

Toxicity Detection and Mitigation on Social Networking Platforms

BRIAN TANG*, SHUBHAM ATREJA*, DONALD LOVELAND*, ZIHAN WU*, and RENEE LI*, University of Michigan, USA

The problem of identifying, managing, and remedying toxic content has become prevalent in online communities. If toxic language can be mitigated in these settings, online communities can act in a more inclusive and fair manner, facilitating better discussions and more helpful answers. In practice, content moderation has been a largely manual process, requiring moderators to individually analyze each reported instance of toxic behavior. To overcome this laborious and potentially biased process, we consider an alternative solution where a software tool, powered by a toxicity detection algorithm learned through large amounts of data, can both identify toxic language as it is typed, on the user-end, and suggest alternative language. To determine the usefulness of such a tool, we gathered data from questionnaires interviews centered around users' experiences regarding toxic content in online platforms. We scoped our tool to more broadly capture online forums that share anonymity as a common feature. We also further refine the scope of our tool's capabilities to focus primarily on notifying user's of perceived toxicity in their text. Contextual interviews were performed to further understand user requirements. After developing a wizard-of-oz prototype, we recruited three users and performed simplified user testing, and conducted heuristic evaluation with one usability expert. After improving our system and iterating on user requirements, we conducted quantitative user evaluations against two user requirements with ten participants.

CCS Concepts: • **Human-centered computing** → **Interaction techniques**.

Additional Key Words and Phrases: text entry, language, toxic, detection, piazza.

ACM Reference Format:

Brian Tang, Shubham Atreja, Donald Loveland, Zihan Wu, and Renee Li. 2021. Toxicity Detection and Mitigation on Social Networking Platforms. In *EECS 598: Human-Computer Interaction (HCI), Fall 2021, August 30–December 15, 2021, Ann Arbor, MI*. ACM, New York, NY, USA, 75 pages. <https://doi.org/10.1145/1122445.1122456>

1 INTRODUCTION

Toxicity has become a significant problem in many online communities and platforms. Prior research shows that toxicity is on the rise even in professional spaces like online college communities [18] and technical forums [4]. When these communities are directly managed by the users (e.g. Piazza, StackOverflow), it is up to community moderator or administrator to take appropriate action against toxic content. Since a manual review can take some time, there is some delay before toxic content gets flagged and removed. Research also suggests that exposure to toxicity can increase the toxicity in subsequent comments [10], which can negatively impact the engagement and growth of a community [12]. By developing a method to circumvent the need for manual review, toxic language can be more quickly addressed and even removed from the platform before it has a chance to disrupt another user's experience, promoting a more fair and equitable space.

With the need for more proactive intervention mechanisms, researchers have searched for methods that instead rely on the user to perform self-moderation. For instance, Twitter has launched an initiative that prompts users to reconsider their response before posting if it contains potentially “harmful” language [2]. Twitter claims that their initiative led to 34% of people revising their initial response. While results seem promising, the initiative is limited to one platform and still requires that users responses be shared with the platform, even if they are not visible to other users.

* All authors contributed equally to this research.

Inspired by the success of that initiative, we focus on modeling toxicity-reduction in online communities as a text-entry problem, that can be solved at the level of individual users. We envision a system that can proactively nudge users away from toxic language, as they are typing out their content. Remedying this behavior instantaneously at the user-level, rather than in a post-hoc fashion through moderators, can significantly improve how quickly toxic language is handled.

Furthermore, the intervention on Twitter simply prompted users to reconsider their text without providing any additional information or guidance. Taking a step further, we may also consider recent advances in style transferring [5, 15] and paraphrasing approaches [6] to suggest edits in the users' content to reduce the toxicity, actively showing the user positive behavior examples. We intend to generate more insights on the exact scope of our intervention through further contextual inquiry while also considering other prominent concerns in technology such as privacy, transparency, and trustworthiness.

For our initial research, we focus on the online learning community, Piazza, which is focused on providing a question and answer platform for individual college courses. While prior research underlines the growing popularity of Piazza and the benefits of using Piazza discussion boards [7], the presence of toxicity, and how that may impact a user's experience on Piazza has not been explored. We take a user-centered design approach, including surveys and contextual inquiry to understand the context of use. The final design is also informed by multiple rounds of analysis and prototyping. Looking at our data, we find that both survey and interview participants reported that toxic content is rarely seen on Piazza. Therefore, we decided to shift our focus to online toxic content interventions that are independent of any platform. From our survey results, we also observe alignment on how users prefer to be notified of their potentially toxic behaviors and their general attitude of distrust towards toxicity detection models. These insights inform the design of future systems, specifically the need to build a transparent toxicity detection system. In the phase of contextual interview, we collected interpretations from users' experiences in three scenarios regarding toxic content online: reading toxic content, intentionally posting toxic content, and unintentionally posting content that drove people away from discussion. Next, we consolidate the data using sequence diagrams, flow charts, and affinity diagrams to further identify user requirements. We identified 7 user requirements for the comprehensive scenario of reading and posting toxic language online. For instance, users want the ability to send, edit, or delete their message despite the circumstances of the toxic content, and the intervention mechanism should be flexible enough to accommodate the different reactions from different individuals. We then design different prototypes keeping in mind the different requirements.

To evaluate and improve our design, we first developed a low-fidelity prototype which allows wizard-of-oz simplified user testing. We conducted simplified user testing of five interaction goals with three participants, and discovered that there are certain interfaces and interactions could be improved. For example, our initial design for the user options is unclear to some users, and the interaction to reveal toxic content is not intuitive. We also demonstrated our system to a usability expert whom provided heuristic evaluation of our system.

2 RELATED WORK

As toxicity becomes a major problem in online communities, it is no longer restricted to contentious topics or contexts. Prior research has found toxic behavior in college communities [18], technical forums like Stack Overflow [4] and even in GitHub issues on open-source projects [17]. In professional contexts, such as GitHub, toxicity can worsen work-related stress and burnout that members may already be experiencing, and adversely affect their productivity [17].

Specific to Piazza, toxicity has not been a topic of prior research. Initially, the first step in our project is to discover whether toxicity is a concern on Piazza or not. It can be argued that some of Piazza's features, such as a discussion board that allows anonymous contribution, make it susceptible to toxic behavior [7]. Prior research

finds that online anonymity can contribute to toxic behavior [3]. At the same time, the prevalent norms on Piazza, which are often dictated by the norms of classroom learning may discourage such behavior, as prior research also finds that norms prevalent in an online community can both discourage and encourage toxic behavior [16]. After performing a survey and multiple contextual inquiries, we failed to attain enough information specific to Piazza. Thus, we have broadened the scope to text-entry in online discussion boards and forums.

When exploring participants’ experiences of toxicity on a platform, how we define toxicity may impact the experiences they describe. And the notion of toxicity itself can be subjective, where different people consider different things to be toxic. We use the definition of toxicity provided by the Perspective API [1], as it defines toxicity by focusing on the outcome itself. Specifically, toxicity is defined as “a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion.” Focusing on participation as an outcome is essential on forums and discussion board platforms such as Reddit, GitHub, NextDoor, etc.

Given the severity of online toxicity, different models have been designed to tackle toxicity in online communities [1, 9, 14]. For instance, Google has released their Perspective API that uses machine learning to score the perceived impact a comment might have on a conversation, as a way to detect toxicity [1]. Noever further presents a comparison between different approaches to toxicity detection, while also highlighting the transparency and explainability of the rules across different models. In the context of our design, it remains to be seen the extent of transparency and explainability expected by the users with regard to toxicity detection.

Beyond just detecting toxicity, prior research has also explored proactive approaches to discourage toxic behavior. In an experiment on Twitter, an intervention was introduced where a bot account would reply to users who posted offensive tweets – highlighting the harmful nature of their content and sanctioning their behavior [13]. Users who were subjected to the intervention displayed less toxic behavior over time. While the approach shows the promise of nudging users away from toxic behavior, the intervention happened after toxic language was used and after other users were already exposed to the toxic content. Along similar lines, Twitter has developed tools capable of asking users to reconsider their response before posting, if it contains potentially “harmful” language. Twitter claims that following this intervention, 34% of people revised their initial response [2]. The intervention further led to 11% fewer offensive replies in the future by those who were subject to the intervention. While this intervention occurs before the toxic content gets posted, their design is platform-specific and likely requires that user content be sent to their servers first. We envision our tool as a browser plugin which can be configured across multiple websites and doesn’t rely on the users’ original content being posted to the platform servers.

As an alternative approach, researchers have also explored whether textual content can be modified without entirely changing its meaning. For instance, Prabhunoye et al. developed a model to change the sentiment of a text input without affecting the intended meaning [15]. Santos et al. gathered a dataset from Reddit to demonstrate how style transfer can make offensive content less offensive [5]. Fu et al. uses paraphrasing techniques to make sentences more polite and demonstrates the effectiveness of the method in translated communication as well as when the perceptions of the poster and the reader may be misaligned [6]. These approaches demonstrate the feasibility of designing interventions that can adjust user’s text based on a specified characteristic, such as toxicity, without changing the user’s intended meaning. However, it remains to be seen if such interventions that suggest edits to the users’ text are considered intrusive by the users, and more generally, what is the extent to which users can be nudged away from typing toxic content.

3 ESTABLISHING FOCUS: INITIAL SURVEY

We created an 18 question survey to understand peoples’ attitudes and experiences towards online toxicity. In addition, we also used this survey to understand people’s beliefs about toxicity moderation online. From these responses, we began to piece together how experience has driven people’s perspectives, as well as mitigation

strategies that might work best for users of the platform. We distributed the survey virtually via Qualtrics to a total of 53 participants, or 50 responses after filtering for response quality.

3.0.1 Method. To better understand user needs, we administered an 18 question virtual survey on the Qualtrics platform soliciting experiences and attitudes towards toxic language online. We designed the questionnaire to cover three focuses: user’s perception on toxic language, user’s explicit experience with toxic language, and user’s thoughts on the moderation of toxic language. The first focus provides us with a foundational understanding of the importance a user places on non-toxic environments. The second focus provides examples of how past toxic experiences have affected the user as well as how they have engaged with past toxic behavior. We intend to use this information to determine specific groups within our target audience who may benefit more from a moderation tool. The final moderation focus serves to enlighten our options and methods for correct toxicity, given the possible drawbacks that may occur such as privacy or transparency. All but two questions utilize the Likert scale for responses, or some variant of the Likert scale. One question, which simply asks if the users have previously used piazza, only requires a yes or no answer, while another which asks for a preferred mode of intervention, lists three possible interventions. [11].

3.0.2 Tasks and Procedures. The researchers conducted five separate pilot runs prior to finalizing the questions. These pilots were done with another individual, in person, to better understand not only how the person would answer a question, but any complications that arose when answering. During the pilot, we observed the length of the original questionnaire to be too long, usually extending closer to ten minutes per person. To remedy this issue, we limited the questionnaire to 18 questions, two of which were simple demographic based questions. Likewise, we worked to make each question have a similar answering scheme, allowing the participants to more rapidly answer the questions. In addition to length, the questionnaire’s original set of questions lacked a specific use case, causing the questions to feel vague and abstract. To ground the responses in reality, participants were instructed to recall their prior experiences with toxicity in an educational online question and answer spaces known as Piazza. Lastly, participants noted a lack of questions that gauged the user’s attitude on toxicity interventions, which was important when thinking about how someone would be moderated. In a particular instance, the participants from the pilot reported that their attitude towards a suggestion of less toxic language would significantly depend on the source of the feedback. Specifically, they would be more likely to feel defensive if the suggestion came from another user with an opposing view, while their attitude would be positive if the suggestion came from a neutral user or an objective system. As a result we have modified the questionnaire as necessary to optimize clear responses from participants. Toxic language high potential to be a sensitive topic for individuals who have experienced abuse. Given the sensitivity of our research topic, we provided a clear description of research goals and privacy protections to ensure informed consent. Prior to beginning the questionnaire, all participants are provided with the purpose, risks, and benefits, and point of contact of this survey. Participants provide consent by explicitly continuing with the survey following reading all the information.

In total, 53 participants started the survey. We used a number of methods to ensure data quality. First, we removed all 3 incomplete responses from the data set. Of 50 complete responses, the median time to complete the survey was 189 seconds - about 3 minutes. We removed two extremely high outliers (approaching one hour for response time, presumably the users left the survey open before responding). After all processing, the average time to complete the survey was 271 seconds, or about 4.5 minutes per participant. This is noted to be significantly less than the roughly 10 minute completion time seen on the pilot.

3.0.3 Participants. The survey was sent out to friends and family with relevant background, as well as various tech focused social media channels. Quality was maintained by restricting the survey to groups that we have spent considerable time (6+ months) as a simple vetting process. We collected the participants’ current education experience in order to better understand if individuals were in school and if so, their current year. This was

useful as it helped determine whether someone would likely be exposed to Piazza through coursework, or if their responses should be considered in a more platform agnostic fashion. Of 50 completed responses, 17 participants are not currently enrolled in school, 2 are in high school, 3 are undergraduate students, 7 are masters students, and 21 are PhD students. We did not collect gender, age, or racial data from the participants as we did not intend to use this information when designing our tool. Specifically, we assume that this information would likely be unavailable to the tool once it is actually deployed, making it difficult to necessitate the data in the design phase. That said, we do recognize that toxicity can disproportionately impact specific gender, age, or race, and has the potential to improve toxicity monitoring. The study was conducted virtually and did not offer any financial incentives for participants.

3.0.4 Results. When analyzing the data, we convert each of the Likert scale values into ordinal variables denoted one to five. For the yes or no question, answers are denoted as 1 and 0 respectively, while the intervention preferences are processed through an ordinal encoding between 1 and 3 (where 1 is just a notification, 2 is a suggestion, and 3 is an automatic change). We first note that half of the respondents have not used Piazza during their time in school, and for those who have, most have not witnessed toxic language usage on the platform (fig. 2). That said, there are a small group who do experience some toxicity on Piazza, which we believe indicates the platform can still improve in terms of equity. We also notice that while the impact of being affected by toxic language varies by user, (fig. 6) those less affected by toxic language tend to prefer a tool with more intervention. This counters the general trend seen in (fig. 5) where users prefer a tool which simply notifies rather than one that offers text suggestion or automatic changes. When considering who might use a tool, we note most users would not trust a tool to determine whether language is toxic potentially due to both privacy concerns and the required transparency that comes with it (fig. 7). Another interesting insight stems from participants who have been negatively affected by toxic language in the past - these participants do not believe pointing out toxic behavior would dissuade future usage of toxic language. We hypothesize this could be indicative of either a) users simply believing that toxicity is commonplace and people will continue to use toxic behavior no matter the consequences (e.g. "trolls" online who use toxic language for fun), or b) users know moderating systems currently exist, yet, they do not seem to work given their personal exposure to toxicity. We will need to further dissect this finding during our contextual inquiries to determine which components of the two cases are true. Our last set of discoveries comes from participants who have used toxic language against others in the past. These users tend to be more defensive about their toxicity and less willing to change their behavior (figs. 9 and 10), suggesting our tool will mainly focus on notifying users who unintentionally use toxic language against others.

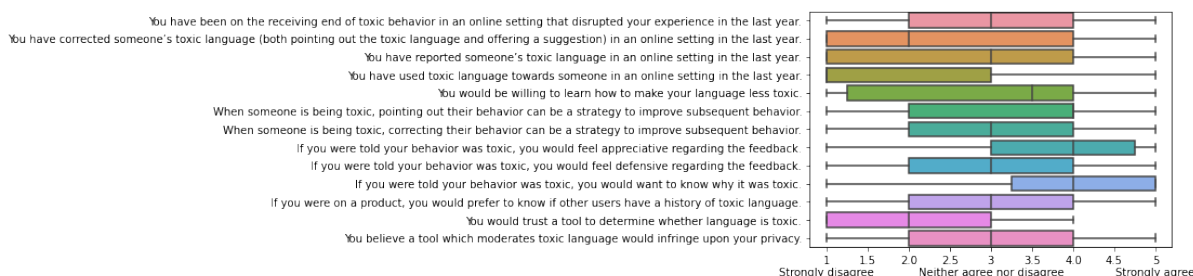


Fig. 1. Box and whisker plot to questions assessing users' experience and attitudes towards toxic language usage online

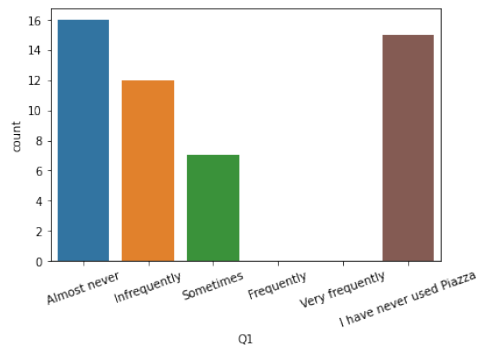


Fig. 2. "How often have you witnessed toxic language on Piazza?" Labels are in order of Likert scale: strongly disagree, somewhat disagree, neither agree or disagree, somewhat agree, strongly agree

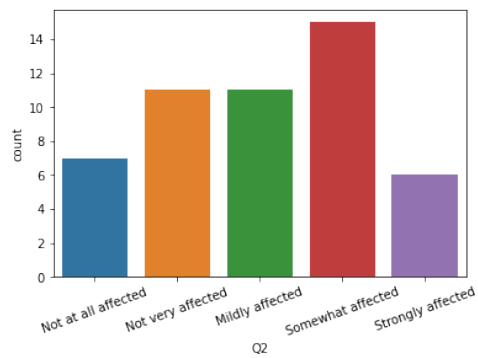


Fig. 3. "How strongly would you be affected by toxic language if the toxic language was directed at you?" Labels are in order of Likert scale: strongly disagree, somewhat disagree, neither agree or disagree, somewhat agree, strongly agree

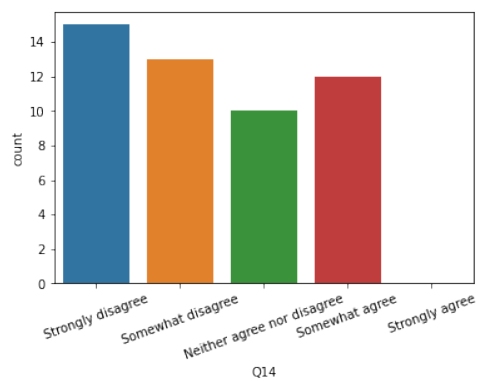


Fig. 4. "You would trust a tool to determine whether language is toxic."

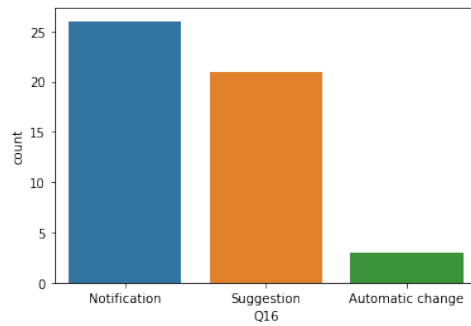


Fig. 5. "Imagine you just typed potentially toxic language, what would you prefer to happen?" 1-Notification of toxic language 2-Suggestion of toxic language 3-Automatic correction of toxic language

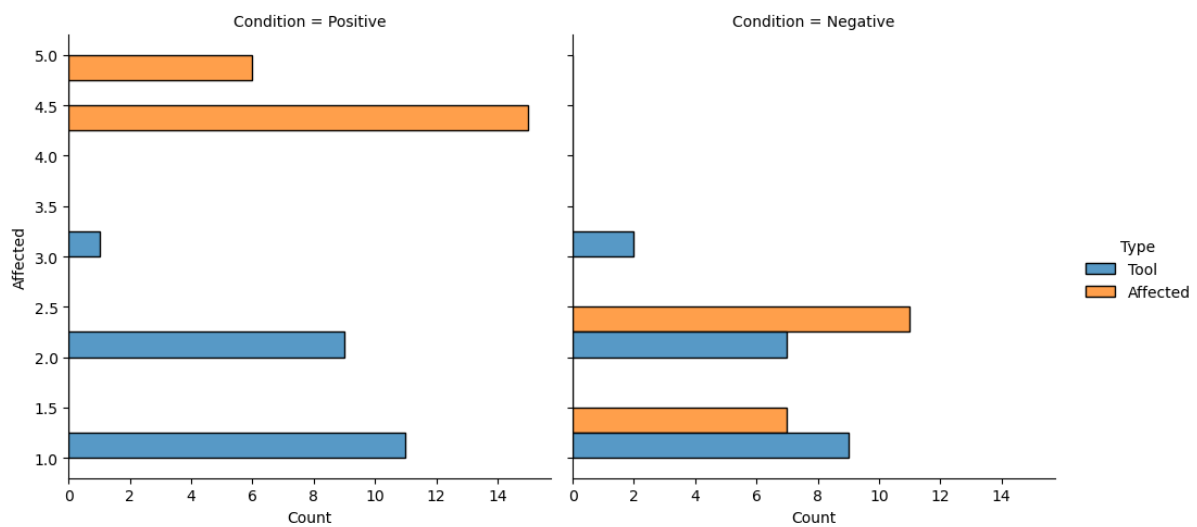


Fig. 6. Participants less affected by toxic language tend to prefer more intervention in a toxicity

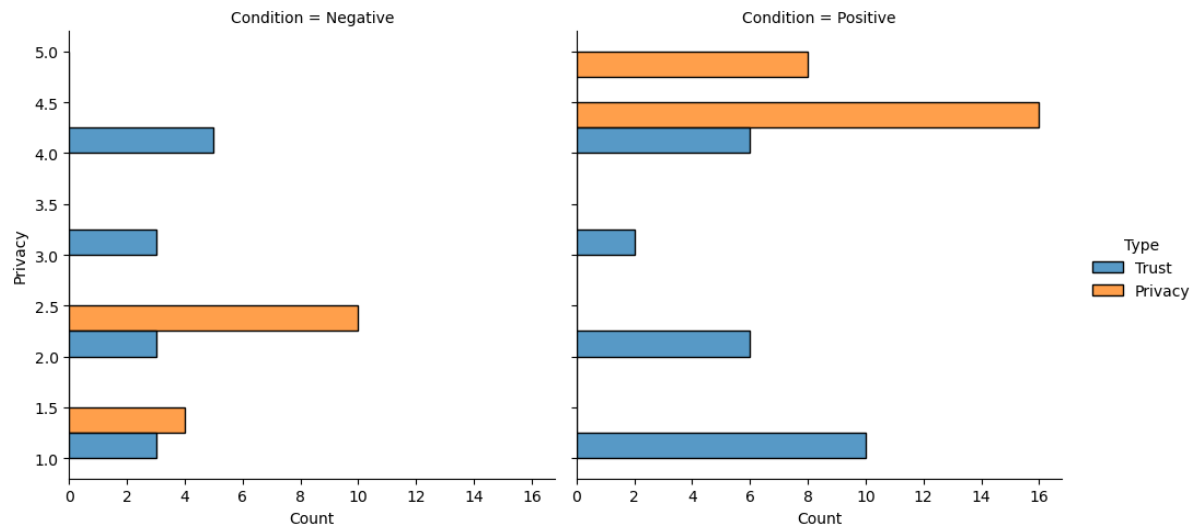


Fig. 7. Participants who have indicated privacy concerns for toxicity detection tend to be less trustful of using such a tool.

