kaggle competitions, one can infer that a Likert scale [very toxic, toxic, hard to say, not toxic] has been used to collect annotations for toxicity [14, 21]. These ratings are then converted into an aggregated value that is the fraction of the annotators who consider the comment as toxic or very toxic. There does not seem to be data available about the fraction of annotators who consider the comment as Not toxic and Hard to say. Still, analyzing whether using anchor comparisons to elicit toxicity judgements improves the quality of toxicity annotations compared to using a Likert scale is also an important direction for future work. Such a comparison would inform ongoing research in crowdsourcing annotations for subjective tasks [27].

We could also cross-annotate some of the original training data for formality, respectfulness, or stereotypes to further understand the prevalence of these attributes in the training data and connect concepts like presence of stereotypes to different target identities (which are already annotated in one of the the original training datasets) [14]. Since most of the data used to develop Perspective is open source, one could also use them to train models for toxic speech detection and evaluate or compare them to Perspective.

9 CONCLUSION

Our human-centered evaluation of Perspective showed that high toxicity scores align with human ratings of toxicity and disrespect for news discussions across three different platforms. While disrespect was part of how Perspective defines a toxic comment, formality and stereotypes were not. Yet, Perspective's high toxicity scores correspond to human ratings of the informality of text and presence of stereotypes. Uncovering such latent attributes helps understand model behavior better, question whether such behavior is even desirable from the model, and investigate for latent biases. For example, only if we uncover stereotypes as a latent attribute can we begin to investigate if Perspective considers some stereotypes more toxic than others. Our evaluation is human centered not simply because we examine if users agree with Perspective but also because it surfaces two latent attributes that begin to explain how the model might construe (or misconstrue) comments as toxic. In this work, by empirically investigating Perspective's transferability on different domains, we illustrate one approach to human-centered evaluations of AI/ML models. Such an evaluation offers additional, useful information on how an AI/ML model works and is distinct from the information provided by traditional, technical evaluations. Human-centered evaluations should bring us closer to understanding how such models work in real-world, sociotechnical settings and help us understand a model's strengths, limitations, and biases.

A APPENDIX

A.1 Survey Layout

A.2 Criteria for Accepting/Rejecting Work in AMT

As we describe above, Mechanical Turk was used to collect human ratings of the comment corpus. The collection method included the use of gold standards, but there were also consistency checks on the answers. We defined a set of criteria for approving or rejecting submitted work. We

Table 11. Rejection Criteria for Different Combinations of (1) Inconsistent Gold Standards Rated through Anchor Comparisons, (2) Responses to Gold Standards That Did Not Match the Pre-existing Label, and (3) Inconsistent Responses for Actual Comments through Anchor Comparisons

Total Missed	Inconsistent gold standards	Wrong gold standards	Inconsistent responses	Decision	Data Use
<= 2	0	0	<= 2	Approve	Use
3	0	0	3	Approve	Use
2	0	1	<= 1	Approve	Use
3	0	1	2	Approve	Use
2	0	2	0	Approve	Use
3	0	2	1	Approve	Don't use, warn
4	0	2	>= 2	Reject	
2	1	0	<= 1	Approve	Use
3	1	0	2	Approve	Don't Use, warn
2	1	1	0	Approve	Use
3	1	1	1	Approve	Don't use, warn
3	1	2	= 0	Approve	Don't use, warn
>= 4	1	2	>= 1	Reject	
2	2	0	0	Approve	Don't use, warn
3	2	0	1	Approve	Don't use, warn
4	2	0	2	Reject	
3	2	1	0	Approve	Don't use, warn
>= 4	2	1	>= 1	Reject	

The first column measures the total number of invalid responses in a total of 15 comments.

recognize MTurk workers are in a hurry and might make mistakes while still behaving in good faith. Therefore, our threshold for rejecting work was set quite low. We evaluated participants' performance on gold standards and checked if their responses were consistent with respect to the provided anchors. We took one of the three decisions based on the below table: (1) approved participants' work and used the data, (2) approved participants' work and did not use the data, and (3) reject the participants' work and did not use the data.