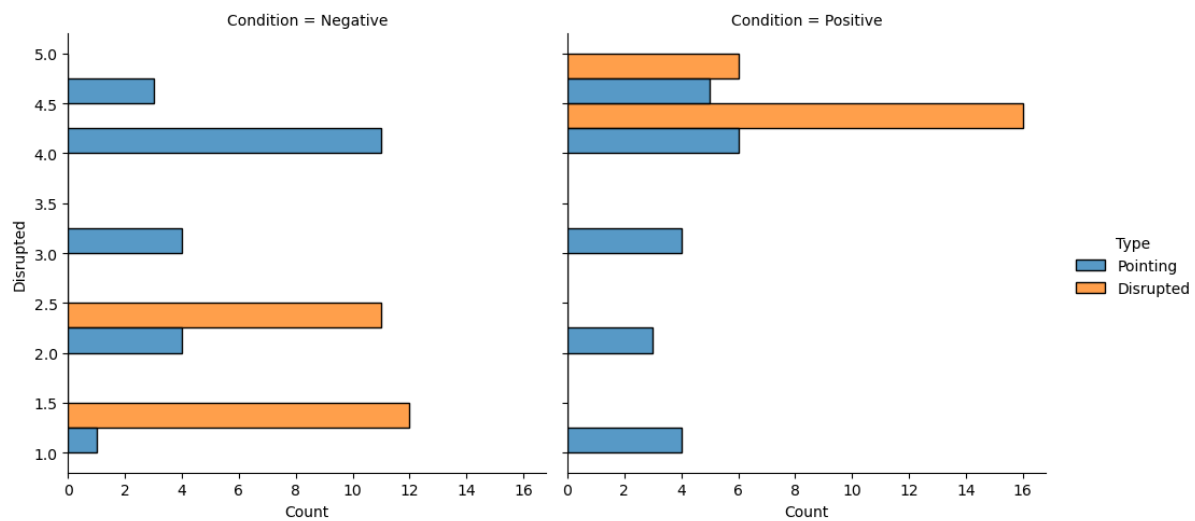


Fig. 8. Participants who have been disrupted by toxic behavior in the past show less optimism that pointing out such behavior can improve subsequent behavior.

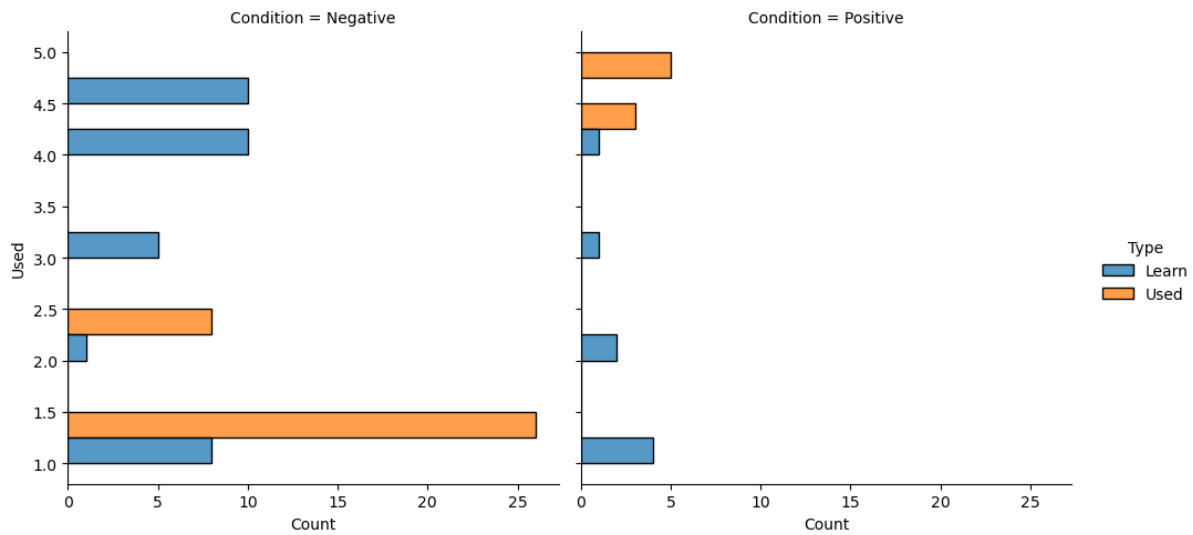


Fig. 9. Participants who have used toxic language in the past tend to be less willing to change their behavior.

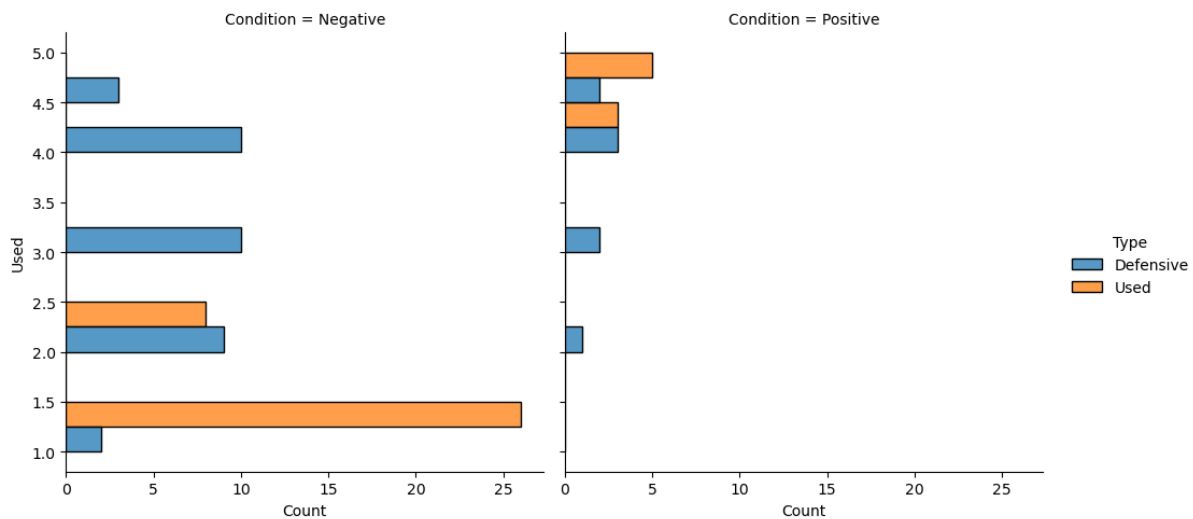


Fig. 10. Participants who have used toxic language in the past tend to be more defensive about their toxic behavior.

4 UNDERSTANDING CONTEXT OF USE: CONTEXTUAL INQUIRY

To understand the context of use for a toxicity-reduction text entry tool, we conducted contextual inquiry on five participants. The purpose of the contextual inquiry is to understand users' intentions, ways of communicating, and breakdowns when performing text-entry tasks in online platforms, especially with regard to dealing with toxic language. One point of interest was in regards to how people receive feedback in regards to their own

written content, as this was an insight found in our questionnaire. Likewise, we were also curious to understand how people interact with systems meant to handle toxic language. We provide details on these two questions in our discussion. All participants reported to have seen some kind of toxic language online and were comfortable with sharing their experiences. The participants comprised of one male and four females, all between the ages of 22 and 25.

4.1 Methods

Before we started any kind of tasks and activities with the participants, we acquired their verbal consent of performing the tasks while the experimenters observe and take notes of their behaviors and responses. Before each contextual interview, we asked the participants to recall the last time they have posted something or saw something that is toxic in Piazza, which is our original focus of platform. If they responded that they did not have the experience in Piazza, we would ask them to recall the last experience regarding toxic language in any platform, and use the platform they mentioned as the context.

When conducting the interview, the participants were asked to navigate to Piazza or the platform of their choice to reflect on the last time they performed the task. To engage the participants in the context, they were also asked to interact with the features of the page that they mentioned such as viewing different posts, replying to posts, and reporting things they found toxic (only as an exercise, didn't find toxic content during interview). While the participants were being interviewed, researchers made interpretations based on participants' description, and discussed with the participants to clarify or correct their interpretations. The researchers finally generated sequence models and flow diagrams for each participant from the interview data.

Based on the similarities and overlaps between the flow diagrams of the five participants, we combined the diagrams of different users and generated the consolidated flow diagram. Similarly, we also combined the information from sequence diagrams of the same tasks to generate consolidated sequence diagrams for different tasks. Finally, we formalized the specifications of users' goals and tasks, and generated seven user requirements in the context of receiving and generating toxic content online based on the breakdowns. These specifications were informed by an affinity diagram we created through our collected interpretations across the five participants.

4.2 Tasks

We asked participants to recall completing tasks in two different scenarios. The first scenario is when the participant reads a toxic post online that is not directed at them. Specifically, the participants were asked to describe the last time they saw posts that "had language directed at them online that made them not want to participate in the discussion". The second scenario is typing toxic languages, which includes two different variants based on if the user was intentionally being toxic or if the user did not realize they were being toxic. Specifically, the participants were asked to describe the last time they were involved in an argument or felt angry online and "had used harsh language online that is likely to make other users not want to participate in the discussion", and "was unintentionally using language that could make other users not want to participate in the discussion". If they did not have the experience and could not answer the question, they were asked to recall the time when they almost sent posts that contained toxic languages or imagine they were in the situation. The researchers also asked the participants to "recall a time when they self-edited something they had written to change the tone or how the recipient could see it" including editing while typing and editing their language after it was posted. The participants were asked to recall their editing behavior in both scenarios of posting toxic languages.

To reduce the priming effect of different scenarios on participants, 2 participants were asked to recall the reading toxic language scenario first and then recall their experience of typing toxic languages, and 3 participants were asked to first recall typing and then reading. In both scenarios, the interviewers asked the participant to walk through the complete process of viewing the post and their behavior and feelings at the time of viewing

and afterward. The interviewers asked follow-up questions based on the participants' description of the task and their responses to previous questions.

4.3 Results

- (1) Participant 1: The participant had not experienced toxic language on Piazza before, so we focused on the participant's toxic experiences on another social media platform, Nextdoor. As the participant browsed through their newsfeed on Nextdoor, they found many instances of controversial content and considered them toxic as it discouraged them from participating in the discussion and in some cases, affected their use of the platform as well (U01-10). Upon discovering such toxic content the participant wondered if they could find someone to talk about such content to articulate their shock. The participant recalled an incident when they had to go and talk to their sister about such content and then they completely forgot about using the platform and went on to do something else. The participant emphasized that in most cases, they do not want to respond to controversial content. In their own words, anything they adds will be like "adding fuel to the fire." (U01-15) Upon further reflection on controversial content, the participant recalled instances when they saw a post that they cared about deeply and therefore, really wanted to contribute. In that instance, the participant spent considerable time to reflect on what they was typing. they would type something, take a pause, re-read it, and then make changes (U01-24,25,26,27). While some changes were benign, such as correcting spelling mistakes, most changes were aimed at limiting any potential fallout from their comment. That included considering the potential audience of the text, limiting any personal references to other commenters, and making sure they were not going to regret it later on. While the participant was able to formulate a response in this case, they reflected upon instances when they simply deleted what they intended to type as they could not make it any less controversial. The participant described regretting later on about not contributing to that discussion (U01-31). In general, the participant felt that their strategies were effective at preventing them from typing anything potentially toxic but that they felt that oftentimes for them, "the pendulum would swing too far in the other direction."
- (2) Participant 2: This participant had used Piazza previously for one class. They had witnessed toxic language on Piazza and did notice the anxious tone conveyed in some piazza posts by their peers. This experience has motivated the participant to be more careful when thinking of making a post on piazza, in an effort to avoid getting a toxic response due to lack of relevant information(See U02 sequence diagram). After this conversation, they went on to describe toxicity on another platform: Facebook Marketplace. The participant had both sold and bought items from Facebook Marketplace and have received toxic responses and given toxic responses themselves. As a seller, they had created a listing with basic requirements, but had many people who misread the post or did not read the post. They said they became frustrated with the flood of responses which did not not put in a baseline effort, and used toxic comments as a way to vent frustration at the moment (see U02 flow diagram 01). They believe a toxicity detection system could be useful when a user is only slightly frustrated but does not believe they would useful in the presence of extreme frustration. As a buyer on Facebook Marketplace, this user also has experienced toxicity when they misread a post of something for sale (see U02 flow diagram 02). Since that encounter however, they have made sure to read posts carefully to ensure smooth transactions on items they really want to buy. They note that toxicity in some doses may be self-correcting behavior for increasing quality of responses on an online platform.
- (3) Participant 3: This participant did not use Piazza, and as such, we discussed various online platforms where they experienced toxicity. These included Reddit, Discord, and multiplayer video games. Overall, they described the various scenarios leading up to the usage or reading of toxic language in each setting. Encountering toxic language online can span a wide variety of scenarios, such as banter or venting in online video games, toxic users excluding people on forums, and even jokes or opinions which are misconstrued

in a negative way (U03 sequence diagrams). Alternatively, the participant's *usage* of toxic language mainly stemmed from either misunderstanding, ignorance, or trying to be helpful in a tense environment (U03-39, U03-51). The participant tried to reduce their toxicity by rereading messages and explaining them before sending them, so as to avoid misunderstandings (U03 sequence diagram 2, U03-26). This has the unfortunate consequence of resulting in messages that are overly verbose (U03-61). They cite the hostile environment created by toxic language as the main downside to its usage in each of the platforms (U03 sequence diagram 4, U03-59). This was observed to reduce interaction from new or outside participants. They detailed how toxic language can ruin a fun gaming environment or lead to a cycle of toxic language being reciprocated between users (U03 flow diagram, U03-13). Additionally, they observed that this toxicity cycle may last much longer in asynchronous discussion forums (U03-62). The participant also described how there is typically one noteworthy message sent which starts these "chains" of toxic language usage, and that the consequences are usually only considered after damage has already been done (U03-49). Finally, they provided their opinion that while a text-entry extension which detects toxic behavior is not useful to those being intentionally toxic, there are many potential benefits in a tool that identifies messages that may be toxic to people of a different background (U03-65, U03-66). The ability of such a tool to explain why a certain message is toxic would also provide some intrinsic value to most users (U03-67, U03-68).

- (4) Participant 4: The participant did not use Piazza as much, and also had not experienced toxic language on Piazza. The platform they chose for the task of browsing toxic content was "Hupu", a Chinese online forum that has different topics but is most famous for gathering sports lovers. This platform is known to have content that are gender stereotypical, especially towards female. Although participant 4 is a female user, they reported that they were less offended by the content in this platform as they already had an expectation of the potentially toxic content (see U04-17), and is thus mentally prepared. They took different methods to reduce the effect toxic content has on them, including getting to know the toxicity level of the platform in general and avoiding reading replies that have too much down-votes. As for producing toxic content, the user chose another Chinese platform which is similar to Twitter. A main concern for the user is to avoid being targeted in heated discussions. Thus, they usually try to make their posts sound reasonable. In the rare case where they wanted to reply harshly because of irritation, they would post their comments under posts that has similar comments so they would not get targeted. In this case, they rarely modify their posts as they do not think it would be seen by a lot of people or be recognized. The participant also recalled a scenario that they unintentionally posted things in a group chat that are harsh, and did not realize it until she got feedback from another person. They did not edit or recall the message as the damage is done, and had to find other ways to make up for it. They would have edited or recalled what they said if they realized it was harsh before the receiver had read it (see U04-35 and U04-36). Thus, the participant needs feedback of if the language is harsh and whether the receiver has already read it.
- (5) Participant 5: This participant had not experienced toxic language on Piazza before. Instead, we focused on a common social media platform Reddit and focused on communities, usually referred to as "subreddits", geared towards learning. We began by walking through a scenario where they had decided to type toxic language towards another user of the platform. The participant noted that they would generally only engage with toxic behavior when toxic language was already present. In other words, they would not organically write a toxic statement. The user then reflected on two instances where they would either write toxic language publicly, or engage with the user in a private message, depending on the severity of the original toxic language and the user's comfort in the community. Given the user was going to respond, the nature of the response would also depend on how the participant was feeling in that moment. This mood could depend on how their day was going, other content they had read, or the level of toxicity in the post (U05-03 and U05-04). The emotional response elicited would also dictate how the participant would handle an apology and subsequently self-edit. If the user apologizes or it was a misunderstanding

and the participant was particularly harsh, they would be more inclined to edit their response (U05-11). On the other hand, if the participant received an apology and was relatively neutral in their response, they would likely keep the response up to provide solidarity to the community. It would be unlikely for the participant to delete a response unless it was severely toxic and completely unrelated to the topic of discussion. The user would continue to use the platform, unless this was a common occurrence, to which they would consider leaving the community (U05-16). For our second scenario, we asked the participant to consider an instance where toxic language had been directed towards them, for instance as a response on a post they made. While there were similar findings in terms of how they would handle the situation from a written response, the participant would also utilize the reporting system embedded in the platform as well as confide in a trusted individual (U05-25, U05-26, and U05-27). In the case of reporting, if the toxic language was unwarranted or extremely harsh, the participant would engage with the reporting capabilities of the platform, particularly if the platform was user-moderated. The participant noted it was important to them that another person would see the report as they were unsure automated systems could accurately understand the toxic language. The occurrence of this case is documented in the U05 sequence 2 diagram, and indicates the importance of trust and accuracy in a system. As a follow up, the user would also share the instance with a close friend or family member to justify their feelings on the situation. Again, the user would likely continue to use the platform, but would exhibit more caution in what they post and how often they post.

Once all five participants' contextual inquiries were analyzed in isolation to one another, we consolidated each participants results into single sequence and flow diagrams. In addition, we also created an affinity diagram, using all interpretations generated during the contextual inquiries, to aid in developing larger themes from the discussions.

When consolidating the sequence diagram, we focused on two overarching scenarios, a) reading undirected toxic language on a site, and b) reading toxic language when it is specifically directed at a user. One difference that is immediately obvious between the two events is that the user is much more likely to respond with toxic content when the language is directed at them, as compared to when they read it online. This is evident through the other mechanisms one engages with toxic content in the general reading case such as reporting or sharing with others (see consolidated sequence diagram 1). This insight indicates a requirement that while a user may want to respond to a message or interact with toxic content, they should also be given other methods of intervention. In addition, both scenarios also call for some form of editing given sufficient conditions are met, such as when someone reflects on their actions or doesn't receive an apology. While it isn't exactly clear yet how a user level system would allow for this modification explicitly, it may be possible to mimic some form of this behavior on the user end through the tool.

When consolidating the flow diagram, we noticed first and foremost that there are many individuals involved when a user is interacting with toxic language online. For example, a user may not be simply interacting with just the other user who either sent or received the toxic language, instead many third parties such as friends/family or site moderators may also be involved (U01-08, U04-11, U05-25, U05-26, and U05-27). This finding will require us to consider how a solution that is meant to target toxic language interacts with those stakeholders, not just the ones sending and receiving toxic language. In addition, we noticed that the user in question has many different ways they can interact with platform (see consolidated flow diagram, particularly the number of arrows coming in and out of the user object) which can be difficult to completely cover in one single solution. With this mind, the user will need a solution that is flexible and provides many means of interaction between the various stakeholders as well as with the toxic content itself.

The affinity diagram provided a more granular overview, compared to the sequence and flow diagrams, of the various interpretations across users and how they align with one another. We originally came up with four large

themes that correspond to a) how users interact with toxic language, b) how users avoid toxic language, c) users seeking third parties to validate or moderate, and d) how users react and reflect on their behavior. From there, we broke down each theme into a set of sub-themes that can be found in the affinity diagram. One theme, that received a lot of attention was when users try to avoid toxic content, where users can do anything from simply scrolling past the content (U01-11), place themselves in the shoes of the other individual (U01-27), or even leave groups all together (U05-35). Likewise, our single most common sub-theme was user's reflecting on their actions, where reactions anywhere from apathy (U02-28) to guilt (U03-47) were present. This finding follows a similar argument as found in both our individual contextual inquiries and questionnaire where interaction with toxic language is a very personal experience and the response somewhat might have can be very different from another person. As such, and similar to our findings with the consolidated flow diagram, flexibility and choice will be a strong requirement. We do note that not all of the interpretations were applicable to the general themes found in our contextual inquiries. Many times, these specific interpretations were observations of a physical action someone was taking, and less about the insight gained from the action, making them not as useful for the affinity diagram. There were also instances where different participants had nearly identical interpretations simply due to how similar they performed certain actions. These duplicates and and less applicable interpretations are placed in their own category outside the affinity diagram.

5 USER REQUIREMENTS AND FUNCTIONAL CONSTRAINTS

Below we list our series of user requirements and an associated explanation for why we included them, given our current understanding of context of use. One general theme we have seen across the board comes from the fact that individuals experience toxicity very differently, and freedom to make various choices must be allowed. In addition, given the emotions involved with toxic content, simplicity and speed is key, i.e., users need to be able to quickly and efficiently change their content if the goal is to limit toxicity. If the users have to go through a significant amount of effort before changing the content, they may be less likely to do so. With this in mind, we consider various user requirements to outline what these different choices should be and detail how they support a use case seen in our contextual interview.

- (1) When the user intends to send a message containing potentially toxic content, the user must be able to edit their message before sending it.
- (2) When the user intends to send a message containing potentially toxic content, the user must be able to delete their message before sending it
- (3) When the user intends to send a message containing potentially toxic content, the user must be able to know whether the receiver would be negatively affected by the message.
- (4) When the user intends to send a response to a message containing toxic content, the user must be able to convey their discontent in their response message.
- (5) When the user sees a message containing toxic content, the user must be able to take actions to mitigate the negative effect of toxic content on them.
- (6) When the user intends to contribute to an online discussion, the user must be able to know whether the discussion is toxic.
- (7) When the user is interacting with toxic content, any moderation or toxic feedback system must be able to accurately discern whether content is toxic with human-like performance.

The explanations and context for each user requirement is provided below.

- (1) From sequence diagrams from multiple users, a common theme occurs where users edited a potentially negative message of theirs to be less abrasive (U01 sequence diagram 2). We found that our users tend to avoid instigating or fueling conflict whenever possible, and as such, the approach of reflecting on their message before sending it was shared by multiple users. Occasionally, this could involve a user altering a

message to contain an explanation or use more passive language (U03 sequence diagram 2). They might modify the tone or content of the post to be less harsh (U04 sequence diagram 2). Additionally, the user may take time to fix grammatical errors and typos within their messages (U01 sequence diagram 2).

- (2) In several situations, the user may struggle with whether a certain message is okay to send, thus they will consider the option of deleting, canceling, or recalling the message in question (U01-33, U05-29, U04-29, U04-32). To avoid harming other users (which could be either friends or strangers), they may reconsider sending a particular message at all.
- (3) For cases where the user is not intentionally trying to hurt someone else with their message, we found through various interpretations (U04-34, U03-50) that users wish to understand whether their message was too harsh. In several sequence diagrams, our users noted they were unsure whether their message would negatively affect the receivers (U04 sequence diagram 2, 3 and U03 sequence diagram 2, 3). For example, after sending their message, one user would need feedback to know whether their message could be construed as offensive by certain people. Another user felt it would have been beneficial if they could have been alerted to the toxicity of their message proactively.
- (4) Through our contextual interviews, we've found that users sometimes rely on feedback to determine whether a message they sent was toxic (U04-34). Without feedback from a reader, users who have used toxic language in a message may not understand why their behavior was hurtful, and they may continue their harmful behavior. When users notice toxic content online, they may wish to intervene or respond by expressing their discontent or refuting the content within the message. Thus, for the recipient to vent and the sender to get feedback, there needs to be communication channels initially.
- (5) One frequently appearing method for dealing with reading toxic content included either venting about it or censoring it in some way. This could include muting the user who sent the message, attempting to forget about the message, reporting the message, downvoting the message, or ignoring the message (U03-06, U04-04, U05-21, U04-20). The user might even take a break from using the platform entirely (U05-15, U03-15).
- (6) To avoid escalating conflicts in online discussions, a user should be able to discern whether a particular discussion or conversation is beginning to get heated (U04 sequence diagram 2, U03 sequence diagram 4). This is important for understanding whether to contribute to or join the discussion at all. If the user still intends to join the discussion, they may opt to minimize any negativity in their messages as discussed in user requirement 1.
- (7) One user specified that it would be important for any moderation system to be accurate in its message removals (U05-26). For example, users have faith in human moderators' ability to curate content by manually removing toxic or controversial posts (U04-11). If we were to implement a toxic feedback system in our design, it should likewise be able to categorize messages as toxic with human-like accuracy to meet these user expectations.