A.3 Finding the Majority Rating for Comments

We aimed to collect ratings from five different participants for each comment. Excluding inconsistent ratings led to working with three or four valid ratings per comment rather than five ratings. We still used the majority rating to resolve different ratings from participants. As Table 12 shows, for most of the observations (266), we were able to find the majority rating by 60% or higher for the toxicity attribute.

We had a few cases where ratings were split between two participants or two groups of participants such that there was no majority rating. The table below (Table 13) shows how we managed a tie-breaker for such cases.

For most of the observations, we were able to find the majority rating by 60% or higher for the formality, respectfulness, and stereotype attributes as well.

In cases where formality, respectfulness, and stereotype attribute ratings were split between two participants or two groups of participants such that there was no majority rating, the below table shows how we managed tie-breaking.

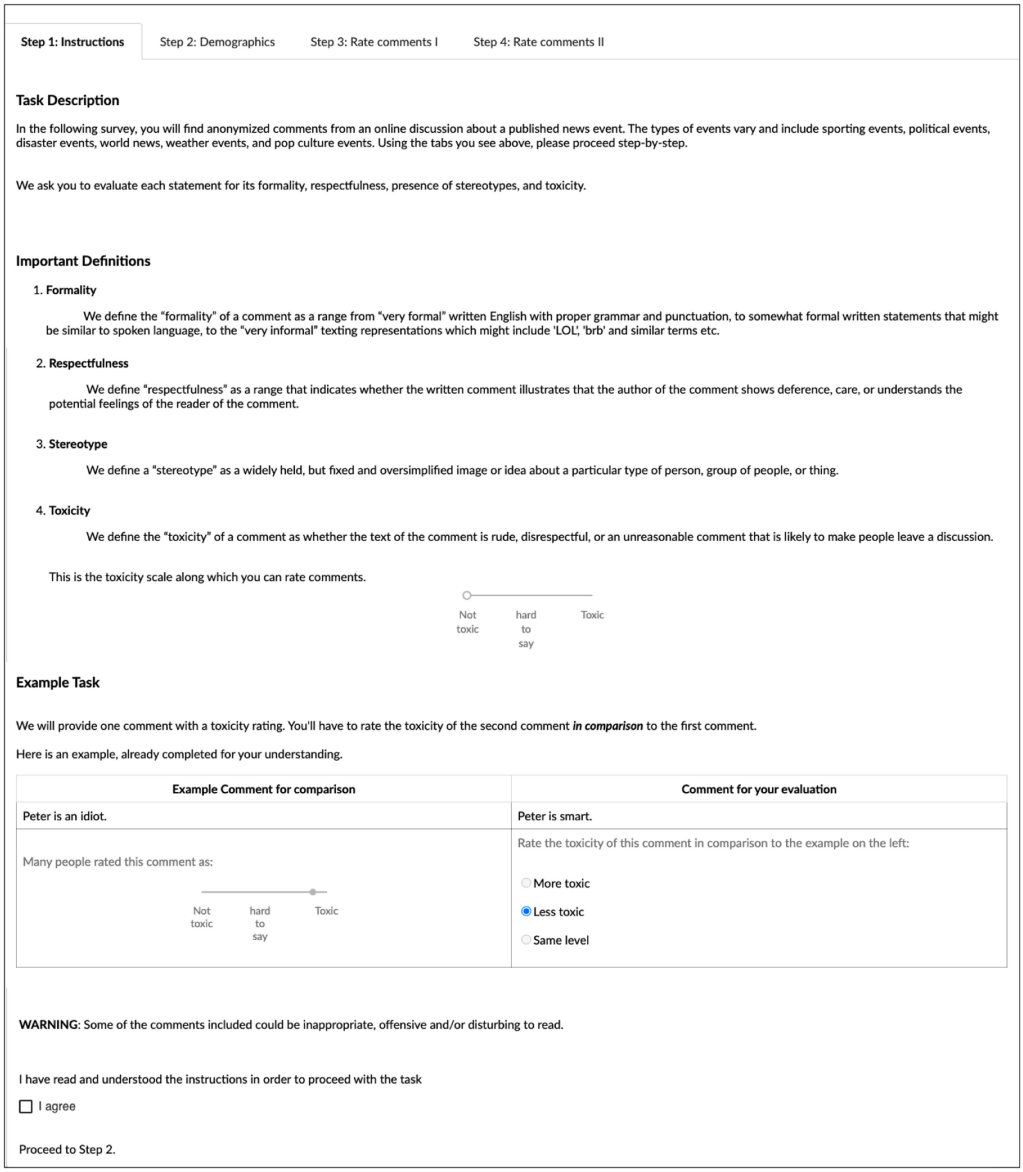

Fig. 15. Our task interface consists of four tabs, with each tab representing a step in the study. This allowed us to set expectations for the participants on the task content and duration. We included the definitions for formality, respectfulness, stereotype, and toxicity in steps 1, 3, and 4 to allow users to reference throughout the task.

Table 12. The Level of Agreement and the Number of Data Points That Correspond to That Agreement

Agreement between participants	Toxicity Attribute - No.of comments
100% (5/5 or 4/4 or 3/3)	76 (62 + 11+ 3)
80% (4/5)	71
75% (3/4)	15
66.66% (2/3)	7
60% (3/5)	97
50% (2/4 or 1/2)	22 (19 + 3)
40% (2/5)	12

Table 13. This Table Details How We Resolved Ties between Different Toxicity Ratings

Categories across which responses were split	Count	Resolved by
['Hard to say' 'Toxic']	15	Collapse to Toxic if the tie is in between Toxic and Hard to say
['Hard to say' 'Not toxic']	8	Collapse to Not Toxic if the tie is in between Not Toxic and Hard to say
['Not toxic' 'Toxic']	3	Collapse to Hard to Say if the tie is in between Not Toxic and Toxic

Table 14. The Level of Agreement and the Number of Data Points That Correspond to That Agreement for Each Attribute

Agreement between participants	Formality Attribute - No. of comments	Respectfulness Attribute - No. of comments	Stereotype Attribute - No. of comments
5/5	24	53	48
4/5	88	65	57
3/5	145	122	117
2/5	43	60	78

Table 15. This Table Details How We Resolved Ties for Each Attribute

Attribute	How it was resolved
Formality	Formal if tie is in between Formal and Very Formal Informal if tie is in between Informal and Very informal Delete others
Respectfulness	Disrespectful if tie between disrespectful & slightly disrespectful Respectful if tie between respectful & partially respectful Delete others
Presence of stereotypes	Possibly Present if tie between Possibly Present & Present Present if tie between Possibly Present & Heavily Present Heavily Present if tie between Heavily Present & Present

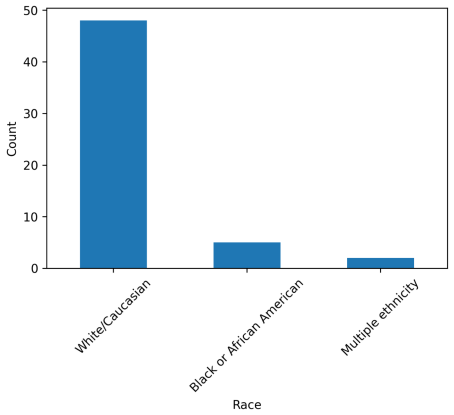

Fig. 16. A bar graph showing racial demographic information.

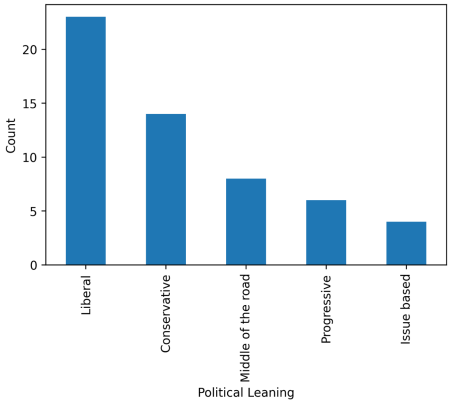

Fig. 17. A bar graph showing the political leaning of different participants.