Upon iterating on our user requirements, we developed a new set of user requirements which are more constrained and more accurately reflect the set of *useful* potential designs.

- (1) The time required for the user to edit (before sending) their toxic message into a less toxic message (that maintains a similar meaning) should be less than manually deleting and entering each character.
- (2) The time required for the user to delete the toxic content in their message should be less than manually deleting each character.
- (3) The user must be notified before they send their message when content in their entered message is toxic. This notification must guide the user to remedy the toxic content.
- (4) When the user intends to send a message containing toxic content, the curious user must be able to understand why the intended recipient might be negatively affected by the message.

- (5) The user must be able to customize the types of toxic content to be notified about in their entered message.
- (6) The user must be able to customize the types of toxic content to mitigate the negative effects of.
- (7) When the user sees a message containing toxic content, the user must be able to preemptively mitigate the negative effect of the toxic content on them using the design.
- (8) When the user intends to contribute to an online discussion, the user must be able to know how toxic the discussion is before contributing.
- (9) When the user is interacting with toxic content, any system dealing with toxic content in the design must be able to accurately discern the extent to which content is toxic with human-like performance.
- (10) The learnability and discoverability of the design must be intuitive to new users.
- (11) The design must be able to meet each user requirement while being platform agnostic.

Our process for revising the user requirements involved asking critical questions regarding the potential design implementation details for each requirements. For example, in the original Req. 2, which parts of the message should be deleted? It only makes sense to delete the toxic portions of the message. Likewise in Req. 1, how easy should it be for the user to edit their message? What is the user's goal with editing the message? It makes more sense to further constrain the requirement so that the potential design requires less workload than manually editing the message. Additionally, the new message should be less toxic than the old message. We also tried coming up with "bad" designs which minimally address the user requirements. For example, in Req. 5, a tool which censors the toxic text after the user has already seen it does not seem as useful as one which preemptively censors the text. Likewise in Req. 1, a tool which allows the user to edit their toxic message but requires lots of overhead from the user does not seem as useful as simply manually deleting and adding characters. Overall, the new set of user requirements more accurately represents the basic requirements for addressing our users' needs in the context of use.

6 INITIAL DESIGN AND LOW FIDELITY PROTOTYPES

6.1 Individual Personas

This outlines the different personas and their expectations for how they would interact with the tool. Further personal details can be found in the appendix.

- (1) This persona is an individual who browses frequently on websites and is less worried about toxic content. As such, they are okay when they see it, but think it would be nice to be notified.
- (2) This persona is a person who commonly and openly writes toxic content. However, they are trying to improve their toxic online behavior.
- (3) This persona is a person who mostly browses online contents instead of posting contents. Because they are aware that online platforms can possibly contain a lot of toxic contents, or they do not often feel targeted by toxic content, they are mentally prepared for toxic content online.
- (4) This persona is a moderator of the site with the goal being to determine whether automatically flagged content is actually toxic. They will interact with the prototype by receiving notifications of newly flagged content and manually determining whether the content should be flagged. The goal is that the tool can help expedite their work.
- (5) This persona is an individual who frequently browses online content but does not want to interact with toxic content each time they come across it. As such, they would prefer the content be hidden from them.

6.2 Individual Sketches

For each sketch, please refer to the appendix.

- (1) The first sketch (Figure 23) shows a mobile plugin that displays the toxicity rating of the text as the user is typing that text. The sketch also highlights the part of the text that is particularly toxic while using colors

within a digital meter to indicate the perceived level of toxicity. Users who want to avoid toxicity may receive useful feedback on how toxic their content is and be able to make changes to particular parts of the text in order to reduce the toxicity.

- (2) Sketch 2 (Fig. 24) is done in the context of Persona 2 writing toxic content on Twitter, although based on the design (a browser extension), this could be done in any online forum setting. In this design, potentially toxic content that is detected within a text entry box would be highlighted. The user would be able to click on a small icon which would pop up an explanation behind why the content is toxic. They would then have the choice to either delete the highlighted content or apply our extension's auto-suggested fix. Note that the post or tweet button would be grayed out or crossed out. While the user can still post the content if they desire, this design nudges the user away from doing so.
- (3) Sketch 3 (see Fig. 25) is done in the context of Persona 3 browsing toxic content in an online forum, and have the system installed in their browser. The system automatically blurs out toxic content. When user hover on the blurred out area, the system will prompt the user that this area contains potentially toxic content. However, it also gives users the freedom to read the original content despite the warning. To avoid accidentally revealing toxic content that could be harmful to individual's mental health, the user needs to double-click to reveal the content. The revealed toxic content will have gray background to remind users that this is potentially toxic, and there is a button for users to block out the content again.
- (4) Sketch 4 (Fig. 26) is done in the context of the moderator and focuses on the interface they might use to override automatic flags. In the panel, the moderator is provided with a series of posts that the automatic flagging tool caught. With each post, there is the perceived associated toxicity which comes from the associated model. To perform an override, or keep the content as flagged, the moderator is provided a green and red button, which performs the respective action. Lastly, the moderator is given a blue button that allows them to visit the full post in case they need more context to make a decision.
- (5) Sketch 5 (Fig. 27) describes a design which protects the user from engaging with toxic content if they so desire. The user can set the level of system sensitivity, depending on personal preference. The system will hide toxic content when detected, and will not be revealed unless the user clicks to reveal the toxic content. The user can also choose to not engage and continue scrolling, and have the ability to never witness the toxic content.

6.3 Individual Storyboards

Please refer to the storyboards in the appendix.

- (1) Storyboard 1 (Figure 28) shows a user browsing through their neighborhood app who suddenly comes across potentially toxic content. Since the topic is of personal relevance to the user, she feels like participating in the discussion. She types out her message but decides to not participate as she fears that her contribution will make the discussion even more toxic.
- (2) In this storyboard, Persona 2 is thinking of writing something slightly hateful and posting it to twitter. The solution (in the form of a browser extension) highlights toxic content as they write it. The user, curious, clicks on one of the highlighted phrases to understand why this content is potentially toxic. The user then decide to post it regardless. Clicking on the grayed (in this case, red) out post button reveals the actual post button. The user has posted their content, but since they still have the extension installed, the user notices that their post has been automatically censored by the extension. Since the user still wishes to see the content, they hover over it, revealing the potentially toxic content.
- (3) In storyboard 3, the user is browsing casually in an online forum. While they are browsing, they encountered something that are potentially toxic and is blurred out by the system. When they hover on the area, they see the hint provided by the system. After they become mentally prepared and decided that they are curious

about the original content, they double-clicked to reveal the original toxic content. After reading it, they decide to leave it blocked out so they won't see it repeatedly.

- (4) In storyboard 4, the moderator of the site is interacting with content that has been automatically flagged as toxic. They begin by first clicking on a link that notifies them of the newly flagged content. In the link they find a series of posts that have the associated toxicity level, buttons to override the automatic flag, as well as a way to visit the post. In this instance, the first post was incorrectly deemed to be toxic and thus the moderator overrides the flag. The moderator is then able to continue interacting with the content.
- (5) Storyboard 5 (Fig. 32) describes a scenario of avoiding toxic content. A user can have complete control over their browsing experience. If they are using a toxicity detection system, all toxic content will be covered. Upon hovering over the content, the user will be informed that this content is toxic. The user can choose to click an additional time to reveal the toxic content itself. If in this instance the user does not wish to engage in such content, they will simply navigate away and continue browsing.

6.4 Design Critiques

6.4.1 Critiques of Sketch 1.

- (1) Toxicity meter might obstruct important information. Maybe highlights could be color-coded?
- (2) Ambiguous in terms of if it is the whole message or if only part is deemed toxic.
- (3) Maybe the cancel button should be on the right side, and send could be on the left; would be harder for users to press the send button.

6.4.2 Critiques of Sketch 2.

- (1) "Fix" phrase is ambiguous, would it fix it for you, do you need to fix? Need to clarify.
- (2) Determining why it is harmful may be hard, I would just simply point to which words were useful in deeming it toxic.
- (3) In the popup dialogue box, I am not sure if the reason and the population are picked from what the user types or are they identified automatically like through some language model.
- (4) Will clicking the "delete" button only delete the highlighted part? Will probably make the sentence incomplete.

6.4.3 Critiques of Sketch 3.

- (1) It is unclear how a user can see the content, what is the action.
- (2) Double-clicking on the blurred area is extra effort than a single click. Also the content warning message might not appear on smaller text passages.
- (3) Minor comment but it seems like user actions are also becoming part of the sketch, and I wonder if that's how it's supposed to be or would it confuse people.
- (4) It will take extra effort to reveal toxic content to make sure that is really user's intention.

6.4.4 Critiques of Sketch 4.

- (1) The "are you sure?" message should be optional and be able to be disabled. The buttons seem too small. This toxicity moderation method requires much manual effort.
- (2) Not sure how to link as a platform agnostic tool
- (3) I was not sure what the key meant, specifically that it can be pressed, but I am not sure if there is a way to indicate an unpressed (or unselected) vs pressed (selected) key.
- (4) Increase button size for "is vs is not" toxic to reduce error.

6.4.5 Critiques of Sketch 5.

- (1) Toxic sensitivity meter could be subjective and may not be well defined. Hovering might unintentionally reveal toxic content.
- (2) Should notify the user beforehand of the level of toxicity.

6.5 Final Persona

We consider two personas for our prototype, one being a user who frequents the site, both reading and writing content, as well as a moderator who interacts with the content. For the user, they are expected to be someone who often will engage in discussion with others and occasionally write something toxic. They are likely looking to decrease their usage of toxic content. Likewise, they also often read content which may be toxic and would likely be hurt when exposed to something toxic. They will be expected to interact with the prototype by a) typing messages of varying toxicity, b) self-editing through a less toxic statement recommended by the tool, c) reading messages of varying toxicity, and d) exposing themselves to toxic content. As for the moderator, they are expected to be someone who has a relatively unbiased opinion on the matter and can objectively determine whether a statement was meant to cause harm to others. They will interact with the prototype by manually looking at flagged posts and, with the help of the automated tool, determine whether something is actually is toxic or not. Their choices can be fed back into the model to help train it to better identify toxicity. Some further details can be found in the appendix.

6.6 Final Sketch

We combined features from individual sketches and addressed the critiques in our final sketch.

For common users who read and write content on the platform, we introduced a settings page for users to customize their preference for reading and writing toxic content. Users can customize the severity level of language that the tool should blur when reading, or alert to change when writing. For reading scenarios, an alert for level of toxicity is shown when user hovers on toxic content to address the critique of notifying users beforehand, and to avoid revealing toxic content intentionally. To avoid accidentally revealing toxic content, users need to double-click to reveal blurred content. To address the critique that content warning might not appear on smaller messages, in the case of smaller text passages (not demonstrated in the sketch), a bubble will appear on top of the message with level of toxicity and content warning.

In writing scenarios, toxic content typed into textual input elements (on web pages for example), will be highlighted and underlined according to the degree of toxicity. This initial intervention is relatively unintrusive and simply notifies the user when the system detects something toxic in the user's typed content. If a user then chooses to click on the highlighted phrase, a semi-transparent dialog pops up suggesting how the user can reword the phrase to be less toxic. The user can choose to either apply the suggested changes, delete the highlighted phrase, or close the dialog and submit the text normally.

For moderators who monitor flagged posts and have the right to override system's toxic language detection results, the plugin has an interface for them to review posts that are flagged as toxic. To address the critique of reducing error in clicking "toxic" and "not toxic", we increased button sizes and used vibrant green and red colors to indicate toxicity. We also made the confirmation message optional based on the critiques.

6.7 Final Storyboard

The final storyboard is comprised of a few scenarios, particularly instances of reading, writing, and moderating toxic language on the site. In all instances, we begin with the user logging on the site and arriving at the home page. Here, the storyboards split to either finding a toxic post, writing a toxic message, or receiving notifications of newly flagged content for the reading, writing, and moderating instances respectively.

For reading cases, the user is shown browsing Twitter and coming across a piece of content that is blurred out. Out of curiosity, they hover over the content and are notified of the toxicity level. They decide they still want to read the content, so they double click on it, exposing extremely toxic content that makes them upset. They decide to re-hide the content to make sure they cannot see it. This instance exemplifies the critique of storyboard and sketch 5 well, where the user otherwise would have been exposed to the content without as much consent (by only hovering over).

In the case of writing, the user is shown typing out a post that ends of being toxic. The automated tool then notifies the user that the statement is toxic and recommends an alternative phrasing. The user then decides to apply the phrasing to the content, thus leaving them with a less toxic statement. In this instance, an alternative scenario where they delete the content is also included. Based on critiques of sketch 2, we try to offer an alternative phrasing instead of offer an explanation of why the content is toxic. We believe this can be done with modern language models.

For moderation, the storyboard depicts a moderator looking at their notifications and assessing the flagged content. They notice one content was flagged with low toxicity, and when reading it they determined the content wasn't actually toxic. To fix this, they override the flag and allow the content to be viewed by all. The originally flagged content is removed from the list and they begin to process the others.

6.8 Paper Prototype

The final paper prototype demonstrates how one would interact with toxic content from two different personas, the typical user and the moderator. The typical user reads and writes content and can interact with toxic content through both means. The reading interface is derived from sketch 1 and sketch 5, while the writing interface is derived from sketch 2 and 3. When reading toxic content, the tool blurs out phrases that it deems as toxic. Based on critiques of 1, our method places a small bar measuring the perceived toxicity on the blurred phrase to mitigate taking up too much room and obstructing the users view. Also, based on sketch 5, we allow the user to choose a toxicity threshold beforehand to determine how toxic a piece of content needs to be before blurring. As for writing, we followed a similar design to sketch 2, where a pop up occurs when toxic content is written and notifies the user of toxicity while also giving a suggested rephrasing. Based on critiques of sketch 2, we first changed some of the button phrasing to ensure clarity of the uses. Specifically, the user is provided a toxicity level as well as a possible alternative phrase that they can then apply. Likewise, based on critiques of sketch 5, we try to identify specific phrases in the entire written sentence that is toxic, rather than simply highlighting the entire written content, to promote detoxifying language. As for the other persona, the moderator, we provide an interface that they can manually override content they deem not toxic. In this page, they will be provided all of the posts deemed toxic with the associated toxicity rating and the opportunity to either agree or disagree with the flag. Based on critiques of the moderator sketch (sketch 4), the keys were made larger and made to be more clear in terms of the actions they perform. Likewise, more options such as a way to disable warnings was included.

In order to wizard-of-oz the slide deck created, we plan to use the tool InVision¹. The paper prototype will be converted into slides and the user will be asked to go through these slides. As the user interacts with the slides, such as clicking on various buttons in the interface, a member of the project team will manually change the slides to reflect what the associated action would be. This will provide the user with an experience similar to what the actual product would look like without requiring as much work. The slides that we will use, in roughly sequential order of how an action would take place, can be at the link in the appendix.

¹<https://www.invisionapp.com/>

Issue	Usability Heuristic	Severity
How to access the configuration screen was not immediately clear	1 (Visibility of system status)	0
Need some indicator that the plugin is actively interacting with the users text (how Grammarly shows their brand icon)	3 (User control and freedom)	2
Need consistency in terms of the result of a hovering action vs a click-based action]	4 (Consistency and standards)	2
Browser-based UNDO action should be supported by our plugin	5 (Error prevention)	0
What is considered toxic is not at all clear. We may need personalization	7 (Flexibility)	3
Information on what and why something is considered toxic	10 (Documentation)	1

Table 1. Results from heuristic evaluation

7 USABILITY EVALUATION

7.1 Heuristic Evaluation

7.1.1 Method. We conducted a heuristic evaluation exercise to uncover potential usability issues with our initial design. For this exercise, we used the 10 usability heuristics provided by the NN group and covered as part of the lecture material. The goal of the demo was to reduce potential exposure to toxicity online.

7.1.2 Participants. We recruited our usability expert from Group 3 of the class. One member of our team presented a demo of the design to the usability expert who then took notes on the various usability issues with our design and their severity.

7.1.3 Tasks and Procedures. As noted above, the goal of the demo was to reduce potential exposure to toxicity online. The goal further had two sub-goals: hiding toxic content that is already present online, and preventing further toxicity when contributing new content. These goals were appropriately explained to the usability expert and then the subsequent tasks in line with the goals were performed. As part of the first task, the team member configures the design to set their threshold for toxicity and toggle the various features on/off. For the next task, the team member comes across potentially toxic content and demonstrates how a user may choose to interact with such content. In the next task, the team member proceeds to contribute their own content that is potentially toxic, and shows the different options that would be available to the user in such a context. The exercise ended with the usability expert providing a quick debrief of the issues and sharing their notes on their severity.

7.1.4 Results. The summary of issues identified via heuristic evaluation can be found in Table ?? . The usability expert did not find any issues with a high-level of severity. Some of the issues (heuristic 1 and heuristic 5) were further attributed to a lack of explanation provided during the demo and their severity was reduced to 0 subsequently. The most severe issue was related to a lack of explanation and flexibility offered in terms of what is considered toxic. Even though the system allowed users to set their own toxicity threshold, the usability expert felt that people may have different notions about what is considered toxic, and therefore, simply having a threshold may not provide them with appropriate flexibility.

7.2 Simplified User Testing

7.2.1 Tasks and Procedures. To evaluate our paper prototype with our participants, we screenshared the slides. We additionally asked our users to screenshare once more in order to track their cursor. While the user interacted with various widgets and elements on a particular slide, we would manually move objects or switch to the next slide. The users were instructed to interact with the UI to accomplish the following goals:

- (1) Set the toxicity settings to your desired settings
- (2) Reveal the censored toxic text
- (3) View the suggestion for decreasing the toxicity of your writing
- (4) Apply the suggestion for decreasing the toxicity of your writing
- (5) Remove the toxic content from your message

7.2.2 Participants. Each of our participants were informed of the procedure/process of simplified user testing and gave their consent to participating in our user evaluations. Participants were recruited such that they matched our primary stakeholder group: someone who uses online discussion platforms and has either been exposed to or used toxic content. The participants were briefly primed on the context behind our browser extension tool. They were then asked to accomplish the earlier sub-goals within three contexts:

- (1) They were asked to imagine they just downloaded the browser extension and were setting up their preferred settings.
- (2) They were asked to imagine they wanted to read some potentially toxic content that was censored by the tool.
- (3) They were asked to imagine they had unintentionally typed some toxic content and wanted to fix their typed content to be less toxic.

7.2.3 Results. Here we describe the feedback we received during the simplified user testing across our three participants.

Participant 1

Participant 1, Goal 1: The user had trouble interpreting the sliders on the settings page. They were unable to find any documentation or information for assistance with their issues. At first, they placed the slider position at the opposite to what they intended to accomplish (they wanted to be notified only with the highest level of toxic content, but placed the slider at the "Mildly toxic" section). This is likely due to confusing wording, and the rectangles stretching from right to left instead of the standard design of left to right. Additionally, there should be some sort of feedback indication on what the selected threshold does, e.g., how much content will be censored for this slider position?

Participant 1, Goal 2: The user hovered over the censored text and initially left clicked a single time. Afterwards, they realized they were supposed to double click to reveal the text. The user suggested that an "unhide" button might be more intuitive. They also indicated that they would get frustrated and likely uninstall the extension if any of the interactions with the censoring tools did not line up with their intentions. After revealing the text, they pointed out that there was no point in having a hide button, since the content was already revealed.

Participant 1, Goal 3 4 5: The user hovered over the highlighted text expecting something to happen, but nothing happened. This indicated an inconsistency in the functionality between the censoring and notification components. After clicking on the highlighted text, they viewed the pop-up containing the notification and suggested fix. They indicated that the design seemed to only allow the user to either apply the suggested fix or delete the content (i.e., there was no obvious option to hide the pop-up and do nothing). They indicated as well that they found using the mouse to apply the suggested fix extra effort. They suggested having the "enter", "delete", and "esc" keys correspond to the actions on the pop-up.

Participant 2

Participant 2, Goal 1: When the user was first placed the settings menu, they were stuck in the gulf of execution and confused on what they could actually interact with. They noted having some method to help identify objects that could be interacted with would be helpful. Furthermore, once the user figured out they could interact with the bars, they were unsure what changing the sensitivity of toxicity actually did, thus stuck in the gulf of execution. Given the current inability to provide immediate feedback until the settings are finalized, the user has no real ability to reconcile what the action did. A better form of feedback should be added to help the user understand the tools behavior.

Participant 2, Goal 2: When the user was placed onto the twitter homepage and asked to interact with the blurred content, the user instantly clicked on the toxic content. While they were happy with the safeguard which requires a double click to expose the content, they felt that the requiring to first hover, receive notification, and then double click was not intuitive. They thought having the toxicity immediately displayed could be helpful to

remove the need to first hover. In addition, the user noted that when the content has been exposed, the option to hide was ambiguous. They were unsure if the hide button would hide the entire message, hide the toxic content, or hide the toxicity monitoring system itself. Being explicit about these buttons is a requirement to aid in use.

Participant 2, Goal 3, 4, 5: The user was tasked with writing content and to interact with the tool as it determined their content was toxic. While the user recognized a highlight occurred in the system, they were unsure how to exactly interact with the highlight, leaving them in the gulf of execution. They also were unsure what the color was meant to represent as they recognized the orange from the settings bar. Once able to open the pop up window which provides the various interaction methods, the user was unsure exactly what each button would do. In particular, the delete button provided the most ambiguity as it could have either deleted the entire message, or just the toxic content. Likewise, without a way to easily change the settings of the toxicity detection tool, they felt that they would become frustrated with the highlights. This could mean we need a less intrusive method to indicate toxicity without highlighting the entire phrase.

Participant 3

Participant 3, Goal 1: When introduced to the system settings, the user had trouble understanding what parts were interactive. This ambiguity left them in the gulf of execution where they simply couldn't figure out what to do next. The user was also unsure about what the different colors represented in the bar, indicating an ambiguity that needs to be addressed. Moreover, the user was unsure how actually interacting with the bar would change the threshold and respective content.

Participant 3, Goal 2: When interacting with the content on the page, the user had trouble processing why there was a blurred text. Without notification from the system settings that blurred text would be indicative of toxic content, they were unsure what was going on. In addition, the double-click action was also not immediately clear to the user, leading them to make a mistake in the interaction. The user would have preferred something similar to other products such as a single click to make them not need to learn new interaction methods.

Participant 3, Goal 3, 4, 5: When interacting with the texts that they wrote, the user felt that they wanted more immediate feedback on their toxic content. Specifically, the user felt that double clicking to perform an action was a higher barrier and they would have preferred something simpler. In practice, they would be more likely to disable the tool given the unnecessary tasks required to complete their goal. Likewise, the functions of the apply and the delete button were not immediately clear to the user. They were unsure what each button would do, leaving them in the gulf of evaluation upon in each interaction.