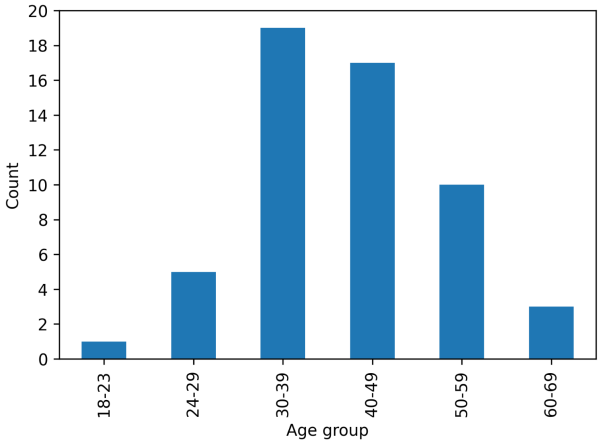

Fig. 18. A bar chart that shows information about the different age groups of study participants.

A.4 Participant Demographic Information

Our survey also collected the following demographic information with the possible choices listed for each demographic dimension:

(1) Age: 18–23, 23–29, 30–39, 40–49, 50–59, 60–69
(2) Education: High school or equivalent (e.g., GED), Enrolled for Associate's degree, Enrolled for Bachelor's degree, Some college no degree, Associate's degree, Bachelor's degree, Graduate degree
(3) Gender: Male, Female, Transgender (M->F), Transgender (F->M), Gender non-conforming, Agender, Bigender, Prefer not to say
(4) Race: Multiple ethnicity, American Indian or Alaskan Native, Asian/Pacific Islander, Black or African America, Hispanic, White/Caucasian
(5) Political Leaning: Liberal, Conservative, Middle of the road, Progressive, Issue-based

Below are the bar charts of participant responses to the demographic questions. We have not listed the response categories that received zero responses for the given demographic dimension.

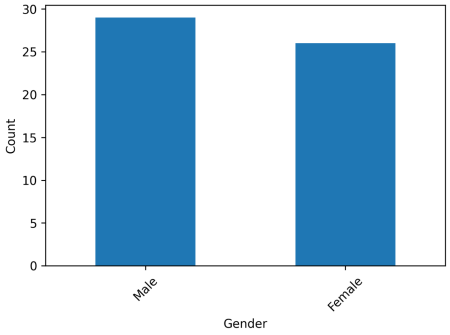

Fig. 19. A bar graph showing gender demographics.

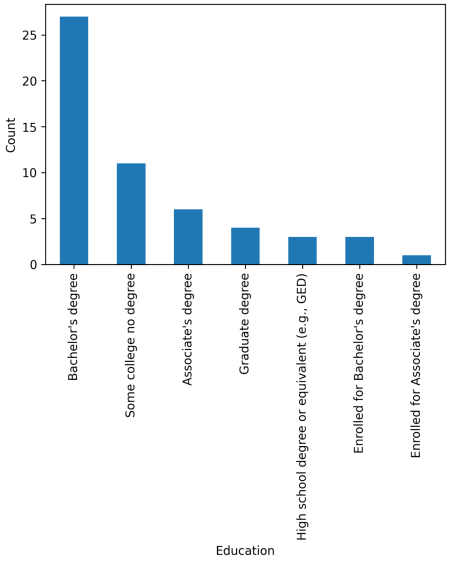

Fig. 20. A bar graph showing the educational level of different participants.

Table 16. A List of Comments that Received (1) Informal/Very Informal Ratings from Participants and Low Toxicity Scores from Perspective and (2) Formal Ratings from Participants and High Toxicity Scores from Perspective

Toxicity Score (from Perspective)	Formality rating (from participants)	Comment Text
0.08	Informal	I enjoyed this video SO Much!!! Great ways that we can take care of each other.
0.09	Informal	Sooo business will re-open BUT IF FOLKS AIN'T GOT NO MONEY THEN WHAT?
0.12	Very Informal	Seattle politicians r giving a Master Course in how 2 destroy a once beautiful, thriving city. My best friend loved it &was driven out after his 8 yr old daughter kept getting accosted at stores, etc, by homeless predators & shop owners & cops shrugged their shoulders in response
0.28	Very informal	Say goodbye to us middle class!! Get ready for war ppl when poor and middle class clash with the rich and political parties!! Soon it either ur rich or poor!! Stay lock and loaded protect your luv ones!!
0.72	Formal	He has said the exact truth. I'm so glad he did. The media were doing exactly what he said. They made a big thing of it accusing him of having brain damage, etc. What a bunch of rats.
0.86	Formal	I'm so sick to see how our people are treated and how our kids are traumatized by these ignorant bigots.
0.90	Formal	Two 12 year old boys just completed a hunter's training course days before one of them pointed a gun at the other kid's chest and shot him dead. Training means nothing with most of these idiots regardless of age.
0.93	Formal	I guess millions of people out of work and thousands losing health insurance and dead from this pandemic are not important to this idiot.