Overall our design failed to meet the following requirements in these particular scenarios:

- *The user must be notified before they send their message when content in their entered message is toxic. This notification must guide the user to remedy the toxic content.* During the simplified user testing, we found that the notification in intense highlight can be confusing and intrusive to users. This undermines the notification's ability to guide the user to remedy the content, but can potentially cause frustration.
- *The learnability and discoverability of the design must be intuitive to new users.* For the settings menu, users were confused about what they could do with the settings menu, which indicates that the learnability of the system has not met the user requirements. We also discovered several designs that were not intuitive to users. For example, using double-click to reveal blocked toxic content is perceived not intuitive, as well as the sensitivity bar in the settings menu.
- *The workload required for the user to edit (before sending) their toxic message into a less toxic message (that maintains a similar meaning) should be less than manually deleting and entering each character and The workload required for the user to delete the toxic content in their message should be less than manually deleting each character.* While the user requirement seemed to be met, they can be improved by adding keyboard shortcuts to enable faster interaction.

8 FINAL DESIGN AND FUNCTIONAL HIGH-FIDELITY PROTOTYPE

Based on the feedback from the qualitative evaluation, we changed our design in the following ways:

- To reduce the negative effect of intense highlight in writing, we improved the highlight by making the color less intrusive to the users.
- In terms of reading toxic content, we added a "reveal" button for users to click to reveal the content such that the interaction would be more intuitive and obvious.
- To make the sensitivity bar more intuitive, we removed the slide bar, and adopted a design using blue to indicate the level selected and gray for available levels, which is more similar to other existing material UI.

After iterating on the design, we implemented our final prototype using Figma, with Twitter as our sample platform. Fig. 11 shows the prototype when user encounters toxic content that are blurred by the system. Users can hover on the blurred area to inspect the perceived toxicity of the content (see Fig. 12), and click on "reveal" to see the original content (see Fig. 13).

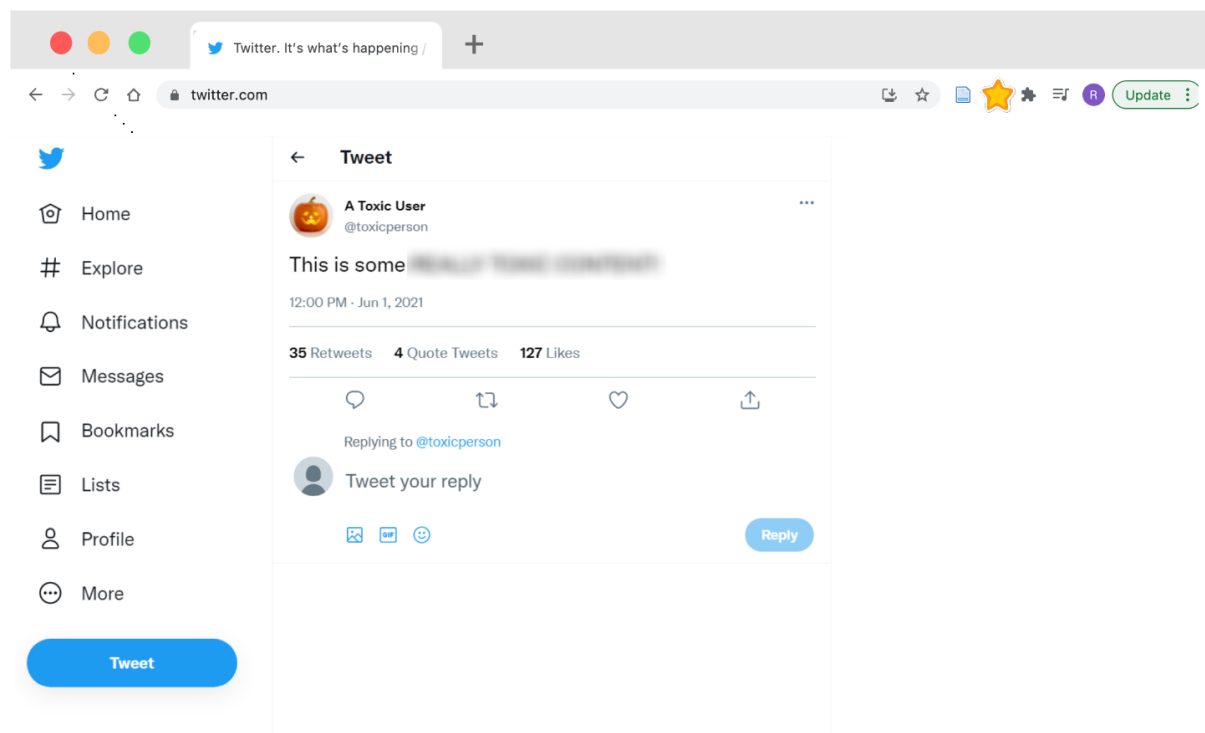


Fig. 11. Screenshot of the tool prototype when user encounters toxic content when they are reading.

In terms of getting feedback on writing toxic content, fig. 14 demonstrates how detected toxic content in users' writing will be mildly highlighted. When users hover on the highlight to look for details of the reminder, the tool will present a suggested fix for the toxic content (see fig.15). If users click "apply", the system will automatically fix the content (see fig.16. If the users click "cancel", the highlight will be removed.

As the tool is envisioned as a Chrome browser plugin, the settings will be automatically embedded in the plugin settings (see the yellow star on the plugin area). When users click on the icon of the tool, the settings will pop up, and users can adjust their preferences for both reading (see fig. 17) and writing (see fig. 18).

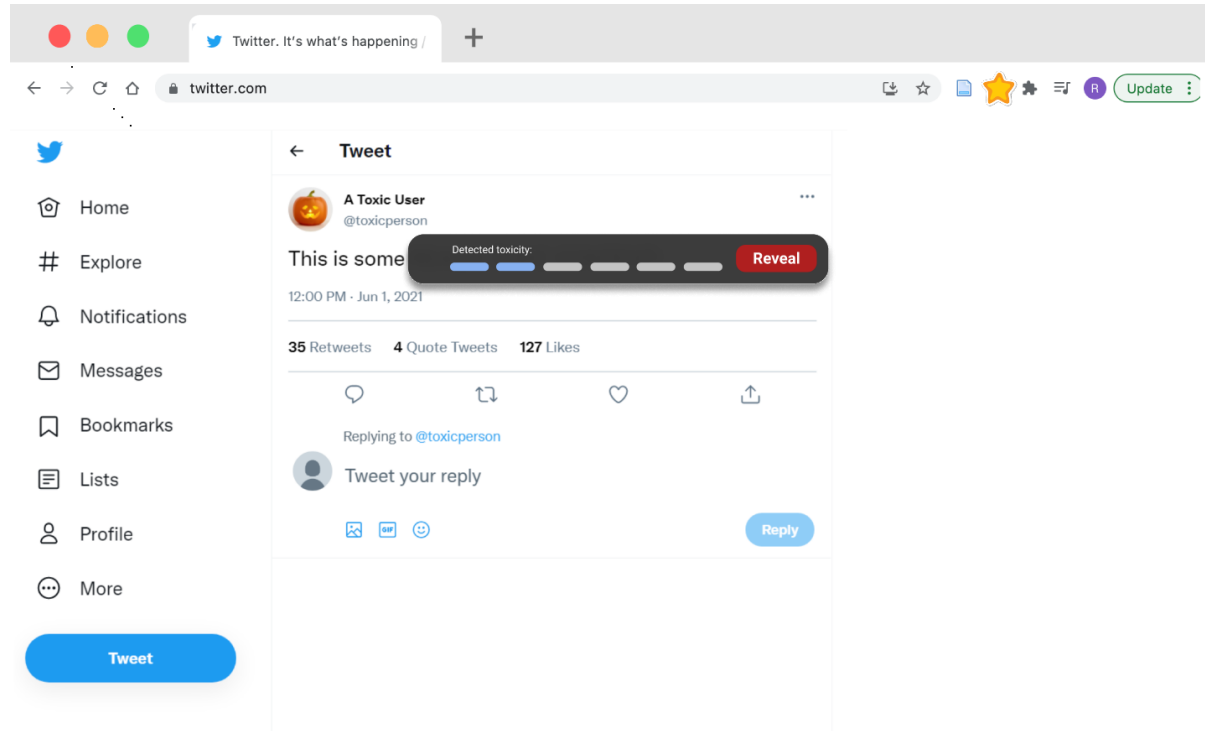


Fig. 12. Screenshot of the tool prototype when user hovers on blurred toxic content when they are reading.

9 USER EVALUATION

9.1 Method

In this user evaluation, we evaluated the first two user requirements: 1) the time required for the user to edit (before sending) their toxic message into a less toxic message (that maintains a similar meaning) should be less than manually deleting and entering each character; 2) the time required for the user to delete the toxic content in their message should be less than manually deleting each character. Based on our understanding of the context of use, our study aims to elucidate how the proposed tool changes the time needed to edit or delete a message, given the need for efficiency in emotional situations. Specifically, if a user is experiencing an emotional reaction, lowering the barrier required to make a change to their message can produce less toxic content. While it is possible to simply force edits automatically, which would undoubtedly be faster than manually changing the content, we also abide by user requirements 4, 5, and 6 to be sure the user can still decide their own actions. To evaluate the two user requirements, we put users in the scenario that they were typing toxic replies on Twitter and decided to edit or deleting the toxic content in their message. We asked users to complete the same task in both a high-fidelity prototype of our plugin and the original Twitter interface and compare the time taken to complete the tasks.

9.2 Apparatus

We asked the participants to complete the task in both our tool and the original input environment. For the toxic content mediating tool condition, we created a high-fidelity prototype of our plugin based on Twitter input

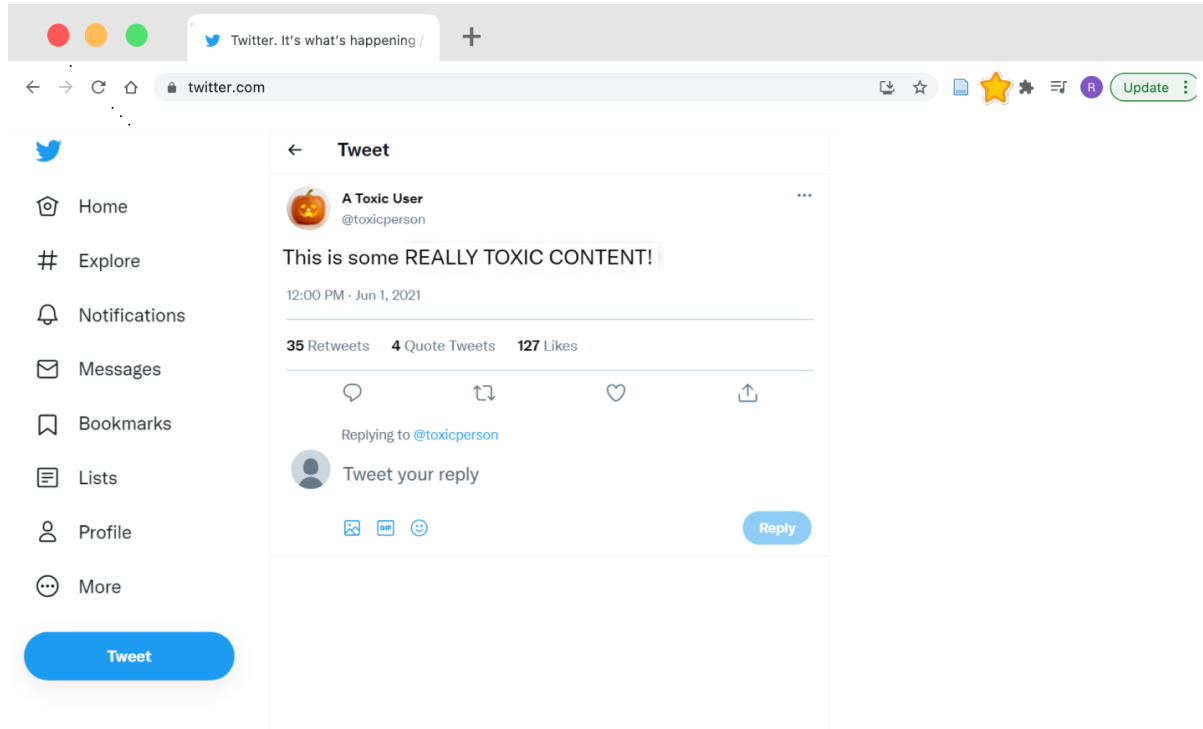


Fig. 13. Screenshot of the tool prototype when user click on "reveal" to read the original content.

context which suggests edits for toxic content using *Figma*. In this condition, users can view the suggestions to change their toxic content and click to replace them. In the condition that users use the original Twitter interfaces, we opened up a Twitter site for them to edit the same toxic reply in the original interface. Given *Figma*'s limited capabilities, we manually timed both the automated change and manual change of the toxic message. While this may introduce some variability, we expect our multiple experiments to average out the introduced noise.

9.3 Tasks and Procedures

Before participants start the task, they were informed of the study process, and that the study will involve toxic content. All participant gave verbal consent for participating in the study. Participants were asked to complete two tasks. For the first task, they were asked to edit three toxic replies to make them less toxic. In the second task, they were asked to delete toxic contents from three replies. Each task was completed both using the interactions provided by prototype of the plugin situated in Twitter and the original Twitter interface. Three participants first completed the task in the prototype of the plugin and then completed in the original interface, and two participants first completed the task in the original interface and then in the prototype.

All six toxic content were extracted from the dataset of the <https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge> in Kaggle, and the researchers came up with the fixes for the toxic comments.

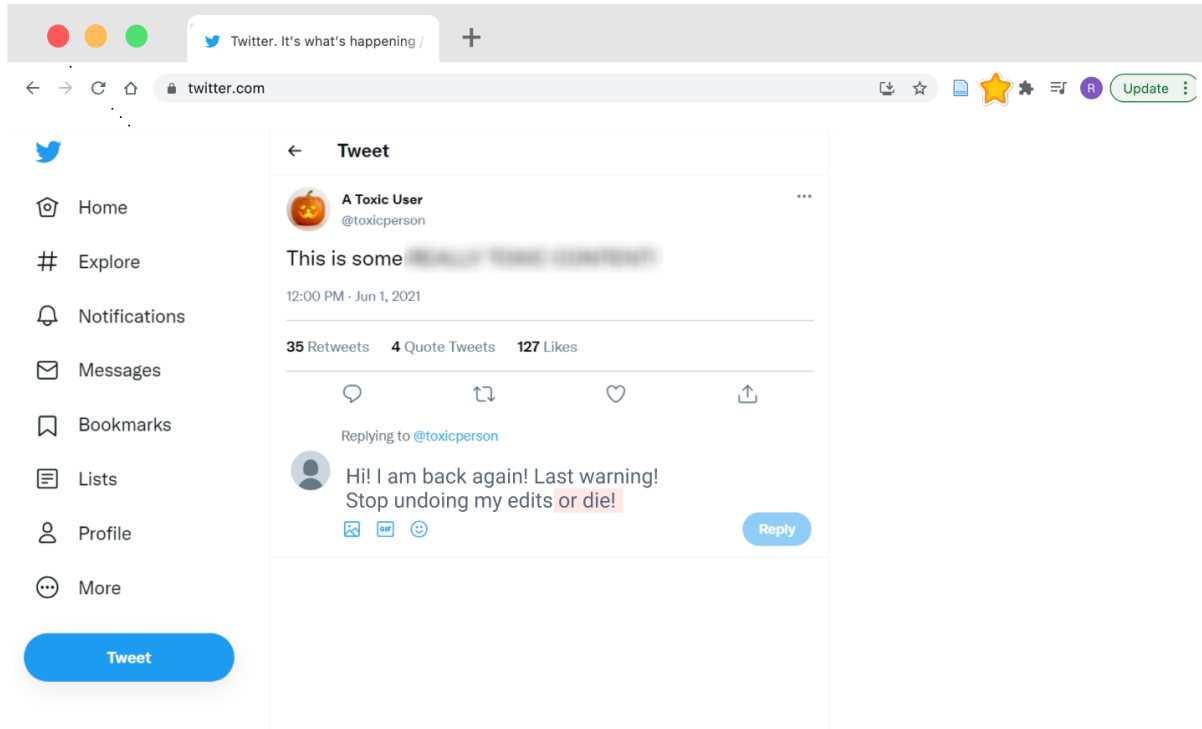


Fig. 14. Screenshot of the tool prototype when users' writings include toxic content.

9.4 Participants

We recruited ten participants. The average age of the participants is 23.4. Five participants identified as female, and five participants identified as male.

9.5 Results

We performed a repeated measurements ANOVA to determine if participants performed significantly different between manually changing their content on twitter and by using our tool. We analyzed the scenarios where users had to delete content and come up with a more positive edit separately. For each condition, we gathered data for 30 trials, as each participant complete three sentences in both deleting and fixing the content. The speed that each user performed the deletion and edit were averaged across the three writing samples to smooth out the signal during the statistical test. While we didn't account for the length of the edit in the average (some people may have chosen to edit with a longer phrasing), most of the edits were of comparable size, meaning we don't expect much deviation across each the individual writing samples. Our statistical analysis showed that participants were able to edit their content much faster with the proposed tool, requiring on average $5.76 (\pm 2.71)$ seconds to automatically edit versus $9.69 (\pm 3.01)$ seconds to manually edit, with $p\text{-value} = 0.0011$ and $F\text{-value} = 22.1363$. Given the $p\text{-value}$ is less than 0.05, we can confidently reject the null hypothesis and assume there is a statistical significance between the automated and manual editing. In addition, since we are only comparing against two groups, we do not need any post-hoc analysis and can confidently assert the automated tool speeds up toxic message editing. The clear separation between the two methods can be seen in 19.

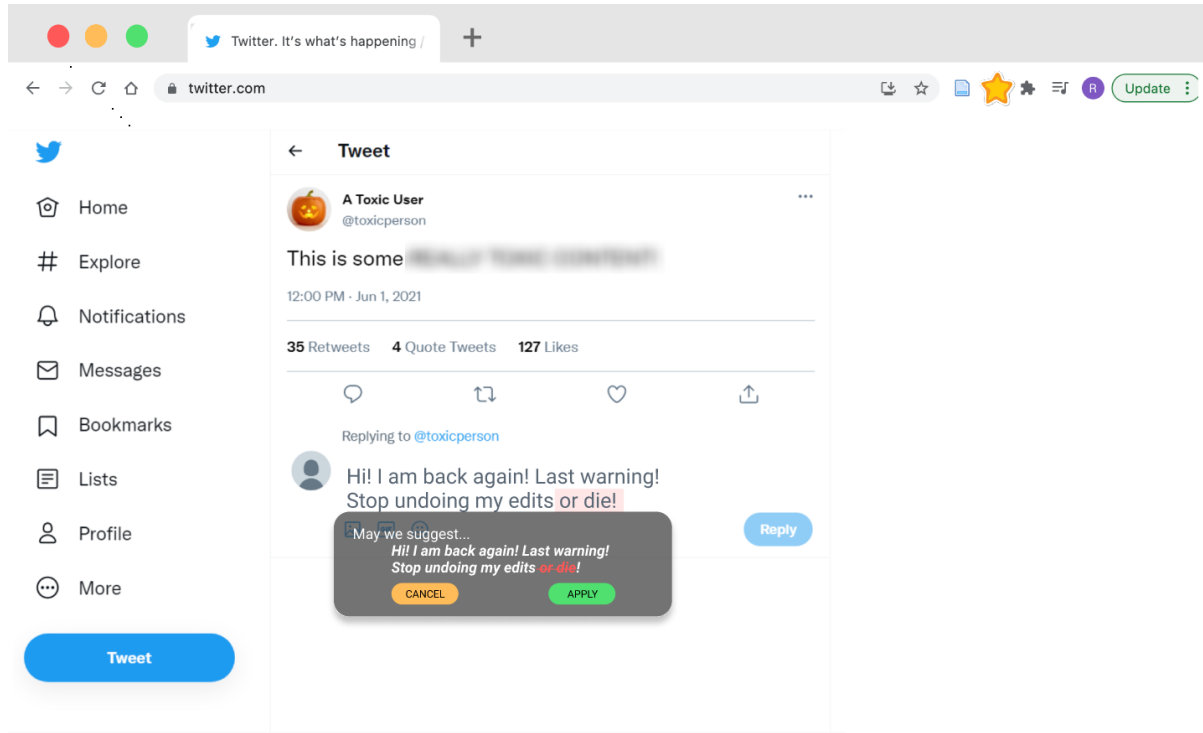


Fig. 15. Screenshot of the tool prototype when user hovers on the toxic content in their writing.

As for the deletion case, using the tool to delete toxic content took participants on average $6.53 (\pm 3.60)$ seconds, a bit faster than the manual case of $8.50 (\pm 2.76)$ seconds. That said, our repeated measurements ANOVA indicated that the tool did not reach a statistically significant difference compared to the manual process; p -value = 0.0926 and F -value = 3.5408. We suspect that this could be due to the ease of performing a deletion – it only requires the user to highlight and delete the content – whereas the editing required users to both formulate and type out a response. The distributions can be seen in 20, and while there is some indication of separation, we can see significant overlap. One hypothesis is that the deletion case is under powered, especially since, intuitively, the manual variant requires significantly less work than the editing scenario creating a more nuanced difference between methods. We perform a power analysis of our ANOVA test for the deleting scenario, only reaching a power of 0.40, indicating we do not have enough data to reliably interpret our p -value. Moreover, to reach a power of 0.8, we would need roughly 23 data points, over double what we collected. In comparison, the fixing scenario had a power of 0.96 and only required roughly 7 samples.

10 DISCUSSION

Based on prior work, we had identified two potential directions to explore within our research project to mitigate toxicity. One approach is to nudge users away from toxic content as they are typing it, by highlighting the toxicity in the content [2, 13]. The other approach would go a step further and actually provide suggestions on how the text could be modified by using techniques such as style transfer or paraphrasing [5, 6, 15]. The responses to our survey suggest that users would be more accepting of a solution that makes them aware of the toxicity in

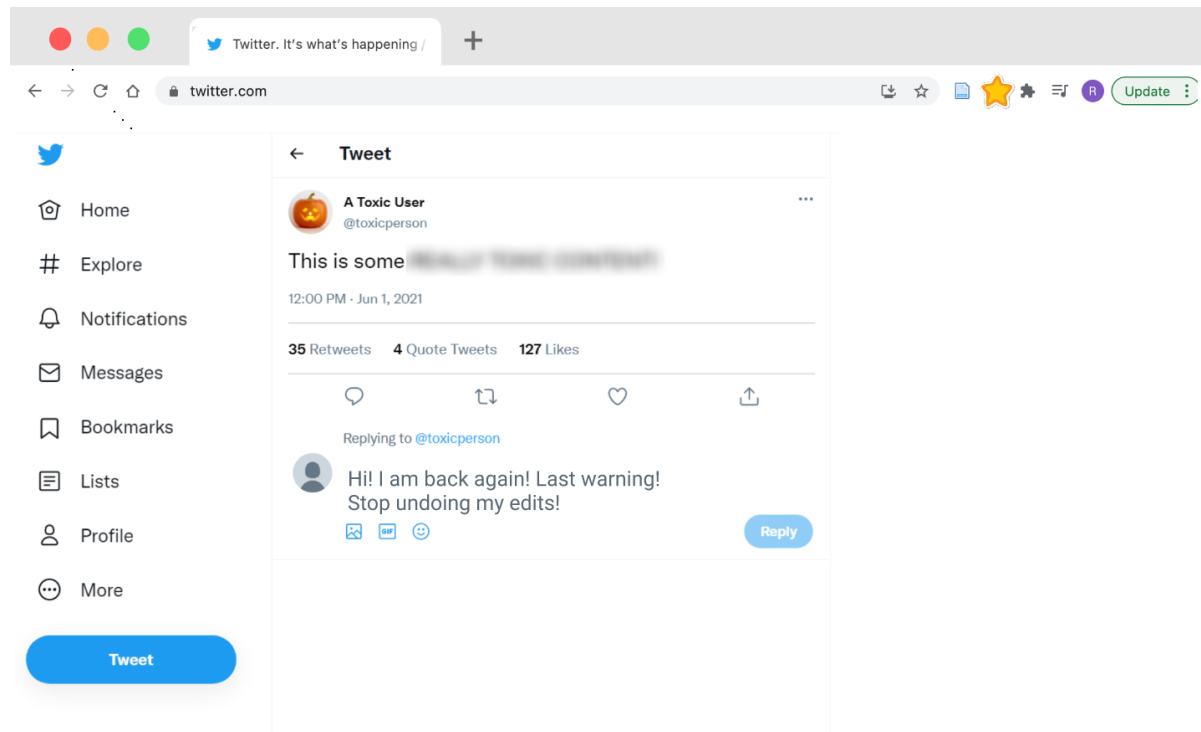


Fig. 16. Screenshot of the tool prototype when user click on "apply" to accept the suggestions from the tool.

their language rather than providing suggestions on how to reduce toxicity. This helped us adjust our scope to focus on solutions which only notify a user of toxic language, rather than trying to correct their toxic language. While this path may be the most readily available of the two mentioned previously, users emphasize that it is important for them to know why their content was toxic. Furthermore, the survey responses suggest that users do not trust tools that can identify toxic language, and we need to develop an understanding of why that trust is lacking and how can we address that gap in our design. Some potential reasons for why users lack trust may stem from privacy concerns behind toxicity detection and censorship [8] or demographic bias issues with hate-speech detection [19] (particularly racial or gendered bias). This further emphasises the need for tools which are both decoupled from the actual product (e.g. a chrome plug in that sits on the user side) as well as models that can offer explanations for their predictions.

10.1 Contextual Interview Insights

The contextual inquiry provided more insights to the gaps in users' current browsing and typing experience regarding toxic content. We identified two sub-goals that are crucial to the main goal of sending messages when users are typing in online platforms through the interviews: users try to avoid being targets when they are posting content, they try to avoid being offensive. Likewise, we identified one sub-goal when receiving toxic messages: having some external source to validate their feelings. We discuss each one of these in more detail and relate them back to our questionnaire, as well as how they might inform requirements for our design.

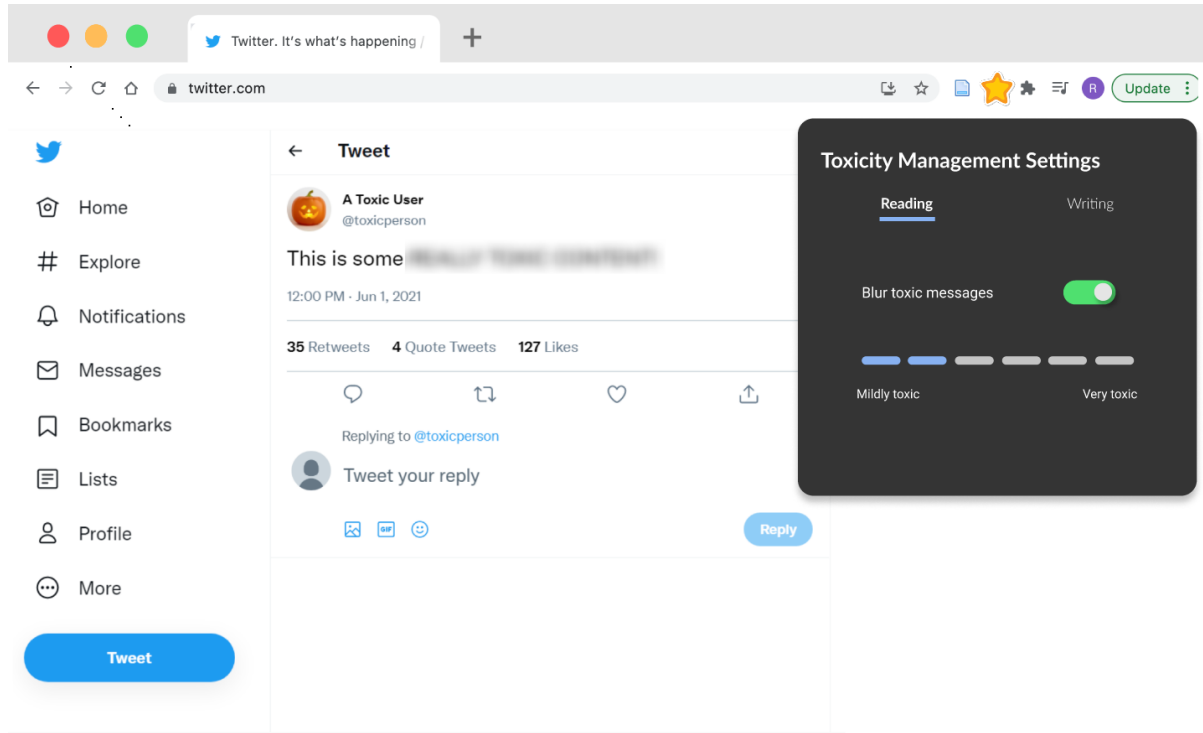


Fig. 17. Screenshot of the plugin settings for reading.

As described by many participants, one of the main reasons that they pay additional attention to their tone and language is that they do not want to become targets of arguments or receive toxic comments. To achieve this goal, users spend time and effort going through the content they have typed multiple times to ensure what they are about to send is "safe" and will not cause trouble. That said, often it is difficult to assure a message satisfies these criteria, and can often still be subject to toxic replies. This indicates that the current typing interfaces in online platforms have a gap of feedback on how likely users will become targeted because of their shared content. Thus, one focus of our solution could be giving user feedback on the likelihood that their language will become targeted. This result is further supported by our questionnaire, which argued that people would appreciate receiving feedback on their language and want to understand when they are being toxic. Similarly, when users are responding to content, they generally do not want to be offensive, and when they are offensive, it is largely unintentional or in response to something even more offensive. This presents two possible remedy scenarios, either the first extremely offensive content could have not been present which would have prevented retaliation, or the user could have been notified that their unknowingly toxic was in fact toxic. Again, this is supported by our questionnaire for similar reasons as above.

On the other hand, when receiving a toxic message, many users discussed avenues to have their feelings validated as well as to determine if there were problems with their original post or response, if applicable. Some users mentioned sharing the content with a friend or family member to get their opinion, while others considered the moderating system built into the platform. In particular, when discussing moderation, participants favored user moderated content rather than automated moderation out of fear that the automated system would be unable

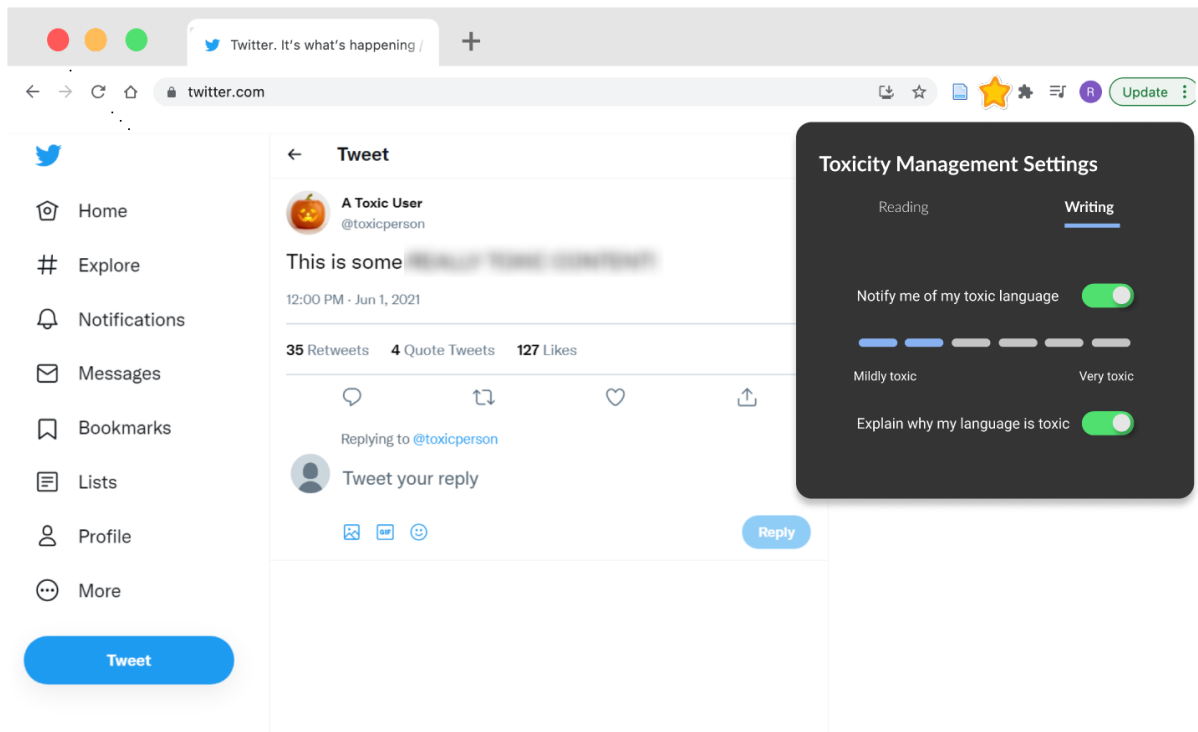


Fig. 18. Screenshot of the plugin settings for writing.

to understand the toxic behavior. Together, these two scenarios of sharing with friends and user-moderated systems argues people prefer to vet their content with systems they trust. With this in mind, it will be imperative we properly scope how much influence an automated system would have over a user's text to limit its power and not dissuade users.

Further consolidation in the affinity diagram reveals several overarching themes and goals of user behavior in the face of toxic content. First, users interact with toxic content in three main ways: reading it when it is not directed at them, sending toxic content to others, or replying to toxic content written by others. As mentioned above, users take certain cautions to avoid interacting with toxic content. This avoidance can take form in one of three ways: avoiding reading toxic content, avoiding sending toxic content, and avoiding being the direct targets of other people's toxic language. However, if such interaction cannot be avoided, users then seek help from either the platform content moderation team by reporting the toxic content, or gathering community support offline or online by showing toxic content to a friend. Finally, users can then reflect on the toxic interaction via feedback on their toxic content, users may receive no feedback, or users may experience various degrees of guilt for discomfort following toxic interactions. These insights will further guide the use cases for the development of toxicity averting tools.

10.2 Prototyping Insights

We utilized the critiques from individual prototypes to inform the design of the final toxicity system prototype. Sketch 2 provided an explanation system for a given toxic tweet. We improved this design to be more clear in

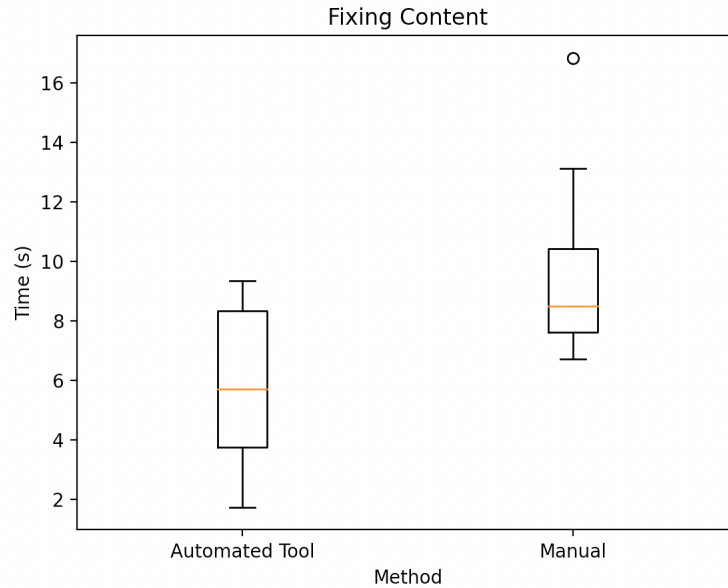


Fig. 19. Box plot which highlight the distribution of average times to fix content with the automated and manual processes.

explaining system capabilities. System 3 displays a toxicity censoring system for users who do not wish to engage with toxic content while browsing. From rounds of critiques, we determined the censoring of the entire content space was too heavy handed, and instead opted for the fine-grained blurring of specific toxic texts instead of the entire message. However, double clicking to reveal the content will remain as it is a good way to confirm the user truly wants to reveal toxic content (as opposed to hovering to reveal toxic content). Sketches 4 and 5 combined to create the moderation side system for our system. Sketch 4 displays the moderator's view for reported toxic content. Sketch 5 contains a toxicity sensitivity meter. Initially this meter was designed to be used by the user to determine what level of toxicity censoring they desired. However, critiques determined the levels of toxicity would be too subjective for end-user usage. Instead, we integrated this toxicity meter into the final design for moderator usage in the content moderation process. In summary, we heavily leveraged design team feedback from critiques to consolidate and improve upon our initial designs. The final prototype includes a toxicity explanation, censoring, and moderation system. This is suitable for the user profile of a user who browses toxic content and the moderator who maintains the platform of toxic content.

10.3 Usability Evaluation

Extensive usability evaluation of our system prototype revealed overlapping problem areas in system set-up, reading toxic messages, and writing toxic messages. In setting up the system for initial use, users found the toxicity bars to be confusing and its measurements subjective. Users do not trust the system to contain an objective evaluation of what is mildly toxic to what is extremely toxic, and thus have a difficult time setting the bar to what they believe their toxicity preference would be. When it comes to the system censoring of toxic content, users found the double clicking (to reveal detected toxic content) to be burdensome and unintuitive. Additionally, hovering over the blurred text was not an intuitive action for users and this caused users to have issue discovering the hover feature. Lastly, users experienced difficulty with the writing toxic content task due to lack of understanding of the pop-up. Users experienced confusion with popup options and could not connect the

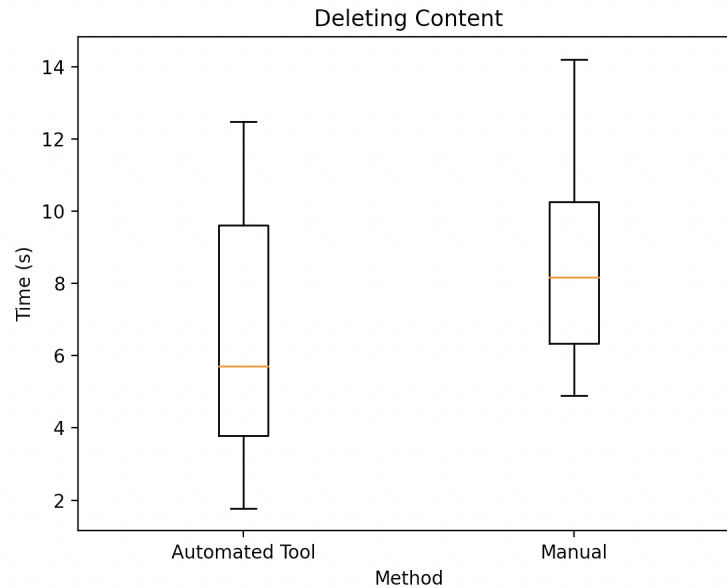


Fig. 20. Box plot which highlight the distribution of average times to delete content with the automated and manual processes.

actions of potential changes to their content. We believe some of the above usability errors could be corrected with a higher-fidelity prototype. For example, a higher fidelity prototype would be able to prompt users to hover or double click. However, the confusion over the pop-up capabilities would require more descriptive words or button design. Results from heuristics evaluation support our findings in the usability studies. We narrow down an extensive list of new user requirements as a result of this round of user studies, including the need for a simple and intuitive way of modifying toxic content writing and explanations of why such content would be toxic to others.

10.4 User Evaluations

In the final stage of our research, we created high fidelity prototypes of the toxicity editing system and performed quantitative user analysis. During the design and creation of high fidelity prototypes, we considered previous feedback from usability testing. We ensured the design of the platform had minimal chance to cause confusion and instruction from the interface remained clear. We conducted the analysis for two different activities: toxic language editing and toxic language deletion. Our findings support that editing time was significantly reduced with our toxicity detection tool. The fact that our statistical hypothesis testing and power analysis showed statistical significance, and strong power, allows us to confidently trust the results. However, the time difference for the deletion task was not statistically significant, which is not terribly surprising given the relative ease of the task to begin with. Our power analysis indicates that if we had collected more data, we may have been able to find a more statistically significant difference, but no exact conclusions can be drawn.

11 CONCLUSION AND FUTURE WORK

Toxic language remains a key problem in online communities given the turmoil it can inflict on other individuals. In this study we set out to explore users' attitudes towards toxicity online and methods for mitigating such behavior. By performing this work, we intend to design a system that can improve the online user experience for individuals impacted by toxic language, ultimately providing a safer and more inclusive environment. Through an online questionnaire we gained several important insights which enlighten our goal. We found that users can be affected in a wide range of ways when they receive toxic behavior. However, we found alignment on how users prefer to be notified of their potentially toxic behaviors and their general attitude of distrust towards toxicity detection models. These insights inform the design of future systems, specifically the need to build a transparent toxicity detection system that is designed with great care towards collaboration and human trust of the system. This involves creating a design that is informed from text, visual, and user-interface perspectives to best support interpretability and give agency to users' to be notified of and correct their potentially toxic behavior.

To further solidify the results from our questionnaire and better understand what features users would be interested in, we performed five contextual interviews. The contextual interviews helped further provide valuable insight into the goals, processes, and mental models of users when faced with potentially toxic language. We locate several overlapping themes from our participants: the allocation of time and effort to ensure they are not contributing to toxic language (receiving or giving), the utility of toxic language to prevent certain dialogue, and the high-contextual dependence of toxicity moderation and detection. Participants 1, 2, and 4 note they spend longer time editing and reading their posts to prevent toxic interactions. As mentioned previously, this effort is motivated by the desire for less toxic or non-toxic interactions. Participants 2, 3, and 4 noted toxicity keeps a group closed, forming an "in-group" vs "out-group" social landscape. This observation directly defines the utility of toxic language. Lastly, participants 2, 3, and 5 observed that whether someone choose to utilize a toxic language detector is highly dependent on the intent, severity, and context of the situation. This observation motivates our future work for explainable and context-dependent designs. While useful to analyze each participant individually, we also considered a consolidated analysis where interpretations were collected in an affinity diagram and individual sequence and flow diagrams were combined into one comprehensive diagram. By doing each of these three steps, we were able to better identify trends that existed throughout all of our participants. One broad theme we determined was that users need flexibility in how they handle toxic language, ultimately giving rise to a handful of user requirements which provide different avenues to interact toxic behavior. In particular, users should still have the ability to send, edit, or delete their message despite the circumstances of the toxic content. Without this flexibility, users may feel restricted by the system and be less likely to use the recommendations provided by it. In addition, users expressed that they can have very different reactions depending on their toxic experience. With this in mind, we identified that having a flexible intervention mechanism after the event is very important to address the needs of the users. For example, users should have the ability to both mitigate the toxic language they see organically, as well as remove any toxic language they put onto the system. With these specific goals in mind, we intend to determine how these would operate in practice and further refine the necessity of the goals.

Despite gaining better understanding on how people organically interact with toxicity, there are negative results in our current questionnaire and surveying data that need to be addressed. The major one revolves around our initial intention to target users from Piazza. Despite this initial focus, roughly 80% of the responses from the survey claimed they never seen toxic language on Piazza before. This may be due to a sampling bias of not having enough users who have used Piazza before, but our results indicate that even those who have used Piazza don't display a significantly higher exposure rate to toxicity (see 2). In the future, to fully rule this out, we will adopt survey distribution methods that are more targeted to our desired audience. This finding was further reinforced by our contextual interviews where each individual noted they had either not seen toxic language

on Piazza, or were unfamiliar with Piazza all together. That said, each user had previous concrete experiences with toxic language in sites they frequent, helping us to realize that a tool that helps mitigate toxic language likely needs to be platform agnostic to be useful. Given toxic language is either not prevalent on Piazza, or we do not have enough data to adequately determine a reasonable context of use surrounding Piazza, we will plan to re-scope our platform to instead contain any online forums where anonymity is present, with a sub-focus being sites that focus on learning. We will ground this scope in social networking websites that our participants noted experiencing toxic language in, such as Reddit, Facebook Marketplace, and Nextdoor. We intend to maintain and ground this sub-focus of learning environments in the fact that the usage of toxic language in an academic or learning environment will likely have more severe consequences. Specifically, it may significantly impact one's academic, social, and mental well-being. In further studies, it would be useful to also ground this impact directly through data which quantifies how toxic language can dissuade students from learning and the subsequent outcomes.

The prototyping process allowed us to further consolidate existing insights gathered from the contextual interviews. Persona analysis provided us with two sets of users: the content browser and content moderator. We created storyboards to determine usage scenarios for the system. In the prototype critique process we consolidated on individual designs and created the final prototype based on previously discovered user needs. We designed a toxicity censoring interface that hides toxic content in a fine-grained manner, and a toxicity moderation system that allows for human-in-the-loop intervention and determination of toxicity. This combined moderator and user facing system will then be tested in context to be further improved.

We discovered more points drive our design during the usability evaluation phase. Users found certain parts of the system to be unintuitive and difficult to on-board. For example, users found the toxicity bars to be subjective and had difficulty trusting the toxicity detection levels and had trouble using the pop-up bars for writing toxic language. These pain points were iterated on and refined before the final prototype was completed. Specifically, we simplified the toxicity bar component of the system as users didn't find it super informative and it only created confusion. Furthermore, we added a reveal button to minimize how often a user would make a mistake based on ambiguous UI design.

Lastly, we created high fidelity prototypes informed from the previous user evaluations. We then performed user evaluations comparing the use of our tool to the use of the existing Twitter web interface, measuring how quickly a user was able to edit and delete toxic content. Given the need to provide simple and efficient intervention mechanisms, we determine speed to be one helpful quality to enable a lower barrier to entry when modifying toxic content. Our statistical analysis of our data determined that the tool allowed for quicker and swifter editing of toxic content. However, the deletion case was determined to be statistically insignificant and under powered when performing a post-hoc power analysis. In the future, we would recruit more participants for the deletion case in order to determine if our insignificant result was due to no actual relationship, or not enough data.

ACKNOWLEDGMENTS

We thank all participants for filling out the survey and participants for their time in the contextual interview. We also appreciate the valuable feedback on survey items given by our friends who completed the pilot survey.

REFERENCES

- [1] [n.d.]. Perspective API. <https://perspectiveapi.com/>.
- [2] 2021. Twitter Rolls out Improved 'Reply Prompts' to Cut down on Harmful Tweets.
- [3] Michael Bernstein, Andrés Monroy-Hernández, Drew Harry, Paul André, Katrina Panovich, and Greg Vargas. 2011. 4chan and/b: An Analysis of Anonymity and Ephemerality in a Large Online Community. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 5.
- [4] Jithin Cheriyan, Bastin Tony Roy Savarimuthu, and Stephen Crane. 2021. Norm Violation in Online Communities – A Study of Stack Overflow Comments. In *Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XIII*.