A.5 Comparing Formality Ratings and Toxicity Scores

In this Appendix section, we offer examples to support findings in the Section 7.5.

Table 17. A List of Comments That Received (1) No Presence of Stereotypes Ratings from Participants and High Toxicity Scores from Perspective and (2) Presence of Stereotype Ratings from Participants and Low Toxicity Scores from Perspective

Toxicity Score (from Perspective)	Stereotype rating (from participants)	Comment Text
0.04	Possibly present	For a sports star to do that, lots of luck had to be in play. Sports stars are known for squandering or wasting away all their money. There are exceptions of course, M. Jordan & Larry Bird.
0.05	Present	So, during this worldwide transfer of wealth, the rich and elite are going to help the masses. Help us to our graves is more like it.
0.15	Present	Michigan should just go ahead and move to the South where they belong.
0.21	Possibly Present	Young people house partying typically means destruction. Not sure she knowingly rented it out for that.
0.88	Not present	Why does this crackpot get primo media space? Most of the black people I know (like most of the white people and most of the Asians I know) think he's a complete idiot. Does he really rank headlines?
0.90	Not present	Good! No one wants to hear his crap! Shut it off for good! One depressing individual. No one wants to be subject to negative lying statements from this person. Good f n bye!
0.92	Not present	Well, people should have stayed their ass in the damn house like the doctor said.
0.95	Not present	imagine that.... their idiot governor who encourages dismissing the virus is complicit in manslaughter.

ACKNOWLEDGMENTS

We thank the reviewers for their thoughtful comments and feedback.

REFERENCES

[1] Perspective API. [n.d.]. About the API. Retrieved November 3, 2022 from https://developers.perspectiveapi.com/s/about-the-api-training-data?language=en_US.
[2] Perspective API. [n.d.]. About the API—Attributes and Languages. Retrieved April 8, 2021 https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages.
[3] Perspective API. [n.d.]. About the API—Best Practices & Risks. Retrieved April 2, 2021 from https://developers.perspectiveapi.com/s/about-the-api-best-practices-risks.
[4] Perspective API. [n.d.]. About the API–FAQs. Retrieved April 2, 2021 from https://developers.perspectiveapi.com/s/about-the-api-faqs.
[5] Perspective API. [n.d.]. About the API–Model Cards. Retrieved November 3, 2022 from https://developers.perspectiveapi.com/s/about-the-api-model-cards?language=en_US&tabset-20254=2.
[6] GitHub. [n.d.]. Conversation AI. Retrieved from April 2, 2021 from https://github.com/conversationai.
[7] GitHub. [n.d.]. Conversationai.github.io/toxicity_with_subattributes.md at Main. Retrieved April 2, 2021 from https://github.com/conversationai/conversationai.github.io/blob/main/crowdsourcing_annotation_schemes/toxicity_with_subattributes.md.
[8] [n.d.]. Developers. https://developers.perspectiveapi.com/s/?language=en_US. (Accessed: March 14, 2023).
[9] The Verge. [n.d.]. Facebook Is Now Using AI to Sort Content for Quicker Moderation. Retrieved September 7, 2021 from https://www.theverge.com/2020/11/13/21562596/facebook-ai-moderation.
[10] Wikipedia. [n.d.]. Gamergate Controversy. Retrieved April 14, 2021 from https://en.wikipedia.org/wiki/Gamergate_controversy.
[11] Wikipedia. [n.d.]. Goldilocks Principle. Retrieved April 8, 2021 from https://en.wikipedia.org/w/index.php?title=Goldilocks_principle&oldid=985929960.
[12] Jigsaw. [n.d.]. Hello Neighbor: How a Local Newspaper Builds Community via Online Comments. Retrieved May 5, 2022 from https://medium.com/jigsaw/hello-neighbor-how-a-local-newspaper-builds-community-via-online-comments-c6a0bbfccb3b.
[13] [n.d.]. Interpreting Reliability Results. Retrieved April 8, 2021 from https://homepages.inf.ed.ac.uk/jeanc/maptask-coding-html/node23.html.
[14] Kaggle. [n.d.]. Jigsaw Unintended Bias in Toxicity Classification. Retrieved May 27, 2022 from https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification/data.

[15] OpenWeb. [n.d.]. Nudge Theory Examples In Online Discussions. Retrieved May 5, 2022 from https://www.openweb.com/blog/openweb-improves-community-health-with-real-time-feedback-powered-by-jigsaws-perspective-api.
[16] Science. [n.d.]. Overzealous Profanity Filter Bans Paleontologists from Talking about Bones. Retrieved September 7, 2021 from https://www.theguardian.com/science/2020/oct/16/profanity-filter-bones-paleontologists-conference#:~:text=Participants%20in%20a%20virtual%20paleontology,beaver%20%E2%80%93%20during%20an%20online%20conference.
[17] Perspective API. [n.d.]. Retrieved April 2, 2021 from https://perspectiveapi.com/#/home.
[18] Perspective API. [n.d.]. Case Studies. Retrieved November 28, 2022 from https://www.perspectiveapi.com/case-studies/.
[19] Perspective API. [n.d.]. Research into Machine Learning. Retrieved November 3, 2022 from https://www.perspectiveapi.com/research/.
[20] Reddit. [n.d.]. Reddit Ban Endangered Thousands of Lives (re: r/ProED): TrueOffMyChest. Retrieved April 2, 2021 from https://www.reddit.com/r/TrueOffMyChest/comments/9xa1dt/reddit_ban_endangered_thousands_of_lives_re_rproed/.
[21] Meta. [n.d.]. Research: Detox/Data Release. Retrieved November 18, 2022 from https://meta.wikimedia.org/wiki/Research:Detox/Data_Release.
[22] Pew Research Center. [n.d.]. The State of Online Harassment. Retrieved April 2, 2021 from https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/.
[23] The New York Times. [n.d.]. The Times Sharply Increases Articles Open for Comments, Using Google's Technology. Retrieved April 12, 2021 from https://www.nytimes.com/2017/06/13/insider/have-a-comment-leave-a-comment.html#commentsContainer.
[24] Buzzfeed. [n.d.]. Tumblr's New Anti-Porn Algorithm Is Flagging Non-Pornographic Content. Retrieved April 4, 2021 from https://www.buzzfeednews.com/article/krishrach/tumblr-porn-algorithm-ban.
[25] Google. [n.d.]. Tune (Experimental)—Chrome Web Store. Retrieved May 5, 2022 from https://chrome.google.com/webstore/detail/tune-experimental/gdfknffdmmjakmlikbpdngpcpbbfhbnp?hl=en.
[26] Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd International acm Sigir Conference on Research and Development in Information Retrieval*. 45–54.
[27] Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel Rosen. 2019. Crowdsourcing subjective tasks: the case study of understanding toxicity in online discussions. In *Companion Proceedings of the World Wide Web Conference*. 1100–1105.
[28] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and its consequences for online harassment: Design insights from heartmob. *Proc. ACM Hum.-Comput. Interact.* 1, Article 24 (2017), 1–19. https://doi.org/10.1145/3134659
[29] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of the World Wide Web Conference (WWW'19)*. Association for Computing Machinery, New York, NY, 491–500. https://doi.org/10.1145/3308560.3317593
[30] Agostina Calabrese, Michele Bevilacqua, Björn Ross, Rocco Tripodi, and Roberto Navigli. 2021. AAA: Fair evaluation for abuse detection systems wanted. In *Proceedings of the 13th ACM Web Science Conference (WebSci'21)*. Association for Computing Machinery, New York, NY, 243–252. https://doi.org/10.1145/3447535.3462484
[31] Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. # thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1201–1213.
[32] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A cross-community learning-based system to assist reddit moderators. *Proc. ACM Hum.-Comput. Interact.* 3, (2019), 1–30.
[33] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proc. ACM Hum.-Comput. Interact.* 1, (2017), 1–22.
[34] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proc. ACM Hum.-Comput. Interact.* 2, (2018), 1–25.
[35] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. The bag of communities: Identifying abusive behavior online with preexisting internet data. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 3175–3187.
[36] Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the Chi Conference on Human Factors in Computing Systems*. 1–14.