- (*Lecture Notes in Computer Science*), Andrea Aler Tubella, Stephen Crane field, Christopher Frantz, Felipe Meneguzzi, and Wamberto Vasconcelos (Eds.). Springer International Publishing, Cham, 20–34. https://doi.org/10.1007/978-3-030-72376-7_2
- [5] Cicero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. 2018. Fighting Offensive Language on Social Media with Unsupervised Text Style Transfer. *arXiv:1805.07685 [cs]* (May 2018). [arXiv:1805.07685 \[cs\]](https://arxiv.org/abs/1805.07685)
 - [6] Liye Fu, Susan Fussell, and Cristian Danescu-Niculescu-Mizil. 2020. Facilitating the Communication of Politeness through Fine-Grained Paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 5127–5140. <https://doi.org/10.18653/v1/2020.emnlp-main.416>
 - [7] Jace Hargis and Torus Washington II. 2019. TODAY’S DISCUSSION BOARDS: THE GOOD, THE BAD, AND THE UGLY. *New Horizons in Education* 9 (July 2019).
 - [8] Yiqing Hua, Armin Namavari, Kaishuo Cheng, Mor Naaman, and Thomas Ristenpart. 2021. Increasing Adversarial Uncertainty to Scale Private Similarity Testing. *arXiv preprint arXiv:2109.01727* (2021).
 - [9] Mai Ibrahim, Marwan Torki, and Nagwa El-Makky. 2018. Imbalanced Toxic Comments Classification Using Data Augmentation and Deep Learning. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. 875–878. <https://doi.org/10.1109/ICMLA.2018.00141>
 - [10] Jin Woo Kim, Andrew Guess, Brendan Nyhan, and Jason Reifler. 2021. The Distorting Prism of Social Media: How Self-Selection and Exposure to Incivility Fuel Online Comment Toxicity. *Journal of Communication* jqab034 (Sept. 2021). <https://doi.org/10.1093/joc/jqab034>
 - [11] Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology* (1932).
 - [12] Shruthi Mohan, Apala Guha, Michael Harris, Fred Popowich, Ashley Schuster, and Chris Priebe. 2017. The Impact of Toxic Language on the Health of Reddit Communities. In *Advances in Artificial Intelligence (Lecture Notes in Computer Science)*, Malek Mouhoub and Philippe Langlais (Eds.). Springer International Publishing, Cham, 51–56. https://doi.org/10.1007/978-3-319-57351-9_6
 - [13] Kevin Munger. 2017. Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment. *Political Behavior* 39, 3 (Sept. 2017), 629–649. <https://doi.org/10.1007/s11109-016-9373-5>
 - [14] David Noever. 2018. Machine Learning Suites for Online Toxicity Detection. *arXiv:1810.01869 [cs, stat]* (Oct. 2018). [arXiv:1810.01869 \[cs, stat\]](https://arxiv.org/abs/1810.01869)
 - [15] Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W. Black. 2018. Style Transfer Through Back-Translation. *arXiv:1804.09000 [cs]* (May 2018). [arXiv:1804.09000 \[cs\]](https://arxiv.org/abs/1804.09000)
 - [16] Ashwin Rajadesingan, Paul Resnick, and Ceren Budak. 2020. Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 557–568.
 - [17] Naveen Raman, Minxuan Cao, Yulia Tsvetkov, Christian Kästner, and Bogdan Vasilescu. 2020. Stress and Burnout in Open Source: Toward Finding, Understanding, and Mitigating Unhealthy Interactions. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER ’20)*. Association for Computing Machinery, New York, NY, USA, 57–60. <https://doi.org/10.1145/3377816.3381732>
 - [18] Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury. 2019. Prevalence and Psychological Effects of Hateful Speech in Online College Communities. *Proceedings of the ... ACM Web Science Conference. ACM Web Science Conference 2019* (June 2019), 255–264. <https://doi.org/10.1145/3292522.3326032>
 - [19] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1668–1678. <https://doi.org/10.18653/v1/P19-1163>

A RESPONSIBLE RESEARCH: HUMAN SUBJECTS RESEARCH PROTECTIONS



B SURVEY AND QUESTIONNAIRE INSTRUMENTS

B.1 Initial Questionnaire Design

Introduction: Welcome to our survey! Following the definition at Perspective API, toxic language is a rude, disrespectful, or unreasonable comment that is likely to make someone leave a discussion. We are hoping to understand how toxic language affects people on online platforms, and what is the potential of using a text-entry tool as an intervention. This survey is completely anonymous and should take about 15 minutes to complete.

Focus 1: User's perception and more abstract thoughts on toxic language

- (1) Which platforms do you perceive to have toxic language? (Select all that apply)
 - (a) Twitter
 - (b) Piazza
 - (c) Stack Overflow
 - (d) Reddit
 - (e) Other: (please specify)
- (2) On which platforms have you witnessed toxic language? (Select all that apply)
 - (a) Twitter
 - (b) Piazza
 - (c) Stack Overflow
 - (d) Reddit
 - (e) Other: (please specify)
- (3) On a scale of 1-10, with 1 being extremely helpful, 5 being neutral, and 10 being extremely toxic, what would you rate your perceived toxicity of the average user of the product. Please use -1 if you refuse to answer. [Do this in context of each answer in question 1]
- (4) How strongly would you be affected by toxic language that is not necessarily directed at you, for example a comment on someone else's post?
 - (a) Not at all affected
 - (b) Not very affected

- (c) Mildly affected
 - (d) Somewhat affected
 - (e) Strongly affected
 - (f) non-applicable
 - (g) refuse to answer
- (5) How strongly would you be affected by toxic language if the toxic language was directed at you?
- (a) Not at all affected
 - (b) Not very affected
 - (c) Mildly affected
 - (d) Somewhat affected
 - (e) Strongly affected
 - (f) non-applicable
 - (g) refuse to answer
- (6) How strongly do you agree with the statement “Toxic content will always be present, in some form and to varying degrees, in online discussion”?
- (a) Strongly disagree
 - (b) Somewhat disagree
 - (c) Neither agree or disagree
 - (d) Somewhat agree
 - (e) Strongly agree
 - (f) non-applicable
 - (g) refuse to answer

Focus 2: User’s explicit experience with toxic language

- (1) You have been on the receiving end of toxic behavior in an online setting that disrupted your experience in the last year.
- (a) Strongly disagree
 - (b) Somewhat disagree
 - (c) Neither agree or disagree
 - (d) Somewhat agree
 - (e) Strongly agree
 - (f) non-applicable
 - (g) refuse to answer
- (2) You have corrected someone’s toxic language in an online setting in the last year.
- (a) Yes
 - (b) No
- (3) You have reported someone’s toxic language in an online setting in the last year.
- (a) Yes
 - (b) No
- (4) You have used toxic language towards someone in an online setting in the last year.
- (a) Strongly disagree
 - (b) Somewhat disagree
 - (c) Neither agree or disagree
 - (d) Somewhat agree
 - (e) Strongly agree
 - (f) non-applicable

- (g) refuse to answer
- (5) You would be willing to learn how to make your language less toxic.
 - (a) Strongly disagree
 - (b) Somewhat disagree
 - (c) Neither agree or disagree
 - (d) Somewhat agree
 - (e) Strongly agree
 - (f) non-applicable
 - (g) refuse to answer


Focus 3: User's thoughts on moderating toxic language

- (1) When someone is being toxic, acknowledging their behavior can be a strategy to improve subsequent behavior.
 - (a) Strongly disagree
 - (b) Somewhat disagree
 - (c) Neither agree or disagree
 - (d) Somewhat agree
 - (e) Strongly agree
 - (f) non-applicable
 - (g) refuse to answer
- (2) When someone is being toxic, correcting their behavior can be a strategy to improve subsequent behavior.
 - (a) Strongly disagree
 - (b) Somewhat disagree
 - (c) Neither agree or disagree
 - (d) Somewhat agree
 - (e) Strongly agree
 - (f) non-applicable
 - (g) refuse to answer
- (3) If you were told your behavior was toxic, you would feel appreciative regarding the feedback.
 - (a) Strongly disagree
 - (b) Somewhat disagree
 - (c) Neither agree or disagree
 - (d) Somewhat agree
 - (e) Strongly agree
 - (f) non-applicable
 - (g) refuse to answer
- (4) If you were told your behavior was toxic, you would feel defensive regarding the feedback.
 - (a) Strongly disagree
 - (b) Somewhat disagree
 - (c) Neither agree or disagree
 - (d) Somewhat agree
 - (e) Strongly agree
 - (f) non-applicable
 - (g) refuse to answer
- (5) If you were told your behavior was toxic, you would want to know why it was toxic.
 - (a) Strongly disagree

- (b) Somewhat disagree
 - (c) Neither agree or disagree
 - (d) Somewhat agree
 - (e) Strongly agree
 - (f) non-applicable
 - (g) refuse to answer
- (6) If you were on a product, you would prefer to know if other users have a history of toxic language.
- (a) Yes
 - (b) No
- (7) You believe it is the job of the product’s moderators to handle toxic language on (name of the platform chosen in Q1)
- (a) Yes
 - (b) No
- (8) You would trust a tool to determine whether language is toxic.
- (a) Strongly disagree
 - (b) Somewhat disagree
 - (c) Neither agree or disagree
 - (d) Somewhat agree
 - (e) Strongly agree
 - (f) non-applicable
 - (g) refuse to answer
- (9) You believe a tool which moderates toxic language would infringe upon your privacy.
- (a) Yes
 - (b) No
- (10) Imagine you just typed potentially toxic language, what would you prefer to happen?
- (a) Notification of toxic language.
 - (b) Suggestion of less toxic language.
 - (c) Automatic change of toxic language into less toxic language.

B.2 Final Questionnaire Design

0% Survey Completion 100%



Research Participant Information and Consent Form

Title of the Study: Toxic Language Online

Point of Contact: Questions Directed To: Brian Tang, (bjaytang@umich.edu)

Description of the research
Following the definition at [Perspective API](#), **toxic language** is a rude, disrespectful, or unreasonable comment that is likely to make someone leave a discussion. We are seeking to understand how toxic language affects people on online platforms and how a text-entry tool could serve as an intervention mechanism. This survey is completely anonymous and should take roughly 10 minutes to complete (18 questions).

What will my participation involve?
You will be asked a series of brief questions about your experiences with toxicity online and on Piazza. You will be asked to rate how strongly they agree/disagree with statements. At the end of the survey, you will be asked a question about your education background.

Are there any risks to me?
Any personally identifiable information (PII) will be deleted. These include IP address etc. This and all other collected information will be securely stored.

Are there any benefits to me?
There are no direct benefits to you.

Whom should I contact if I have any questions?
You may ask any questions about the research at any time.
If you have questions about the research after finishing today you should contact the researcher Brian Tang, (bjaytang@umich.edu)
Your participation is completely voluntary. Your participation in this survey indicates your consent.

PLEASE PRINT/COPY/SAVE THIS CONSENT FORM FOR FUTURE REFERENCE

→

- (1) How often have you witnessed toxic language on Piazza?
 - (a) Almost never
 - (b) Infrequently
 - (c) Sometimes
 - (d) Frequently
 - (e) Very Frequently
 - (f) I have never used Piazza
- (2) How strongly would you be affected by toxic language if the toxic language was directed at you?
 - (a) Not at all affected
 - (b) Not very affected
 - (c) Mildly affected
 - (d) Somewhat affected

- (e) Strongly affected
- (3) You have been on the receiving end of toxic behavior in an online setting that disrupted your experience in the last year.
 - (a) Strongly disagree
 - (b) Somewhat disagree
 - (c) Neither agree or disagree
 - (d) Somewhat agree
 - (e) Strongly agree
- (4) You have corrected someone's toxic language (both pointing out the toxic language and offering a suggestion) in an online setting in the last year.
 - (a) Strongly disagree
 - (b) Somewhat disagree
 - (c) Neither agree or disagree
 - (d) Somewhat agree
 - (e) Strongly agree
- (5) You have reported someone's toxic language in an online setting in the last year.
 - (a) Strongly disagree
 - (b) Somewhat disagree
 - (c) Neither agree or disagree
 - (d) Somewhat agree
 - (e) Strongly agree
- (6) You have used toxic language towards someone in an online setting in the last year.
 - (a) Strongly disagree
 - (b) Somewhat disagree
 - (c) Neither agree or disagree
 - (d) Somewhat agree
 - (e) Strongly agree
- (7) You would be willing to learn how to make your language less toxic.
 - (a) Strongly disagree
 - (b) Somewhat disagree
 - (c) Neither agree or disagree
 - (d) Somewhat agree
 - (e) Strongly agree
- (8) When someone is being toxic, pointing out their behavior can be a strategy to improve subsequent behavior.
 - (a) Strongly disagree
 - (b) Somewhat disagree
 - (c) Neither agree or disagree
 - (d) Somewhat agree
 - (e) Strongly agree
- (9) When someone is being toxic, correcting their behavior can be a strategy to improve subsequent behavior.
 - (a) Strongly disagree
 - (b) Somewhat disagree
 - (c) Neither agree or disagree
 - (d) Somewhat agree
 - (e) Strongly agree
- (10) If you were told your behavior was toxic, you would feel appreciative regarding the feedback.


- (a) Strongly disagree
 - (b) Somewhat disagree
 - (c) Neither agree or disagree
 - (d) Somewhat agree
 - (e) Strongly agree
- (11) If you were told your behavior was toxic, you would feel defensive regarding the feedback.
- (a) Strongly disagree
 - (b) Somewhat disagree
 - (c) Neither agree or disagree
 - (d) Somewhat agree
 - (e) Strongly agree
- (12) If you were told your behavior was toxic, you would want to know why it was toxic.
- (a) Strongly disagree
 - (b) Somewhat disagree
 - (c) Neither agree or disagree
 - (d) Somewhat agree
 - (e) Strongly agree
- (13) If you were on a product, you would prefer to know if other users have a history of toxic language.
- (a) Strongly disagree
 - (b) Somewhat disagree
 - (c) Neither agree or disagree
 - (d) Somewhat agree
 - (e) Strongly agree
- (14) You would trust a tool to determine whether language is toxic.
- (a) Strongly disagree
 - (b) Somewhat disagree
 - (c) Neither agree or disagree
 - (d) Somewhat agree
 - (e) Strongly agree
- (15) You believe a tool which moderates toxic language would infringe upon your privacy.
- (a) Strongly disagree
 - (b) Somewhat disagree
 - (c) Neither agree or disagree
 - (d) Somewhat agree
 - (e) Strongly agree
- (16) Imagine you just typed potentially toxic language, what would you prefer to happen?
- (a) Notification of toxic language.
 - (b) Suggestion of less toxic language.
 - (c) Automatic change of toxic language into less toxic language
- (17) If you are in school, what year are you in?
- (a) High School
 - (b) Undergraduate student
 - (c) Master's graduate student
 - (d) PhD graduate student
 - (e) Not in school
- (18) Have you used Piazza during your academics?

- (a) Yes
- (b) No

B.3 Anonymized and De-identified Questionnaire Data

Raw data shown in the table is represented through a 1-6 ordinal variable when it is a Likert or Likert-like question (Questions 1-15), 1-3 ordinal variable when it is a question referring to intervention mechanism (Question 16), 1-5 ordinal variable when question is referring to year in school/not in school (Question 17), and a binary variable if the answer choices are yes or no (Question 18).

0% Survey Completion 100%


UNIVERSITY OF MICHIGAN

Toxic Language Online

We are researchers from the University of Michigan - Ann Arbor in the U.S. In our research, we are interested in creating a text entry tool which detects and remedies toxic language.

To help us in our research, we need you to answer a set of questions regarding your experience with toxic language online. Additionally, we need your feedback regarding whether you think you would use such a tool.

At the end of this study, we will ask you to answer few demographics questions. Please answer those questions honestly. The answers to those questions will NOT impact your participation in this study.

→

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18
1	2	4	1	1	4	3	1	2	3	5	2	4	5	1	3	1	1	2
2	3	3	3	1	1	1	1	2	5	3	3	1	1	1	2	2	2	2
3	1	4	1	1	4	1	3	4	4	3	2	4	3	1	3	2	4	2
4	1	5	2	2	4	1	5	4	4	4	3	4	4	3	3	1	5	2
5	1	2	2	4	4	2	4	2	2	4	4	5	4	2	4	1	5	2
6	1	2	4	4	4	4	2	2	2	2	3	2	2	2	2	1	4	2
7	1	2	5	5	5	5	2	3	2	3	4	4	5	4	5	3	4	1
8	1	3	1	2	1	1	4	3	3	2	3	5	2	1	4	1	5	1
9	3	1	4	4	4	2	1	1	1	3	3	2	1	1	1	1	5	1
10	1	5	2	2	1	1	3	4	3	3	3	4	4	1	5	1	4	2
11	6	1	3	1	1	5	1	1	1	1	5	1	1	1	4	2	2	1
12	1	4	1	1	1	1	5	4	2	5	2	5	4	2	4	2	3	2
13	2	2	1	1	1	1	4	4	4	4	4	5	4	3	2	3	5	2
14	6	1	3	4	4	5	1	1	1	1	5	4	1	1	4	1	5	1
15	2	3	2	1	2	1	1	4	4	5	2	3	4	4	4	1	4	2
16	6	2	4	5	1	5	1	2	3	1	3	4	3	1	5	1	1	1
17	6	1	4	3	4	1	1	1	1	1	5	3	1	1	5	1	3	1
18	2	5	5	4	5	2	4	1	1	4	3	5	3	4	1	1	4	2
19	6	3	3	2	5	1	3	4	4	5	4	5	4	1	4	1	5	1
20	1	2	4	2	1	1	5	5	5	5	2	5	5	4	2	2	4	2
21	3	5	4	5	3	3	3	3	3	3	3	3	3	3	3	1	5	1
22	2	2	1	1	1	2	3	3	4	5	3	5	3	2	4	2	3	2
23	6	4	4	2	5	4	1	5	5	5	4	5	2	1	5	1	5	1
24	6	5	5	4	5	1	4	3	4	4	4	5	3	2	2	1	5	1
25	3	2	2	1	1	1	1	4	4	4	2	3	4	2	4	1	3	2
26	6	3	4	4	4	1	5	5	4	5	1	5	5	4	2	2	3	1
27	2	4	4	5	3	2	5	1	2	2	5	4	4	2	4	2	3	2
28	6	3	4	1	1	2	1	2	1	2	5	2	1	1	5	1	5	1
29	1	4	1	4	4	1	1	4	4	5	2	5	2	2	3	2	4	2
30	3	4	5	3	5	5	4	4	4	5	4	5	5	4	4	1	4	2
31	2	4	2	1	3	1	4	4	4	4	2	4	4	4	4	2	4	2
32	2	4	5	4	4	1	4	5	5	5	4	5	2	4	4	3	4	2
33	1	1	1	1	1	1	3	3	3	4	3	5	4	2	3	2	4	2
34	1	2	2	1	1	1	4	4	2	4	3	4	2	2	3	2	4	2
35	6	4	3	2	2	3	4	4	4	3	3	2	2	3	3	1	4	1
36	3	3	1	1	1	1	5	5	2	4	2	5	3	4	3	2	4	2
37	6	4	2	1	1	3	3	2	2	3	4	4	3	3	3	2	3	1
38	6	2	2	4	4	2	2	4	4	2	3	2	3	1	2	1	5	1
39	6	1	1	1	1	1	1	3	3	1	1	3	3	1	5	2	5	1
40	6	3	4	4	3	3	4	4	4	5	3	5	3	3	3	1	5	1
41	6	1	4	4	4	2	5	3	3	2	4	5	3	2	1	2	5	1
42	2	4	5	5	5	4	3	5	4	4	2	5	4	2	3	1	5	1
43	3	4	2	4	5	1	5	5	4	4	4	4	2	4	2	2	5	1
44	2	3	2	3	2	1	4	5	5	3	4	4	4	4	4	2	4	2
45	1	2	1	1	4	3	2	1	1	2	3	2	1	3	5	1	2	2
46	1	4	1	1	1	1	5	2	2	4	4	5	2	2	4	1	4	2
47	2	4	4	1	3	3	4	4	2	4	2	5	3	3	2	2	4	2
48	1	3	4	1	1	1	5	4	4	4	4	5	3	3	4	2	4	2
49	1	5	4	2	5	3	4	4	4	3	4	5	4	3	1	2	4	2
50	2	3	4	1	1	1	4	4	2	5	2	5	3	4	2	1	4	2

Fig. 21. Raw questionnaire data

No.	Question Description	Values					
		1	2	3	4	5	6
Q1	How often have you witnessed toxic language on Piazza?	Almost never	Infrequently	Sometimes	Frequently	Very frequently	I have never used Piazza
Q2	How strongly would you be affected by toxic language if the toxic language was directed at you?	Not at all affected	Not very affected	Mildly affected	Somewhat affected	Strongly affected	-
Q3	You have been on the receiving end of toxic behavior in an online setting that disrupted your experience in the last year.	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree	-
Q4	You have corrected someone's toxic language (both pointing out the toxic language and offering a suggestion) in an online setting in the last year.	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree	-
Q5	You have reported someone's toxic language in an online setting in the last year.	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree	-
Q6	You have used toxic language towards someone in an online setting in the last year.	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree	-
Q7	You would be willing to learn how to make your language less toxic.	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree	-
Q8	When someone is being toxic, pointing out their behavior can be a strategy to improve subsequent behavior.	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree	-
Q9	When someone is being toxic, correcting their behavior can be a strategy to improve subsequent behavior.	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree	-
Q10	If you were told your behavior was toxic, you would feel appreciative regarding the feedback.	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree	-
Q11	If you were told your behavior was toxic, you would feel defensive regarding the feedback.	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree	-
Q12	If you were told your behavior was toxic, you would want to know why it was toxic.	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree	-
Q13	If you were on a product, you would prefer to know if other users have a history of toxic language.	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree	-
Q14	You would trust a tool to determine whether language is toxic.	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree	-
Q15	You believe a tool which moderates toxic language would infringe upon your privacy.	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree	-
Q16	Imagine you just typed potentially toxic language, what would you prefer to happen?	Notification of toxic language.	Suggestion of less toxic language.	Automatic change of toxic language into less toxic language.	-	-	-
Q17	If you are in school, what year are you in?	High School	Undergrad	Master's	PhD	Not in school	-
Q18	Have you used Piazza during your academics?	No	Yes	-	-	-	-

Fig. 22. Labels of response values to questionnaire

C CONTEXTUAL INQUIRY

C.1 Individual Interpretations

Interpretations: U01

She opens piazza to read the discussion board when there is a doubt about a course component	She opens NextDoor to see pictures of pets or look for furnitures, but she sees stuff that becomes controversial. (for e.g., benign discussion about yard signs)	When she sees controversial stuff she doesn't feel like participating in it	She doesn't like when she sees things that she wouldn't say to a real person in a real life; such things make her withdraw from the discussions	She finds that there are no clear signs on what can be controversial
She sees that it takes one person to start a controversy on any topic; when they say something controversial, many other people start arguing...and then its just a back n forth	When she sees something controversial, she tries to stay out of it. Sometimes you she would go and talk about it with other people	When something is too problematic, she wants to go and talk to other people about it, as it helps in articulating the shock	When she goes to talk to someone else, she forgets about using nextdoor and starts doing something else	When she sees something problematic, she sometimes continues to scroll

When she reads problematic things online, it affects her real life as she becomes extra cautious about things she does.	Her typing activity on the app is usually limited to benign things like helping people with gigs, buying/selling stuff	She usually doesn't like to participate in more controversial things, because she doesn't see the point of it...	She worries that adding more comments in such a situation can work like adding fire to the fuel	Because of the nature of textual conversations, she is worried that her meaning may get lost and her text may be interpreted in a very different way
When she is looking for toxic posts, topic of the post is a good way to say whether something might have toxic comments	She further looks at the comments to see if the post on a given topic has become too problematic.	She reads a comment about christianity and how other people start arguing about it..	She concludes that she can't really participate in the topic without saying something that would become controversial	In this case, the topic of the post makes it hard to participate in a non-controversial way

She would adopt different strategies to try and not make it further problematic.	Strategy 1: If she had to contribute such a post, she would take a pause before typing anything, and also pause and read again before sending the comment	Strategy 2: she would phrase her message to focus on the topic and not the person	Strategy 3: She would think about how she might feel about posting it after a week and whether there is something that she would regret later on.	Strategy: She would also think about the potential audience who might read this and how they might interpret it.
She feels she has the self-control to be able to implement these strategies...	However, the strategies can be counterproductive as sometimes they make her not post at all or not post what she wanted to	Sometimes, she has regrets about not participating in a discussion, especially when the topic was important to her	She feels that self-regulating mechanisms can have a silencing effect as the pendulum can swing too far	She feels if she had the option to take back something she has already said, it would give her more confidence in posting.

Editing and deleting exist but she doesn't feel like using them because people can still take a screenshot and save it.	A save or preview option that would give you all the feel and experience of posting something (like a qualtrics survey) but not really post it would be helpful	Her current mechanisms make her regret more about not posting something than posting something wrong...	When she is typing to a large group, she usually has to do multiple edits before posting it.	When she reads something that she has already typed, the reflection often leads to new changes.
However, even editing after the reflection also doesn't make her very confident...its satisfying at best				

Interpretations: U02

has used piazza in the past for an especially difficult course.	cannot recall toxic conversations on piazza	but does remember many of the questions had a very anxious tone. Mentions it would be easy for small instances of toxic language to affect the people who made the question post	thinks that toxic language more likely comes from people writing anonymously	thinks toxic language will decrease if there classmates are not allowed to post as "anonymous to everyone"
she opens up facebook on mobile and mentions there is toxic language in facebook marketplace	she has both bought and sold items on marketplace	when buying an item, mentions it is very easy for anyone to send a message asking if item is available	however, very few people actually move forward	Last year she used marketplace to fill a sublease, and received dozens of responses per day

Last year she used marketplace to fill a sublease, and received dozens of responses per day	however very few people even responded back when she messaged back, it made her very frustrated	additionally people did not meet the basic requirements for the room. she felt angry people did not spend time reading the selling post	she had used rude language to show anger at a "buyer" who kept changing their mind	would take suggestions if she was only slightly annoyed, prefers to have the option of a less toxic phrasing or word
however, imagines that if someone was extremely angry and felt justified, the suggestions may not be helpful	this anxiety made her more careful about writing messages, only if she was sure she was interested in the item	fear of experiencing toxic language made her act more carefully to avoid toxic scenarios	would have used a tool at that time to construct message	however, that same toxicity and anxiety also self regulated away some annoying behavior - uninterested buyers, unclear posts

Interpretations: U03

U03-01: The user played a multiplayer video game to have fun.	U03-02: Someone uses toxic language.	U03-03: The recipient is affected by the toxicity.	U03-04: The user suggests the person the language is directed at to mute the perpetrator.	U03-05: The recipient might not mute others since there might still be useful information they missed out on.
U03-06: The recipient mutes the perpetrator.	U03-07: The user played a multiplayer video game to have fun.	U03-08: The user has toxic language directed at them.	U03-09: The user is unaffected by the toxicity.	U03-10: The user plays off the toxicity as a joke.
U03-11: The user is affected by the toxicity.	U03-12: The user mutes the perpetrator.	U03-13: The user reciprocates commonly used inappropriate language.	U03-14: The user offers to listen to the perpetrator's issues.	U03-15: The user takes a break.
U03-16: The user is annoyed by the actions or language of someone.	U03-17: The user uses exclusionary language ("Shut up", "Nobody cares").	U03-18: The recipient is unaffected.	U03-19: The user types a hurtful response listing the reasoning of why the user is annoyed.	U03-20: The user types whatever hurtful things come to mind.

<p>U03-21:</p> <p>The user sends the response.</p>	<p>U03-22:</p> <p>The user is upset with someone's opinions or actions.</p>	<p>U03-23:</p> <p>The user wishes to express their discontent.</p>	<p>U03-24:</p> <p>The user does not wish to come across as hateful or angry.</p>	<p>U03-25:</p> <p>The user types something very negative.</p>
<p>U03-26:</p> <p>The user alters the language to be more passive.</p>	<p>U03-27:</p> <p>The user sends the message.</p>	<p>U03-28:</p> <p>The user wishes to give advice to someone.</p>	<p>U03-29:</p> <p>The user types and sends out the non-aggressive advice.</p>	<p>U03-30:</p> <p>The recipient uses off-handed comments and shows displeasure.</p>
<p>U03-31:</p> <p>The user is affected by the toxicity.</p>	<p>U03-32:</p> <p>The user sees any more advice as repetitive and counterproductive.</p>	<p>U03-33:</p> <p>The user wishes to be funny.</p>	<p>U03-34:</p> <p>The user tells a joke.</p>	<p>U03-35:</p> <p>The recipient responds, offended by the joke.</p>
<p>U03-36:</p> <p>The user wishes to express an opinion or idea that they think is fine.</p>	<p>U03-37:</p> <p>The user types and sends the message.</p>	<p>U03-38:</p> <p>The message is offensive to someone.</p>	<p>U03-39:</p> <p>The message is misconstrued as offensive by someone.</p>	<p>U03-40:</p> <p>The recipient replies and shows their discontent.</p>
<p>U03-41:</p> <p>The user directly disagrees with someone's message.</p>	<p>U03-42:</p> <p>The user expresses their disagreement.</p>	<p>U03-43:</p> <p>The recipient replies and shows their discontent.</p>	<p>U03-44:</p> <p>The recipient labels the user as some negative construct ("People like you").</p>	<p>U03-45:</p> <p>The recipient brings up statistics supporting their message.</p>
<p>U03-46:</p> <p>Damage has been done from toxic language.</p>	<p>U03-47:</p> <p>The user feels guilty about the damage.</p>	<p>U03-48:</p> <p>The user reflects on the situation.</p>	<p>U03-49:</p> <p>The user realizes they hurt someone but didn't think much about it initially.</p>	<p>U03-50:</p> <p>The user wishes they were able to proactively stop this.</p>
<p>U03-51:</p> <p>Someone tries to help reduce the tension of the situation.</p>	<p>U03-52:</p> <p>Someone joins in to contribute.</p>	<p>U03-53:</p> <p>They get attacked for joining someone's side.</p>	<p>U03-54:</p> <p>Any relationship is temporarily or permanently damaged.</p>	<p>U03-55:</p> <p>Someone wishes to remedy a misunderstanding.</p>
<p>U03-56:</p> <p>They explain the misunderstanding.</p>	<p>U03-57:</p> <p>Damages to any relationships are salvaged.</p>	<p>U03-58:</p> <p>People become too afraid or uncomfortable to contribute to the discussion.</p>	<p>U03-59:</p> <p>At the end of the day, no one has fun after toxic language is used.</p>	<p>U03-60:</p> <p>Very often gender minorities plays a role.</p>

U03-61: When trying not to offend someone, language becomes less concise because you are afraid they'll take it the wrong way.	U03-62: The toxicity of asynchronous discussion forums and comments sections last much longer.	U03-63: Usually self-reflective post-mortem that you really consider the consequences.	U03-64: There are scenarios where there is a definitive "thing" that they shouldn't have said to cause everything.	U03-65: People who are intentionally being toxic don't need a browser extension.
U03-66: An extension that points out something is potentially toxic to people of a different background may be helpful.	U03-67: An extension that points out why something is potentially may be helpful.	U03-68: Can get some intrinsic value and thoughtfulness from realizing it may be offensive to other people.		

Interpretations: U04

U04-01: The user scrolls through the list of post to find a post they are interested in.	U04-02: The user reads the posts and replies and decides if the content is toxic.	U04-03: The user "lights up" (upvotes) the post or reply they think is good.	U04-04: The user wants to click "lights off" for (downvotes) the post or reply they think is toxic.	U04-05: The platform has a prestige score for each user, and to downvote a post or reply costs one prestige score.
U04-06: The user cannot downvote toxic content when they have zero prestige score even when she wants to.	U04-07: The user does not care so much about posts that share their view but has toxic language and keeps browsing.	U04-08: The platform sorts the comments by the number of "lights" (upvote minus downvote) from high to low by default.	U04-09: The user only views a few comments at the front under a post to avoid reading potentially toxic posts.	U04-10: The user does not see much toxic language as the platform automatically blocks some toxic words.

Interpretations: U04

U04-11: The user does not see much toxic language because moderators of forums remove controversial and toxic posts manually.	U04-12: The user normally just curse in their head when reading toxic posts instead of replies to them.	U04-13: The user would click into posts with potentially controversial or toxic topics to read into the content, as many posts has non-toxic content while they have potentially toxic title to attract people.	U04-14: The user cannot tell if the actual post is toxic by just by reading the tile in the post list.	U04-15: The first few comments under a post influences later comments, as they become the comments with the most "lights" when there are just a few comments.
U04-16: The user has an expectation of the toxicity level of the platform.	U04-17: If the user knows the platform is likely to contain toxic content, they are prepared when browsing and thus are less affected.	U04-18: If the user knows the platform is likely to contain a certain type of view that are toxic, they are less affected even they are targeted by the view as they are prepared.	U04-19: If the user reads toxic contents that targets a population group that they are part of, they would browse to see if there is anyone on their side.	U04-20: Knowing that some people on the internet are stubborn and their views cannot be changed, the user ignores what they say and do not involve in a conversation

Interpretations: U04

U04-21: The user would potentially get into arguments when they see toxic content.	U04-22: The user would try to make sure to back their posts with facts and be reasonable while getting in an argument.	U04-23: The user try to sound objective and reasonable to avoid becoming targets of toxic languages.	U04-24: If they have friends with different opinions on a subject matter, the user would avoid posting content that are on the other side.	U04-25: If they know that someone they know might see their comment on something and would get upset, they would avoid posting the comment.
U04-26: The user does not edit their comments if they are participating in a heated discussion and there are tons of similar comments, and thus their comments is not likely to be seen and targeted.	U04-27: The user avoid to get involved into arguments that potentially involves several rounds of back-and-forth arguments, but are okay if they are not expecting to be replied.	U04-28: The user sometimes wants to send something that is potentially harsh, then started to edit it to make it sound less harsh, but ended up not sending it.	U04-29: The user would recall their message in group chats.	U04-30: The user makes some jokes that might not be 100% appropriate

Interpretations: U04

U04-31: The user makes sure the other person sees that joke.	U04-32: The user then recall the message to avoid it staying in the chatting history.	U04-33: The user once sent message that are not intending to be offensive but learned from other people that the receiver felt bad.	U04-34: The user needs feedback to know if the message is harsh.	U04-35: If the user knows the message they sent is harsh before the recipient sees it, they will definitely modify the content.
U04-36: If the user knows the message they sent is harsh after the recipient sees it, they will not modify the content because the damage is done.	U04-37: The user looks for other ways to make up for their harsh language if the recipient reads something they sent that are unintentionally harsh.			

Interpretations: U05

U05-01: User would rarely write something toxic on their own	U05-02: Toxic messages would largely be in response to a previous toxic message	U05-03: The user's mood during the day would impact their level of toxicity when responding	U05-04: The user's mood can be impacted by other messages on the platform. More negativity/hostility can make them more likely to elicit similar responses.	U05-05: The user would respond to the message directly rather than message an individual privately
U05-06: Facilitating a safe space for others but openly calling out instances of toxicity is important to the user.	U05-07: The user is generally aware of their emotions in instances of replying to toxicity	U05-08: User is likely to disregard how the message may impact someone depending on the original level of toxicity.	U05-09: The user provides a response to call out the language	U05-10: When applicable, user tells them why the content is unacceptable

Interpretations: U05

U05-11: The user would only edit their statements once they have a response that shows remorse and if they deem they were too harsh.	U05-12: Harshness of comment would need to be determined by an external party to be useful.	U05-13: The user would apologize to the user in order to mitigate more negative behavior, again if too harsh.	U05-14: The user would be unlikely to delete their message.	U05-15: The user would take a break from the site.
U05-16: The user would be hesitant to return to the site if the experience was extremely negative, such as if it explicitly included racism, sexism, or other forms of discrimination.	U05-17: User would re-engage with site after a sufficient cool down period	U05-18: The user enters the platform and first determined if they have notifications on their posts	U05-19: If they have a community they frequent they prioritize replies made in that community	U05-20: If there are notifications, they go into conversations and see what is being said.

Interpretations: U05

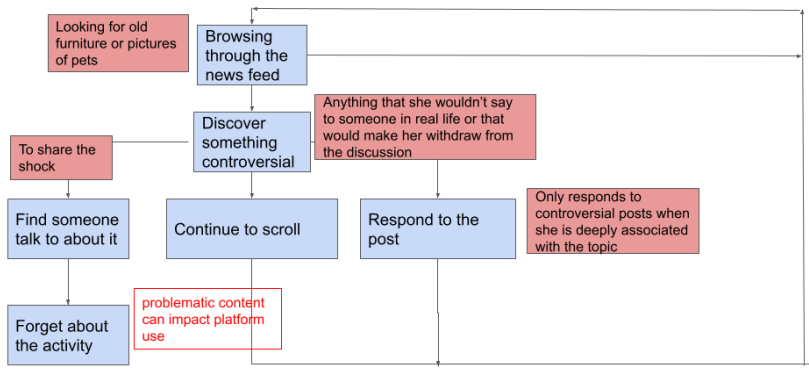
U05-21: Unhelpful comments on their posts are simply ignored	U05-22: If the comments were hurtful, the user would enter the post thread and interact with the content	U05-23: The user would double check their post to determine if they were at fault	U05-24: If unfairly attacked, user would respond to call out and correct behavior	U05-25: The user would assess the platforms reporting system to further handle the hurtful content
U05-26: User is more likely to report on sites that use a user-moderator since they are unsure if an automated moderator could properly moderate content.	U05-27: User would share content with friends or community pointing out the negative behavior	U05-28: If the other user were to apologize and edit their comment, the user would feel better regarding the situation	U05-29: Without apology, user would feel inclined to delete their original post given they felt belittled and hurt	U05-30: Instances of toxic replies are addressed as quickly as possible to set a precedent of acceptable behavior.

Interpretations: U05

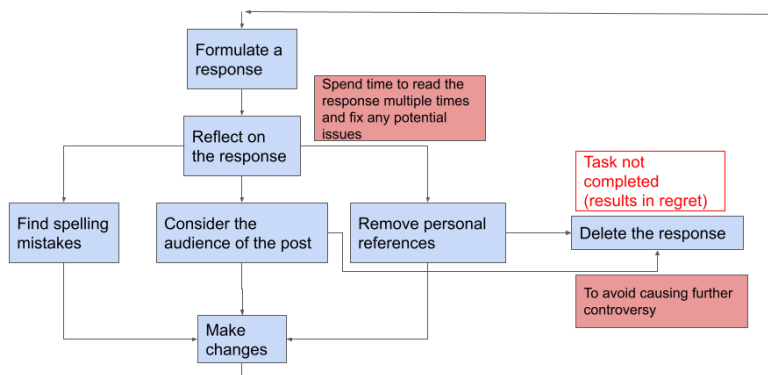
U05-31: User would feel less inclined to post on the community after a negative experience	U05-32: User would feel anxious during next post, worried about similar behavior in replies	U05-33: Numerous instances of toxic replies would cause the user to leave the community	U05-34: When seeing copious amounts of toxicity on a platform, user becomes less likely to browse and engage with community	U05-35: The user would migrate to groups where they felt safe
---	--	--	--	--

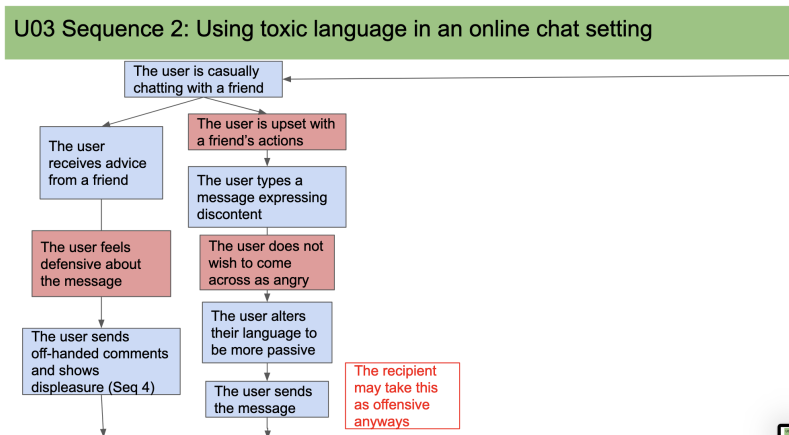
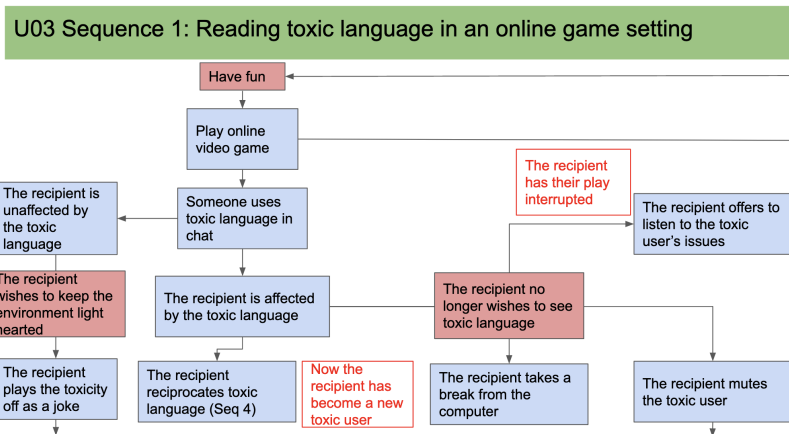
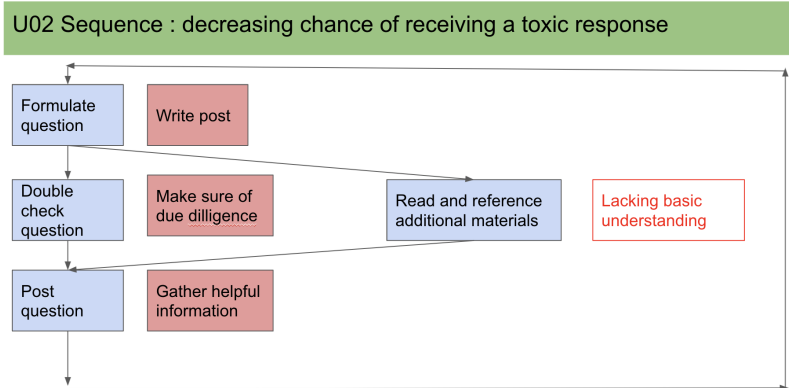
C.2 Individual Sequence Diagrams

U01 Sequence 1: Getting exposed to a controversial post

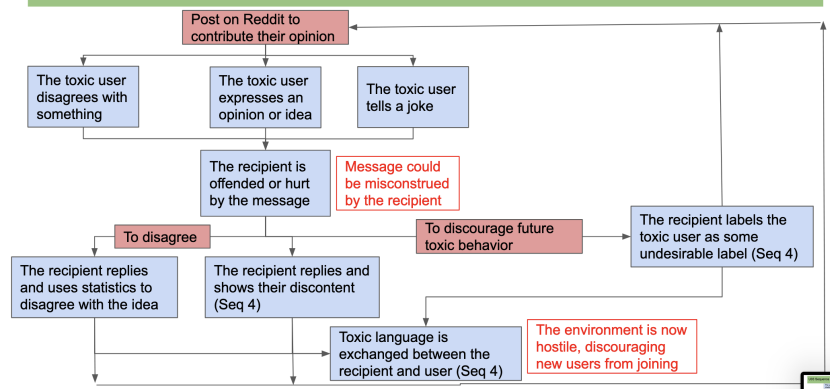


U01 Sequence 2: Responding to a controversial post

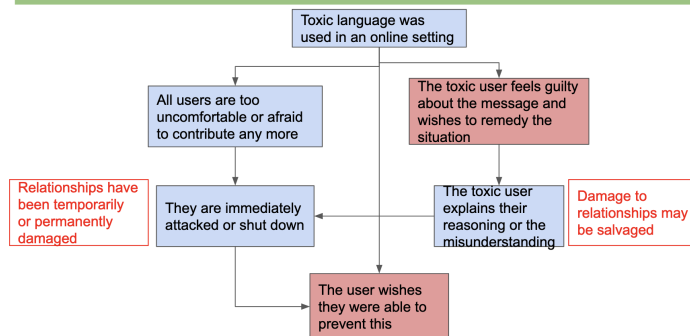




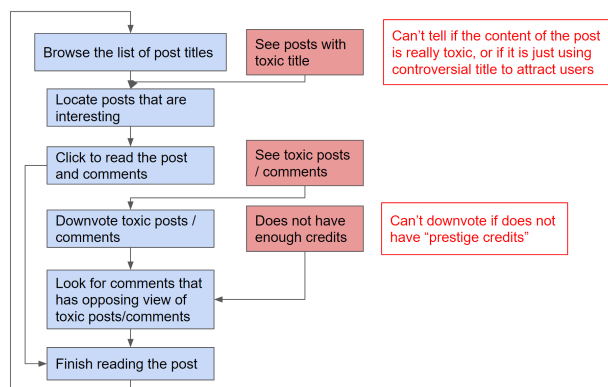
U03 Sequence 3: Using toxic language in an online forum setting



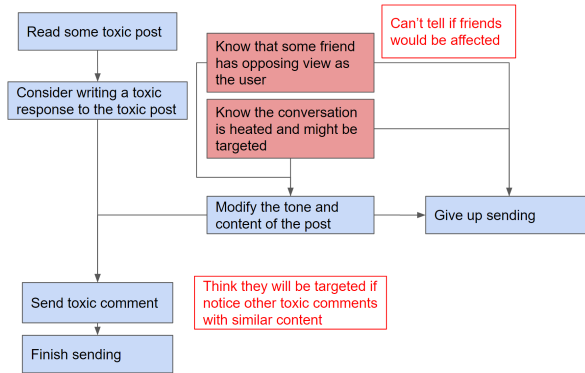
U03 Sequence 4: After toxic language has already impacted an online setting



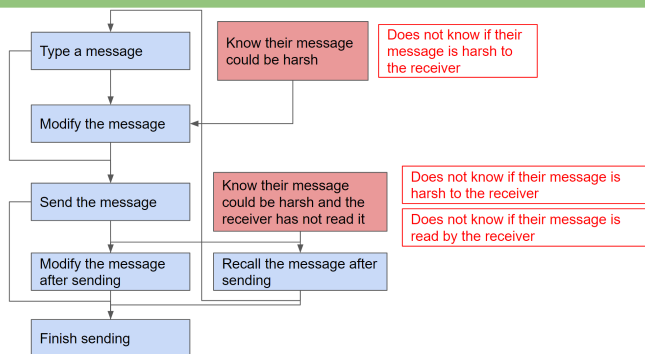
U04 Sequence 1: Reading toxic language in an online forum