[37] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 67–73.
[38] Jessica L. Feuston, Alex S. Taylor, and Anne Marie Piper. 2020. Conformity of eating disorders through content moderation. *Proc. ACM Hum.-Comput. Interact.* 4, 1 (2020), 1–28.
[39] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transact. Inf. Syst.* 14, 3 (1996), 330–347.
[40] Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2023. Handling bias in toxic speech detection: A survey. *ACM Comput. Surv.* Just Accepted (January 2023). https://doi.org/10.1145/3580494
[41] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press.
[42] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data Soc.* 7, 1 (2020), 2053951719897945.
[43] Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N. Asokan. 2018. All you need is "love" evading hate speech detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*. 2–12.
[44] Erin R. Hoffman, David W. McDonald, and Mark Zachry. 2017. Evaluating a computational approach to labeling politeness: Challenges for the application of machine classification to social computing data. *Proc. ACM Hum.-Comput. Interact.* 1, (2017), 1–14.
[45] Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving google's perspective api built for detecting toxic comments. arXiv:1702.08138. Retrieved https://arxiv.org/abs/1702.08138.
[46] Jordan S. Huffaker, Jonathan K. Kummerfeld, Walter S. Lasecki, and Mark S. Ackerman. 2020. Crowdsourced detection of emotionally manipulative language. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–14.
[47] Jialun "Aaron" Jiang, Skyler Middler, Jed R. Brubaker, and Casey Fiesler. 2020. Characterizing community guidelines on social media platforms. In *Conference Companion Publication of the Computer Supported Cooperative Work and Social Computing (CSCW'20 Companion)*. Association for Computing Machinery, New York, NY. 287–291.
[48] Prerna Juneja, Deepika Rama Subramanian, and Tanushree Mitra. 2020. Through the looking glass: Study of transparency in reddit's moderation practices. *Proc. ACM Hum.-Comput. Interact.* 4, (2020), 1–35.
[49] Daniel Kahneman. 2003. A perspective on judgment and choice: Mapping bounded rationality. *Am. Psychol.* 58, 9 (2003), 697.
[50] Gary King, Michael Tomz, and Jason Wittenberg. 2000. Making the most of statistical analyses: Improving interpretation and presentation. *Am. J. Pol. Sci.* (2000), 347–361.
[51] Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing toxic content classification for a diversity of perspectives. In *Proceedings of the 17th Symposium on Usable Privacy and Security (SOUPS'21)*. 299–318.
[52] J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* (1977), 159–174.
[53] Zachary C. Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.
[54] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 220–229.
[55] Tanushree Mitra, Clayton J. Hutto, and Eric Gilbert. 2015. Comparing person-and process-centric strategies for obtaining quality data on amazon mechanical turk. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1345–1354.
[56] Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. arXiv:1808.07231 (2018).
[57] Jessica A. Pater, Moon K. Kim, Elizabeth D. Mynatt, and Casey Fiesler. 2016. Characterizations of online harassment: Comparing policies across social media platforms. In *Proceedings of the 19th International Conference on Supporting Group Work*. 369–374.
[58] John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 4296–4305. https://doi.org/10.18653/v1/2020.acl-main.396
[59] Bernhard Rieder and Yarden Skop. 2021. The fabrics of machine moderation: Studying the technical, normative, and organizational structure of Perspective API. *Big Data Soc.* 8, 2 (2021), 20539517211046181.
[60] Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. Hatecheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 41–58. https://doi.org/10.18653/v1/2021.acl-long.4

[61] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 59–68.

[62] Douglas K. Van Duyne, James A. Landay, and Jason I. Hong. 2007. *The Design of Sites: Patterns for Creating Winning Web Sites*. Prentice Hall Professional.

[63] Alexandros Xenos, John Pavlopoulos, and Ion Androutsopoulos. 2021. Context sensitivity estimation in toxicity detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH'21)*. Association for Computational Linguistics, 140–145. https://doi.org/10.18653/v1/2021.woah-1.15

Received 17 December 2021; revised 15 December 2022; accepted 12 January 2023