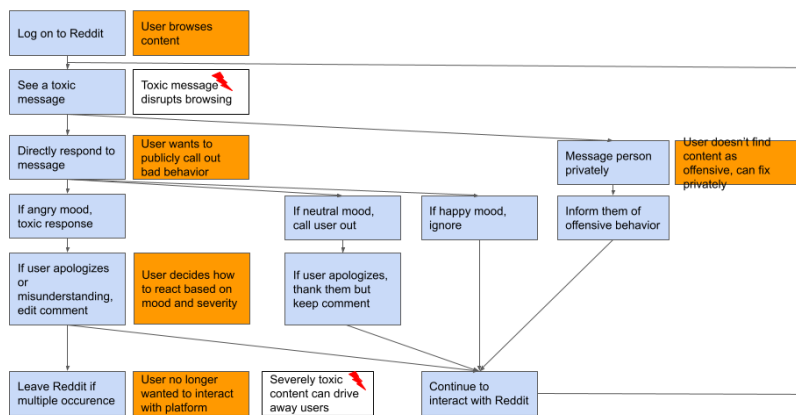
U04 Sequence 2: Involving in argument in an online platform similar to Twitter

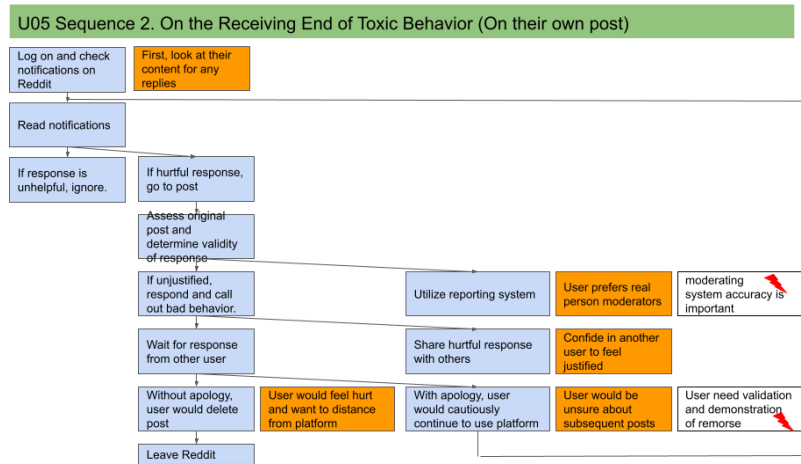


U04 Sequence 3: Talking in group chat and unintentionally sent harsh message



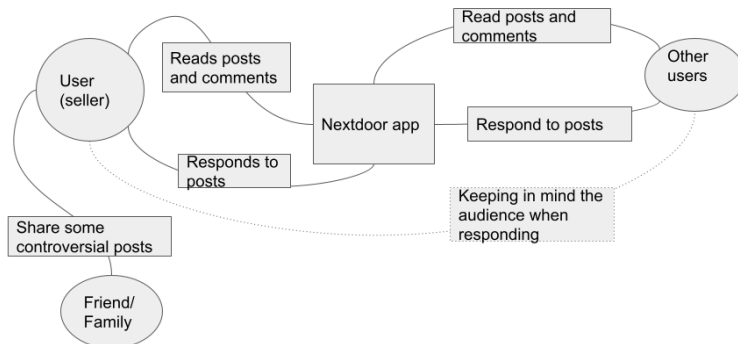
U05 Sequence 1. Reading Toxic Behavior (Undirected at User)



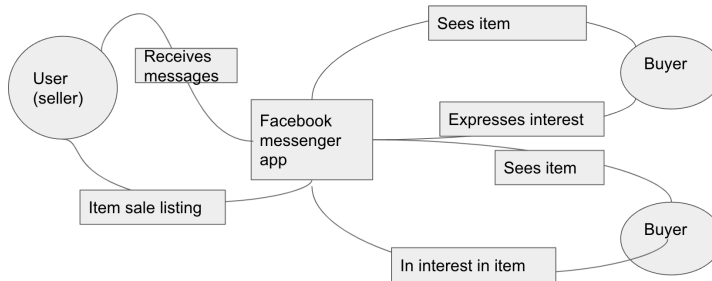


C.3 Individual Flow Diagrams

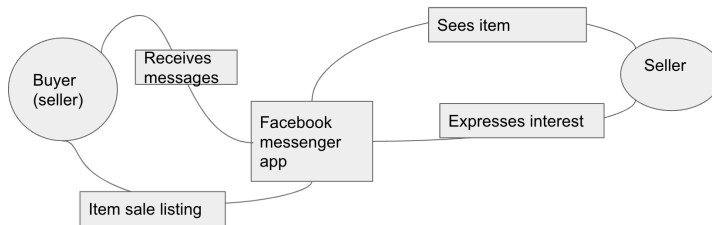
U01 Flow 1: User activity on nextdoor



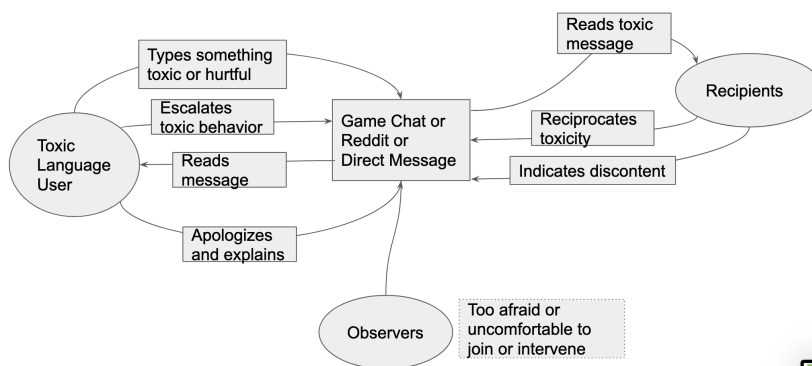
U02 Flow 1: seller experience on marketplace



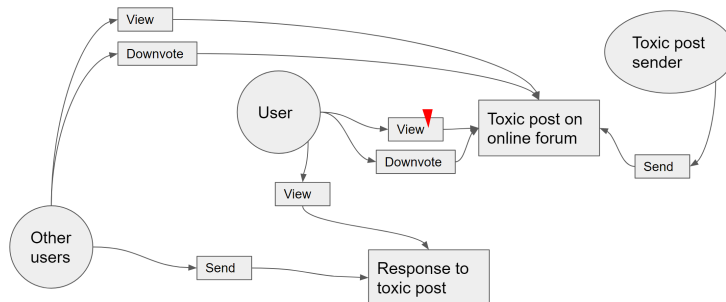
U02 Flow 2: buyer experience on marketplace



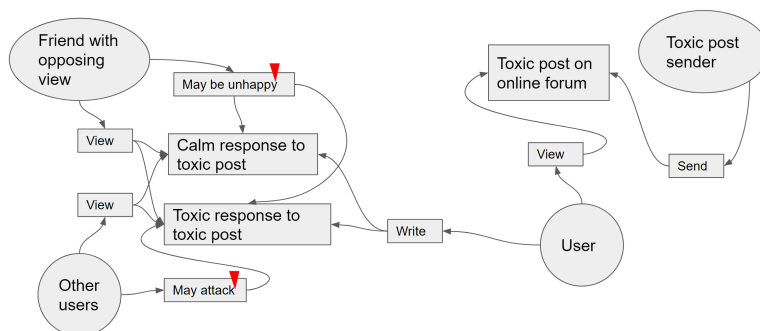
U03 Flow 1: Toxicity Cycle



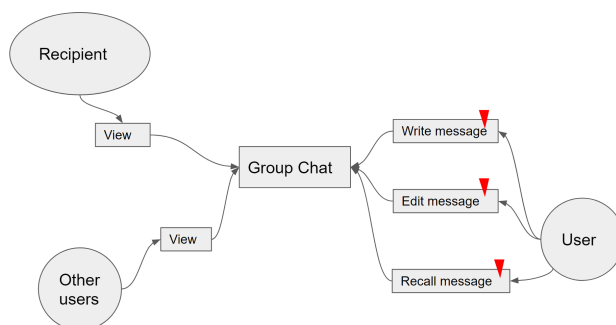
U04 Flow 1: Reading toxic posts



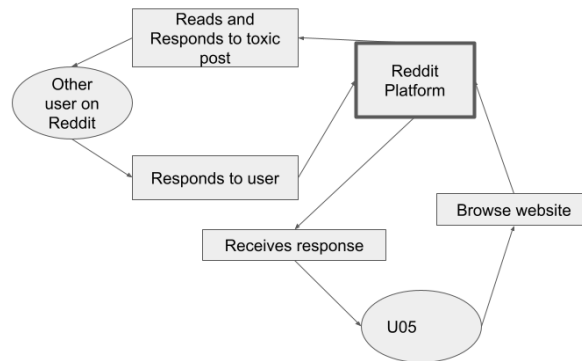
U04 Flow 2: Involving in argument in an online platform similar to Twitter



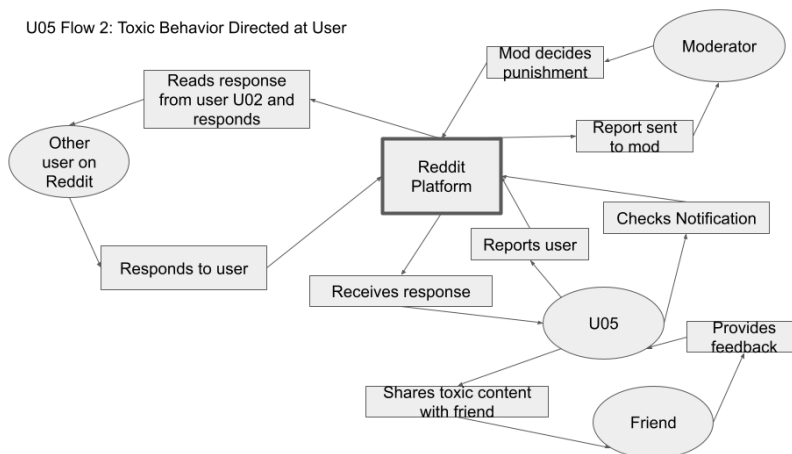
U04 Flow 3: Talking in group chat and unintentionally sent harsh message



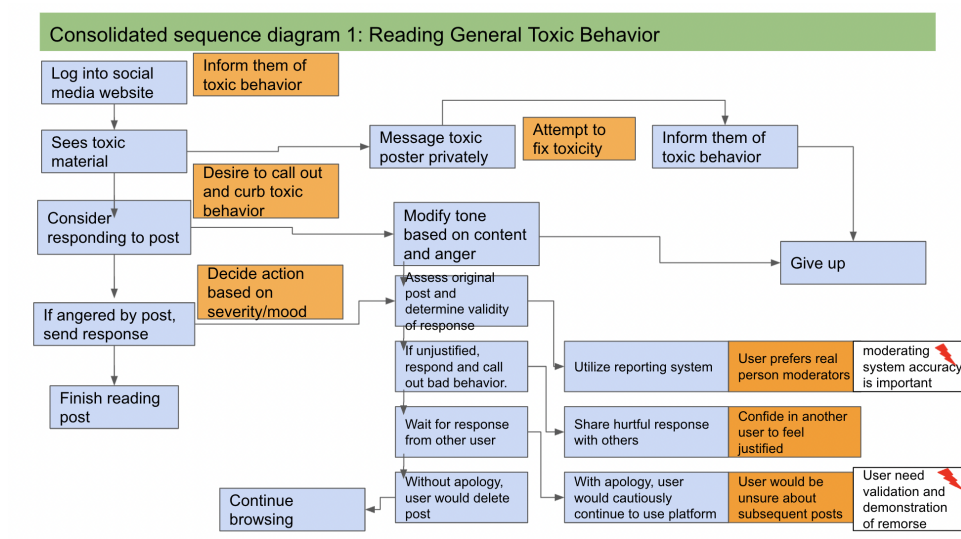
U05 Flow 1: Toxic Behavior Undirected at User



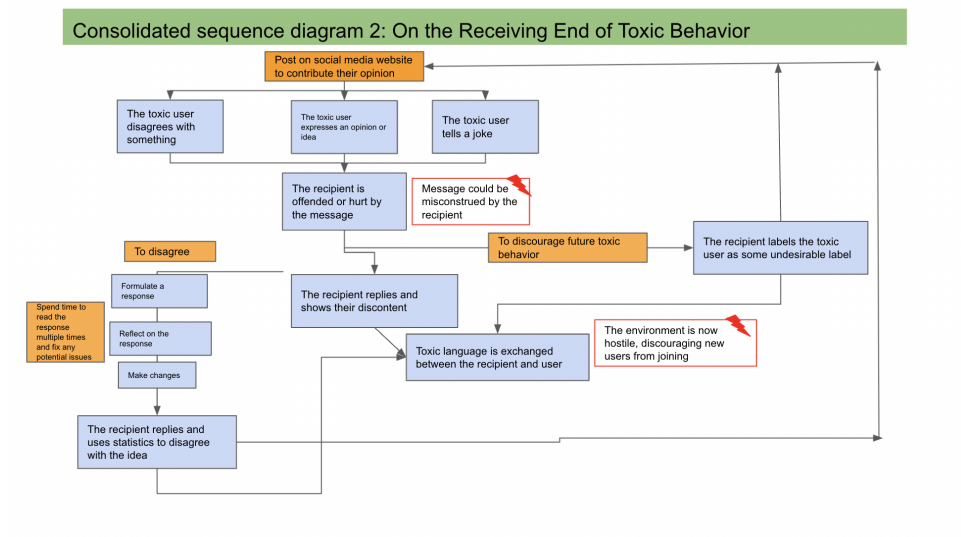
U05 Flow 2: Toxic Behavior Directed at User



C.4 Consolidated Sequence Diagrams

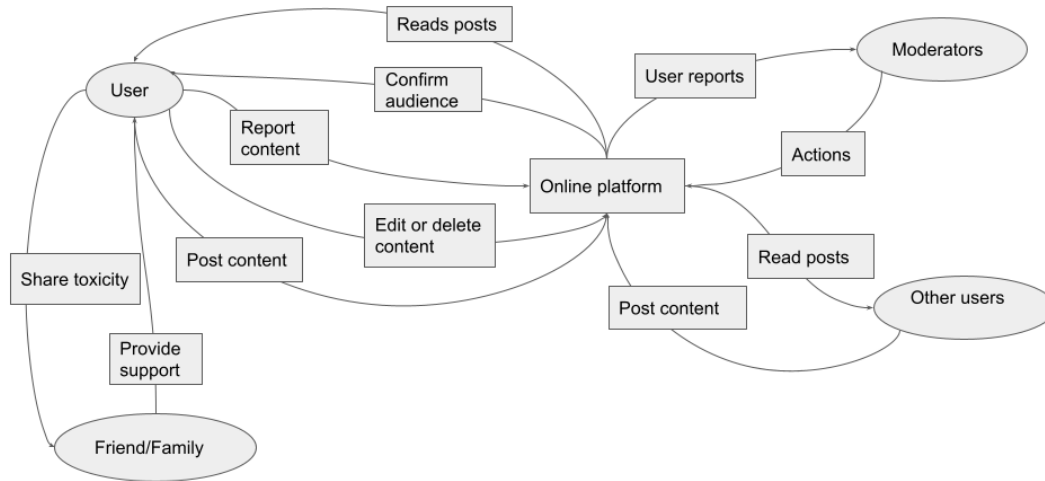


Consolidated Sequence Diagram - Reading Toxic User-generated Content



Consolidated Sequence Diagram - Responding to Toxic Content

C.5 Consolidated Flow Diagrams



Consolidated Flow Diagram

C.6 Affinity Diagram

We created our [affinity diagram](#) through Miro.

D LOW FIDELITY PROTOTYPES

D.1 Individual Personas

- (1) Persona 1: This person is a high school student, 17 years old. They spend a lot of time on social media sites such as Twitter and Instagram. They are used to seeing toxic content online and see the posters as just online trolls.
- (2) Persona 2: This person is a working adult, 26 years old. They spend a lot of time debating in forums based on which video games are the best. They have strong opinions that Nintendo creates the best games and will often get into fights to defend them. That said, sometimes they get too angry and are trying to be better about it.
- (3) Persona 3: This person is an older adult, 40 years old. They usually find themselves on social media apps such as Facebook. They are relatively mature and have a lot of life experience, and thus do not get too bothered by seeing toxic content online.
- (4) Persona 4: This person is in college, 21 years old. They are the student body president and like to keep things in line. They usually report content online and have recently been promoted to the moderator of their site. That said, it is hard for them to go through all of the content and could use some help.
- (5) Persona 5: This person is a graduate student, 25 years old. They loves animals and generally likes to help people. They often browses her neighborhood social media app, looking for pictures of cute animals or any requests for odd jobs. They despises toxicity and believes that she would never write something online that she cannot say to a friend in real life.

D.2 Individual Sketches

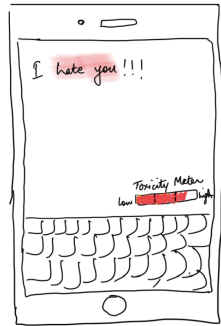


Fig. 23. Sketch 1: A mobile plugin that displays the toxicity rating of the text as you type while also highlighting the part of text that may be contributing to the toxicity

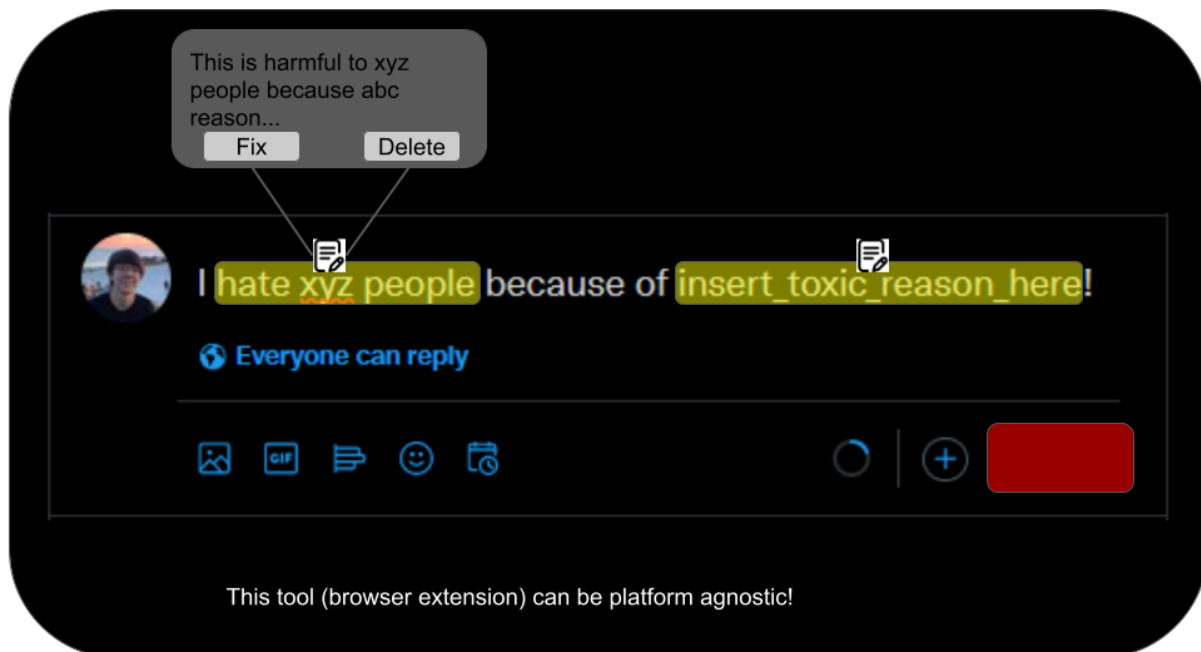


Fig. 24. Sketch 2

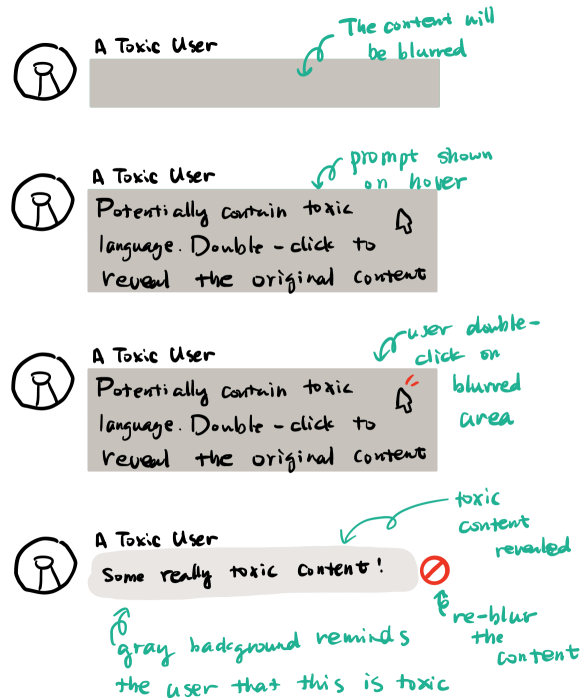


Fig. 25. Sketch 3

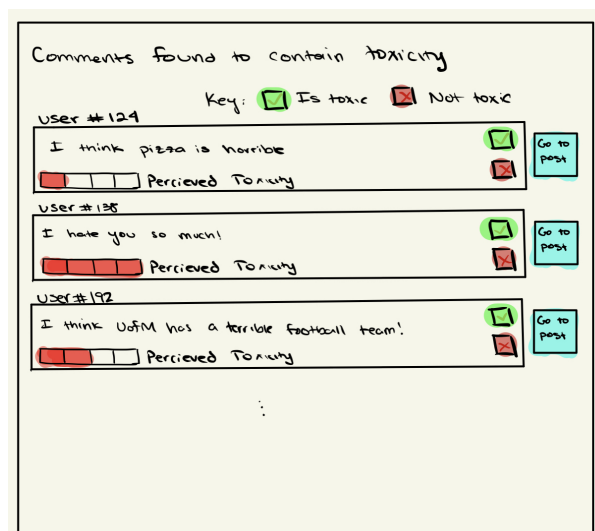


Fig. 26. Sketch 4

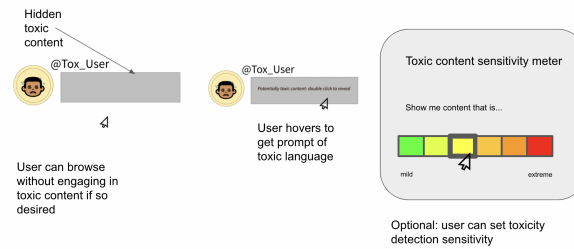


Fig. 27. Sketch 5

D.3 Individual Storyboards

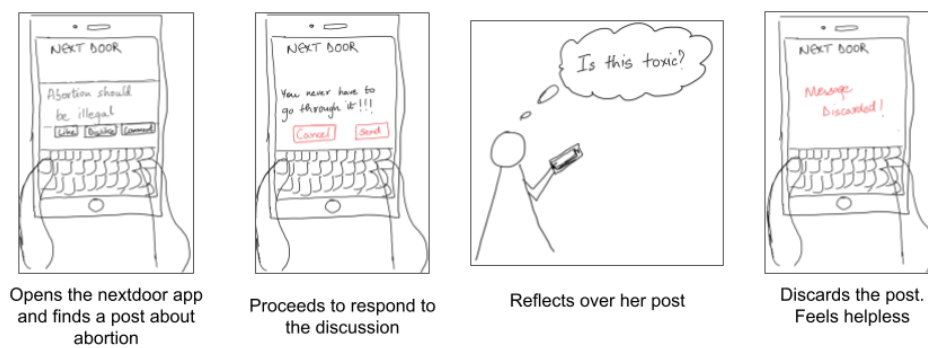


Fig. 28. Story Board 1: Emma wants to participate in a community discussion about women's abortion rights

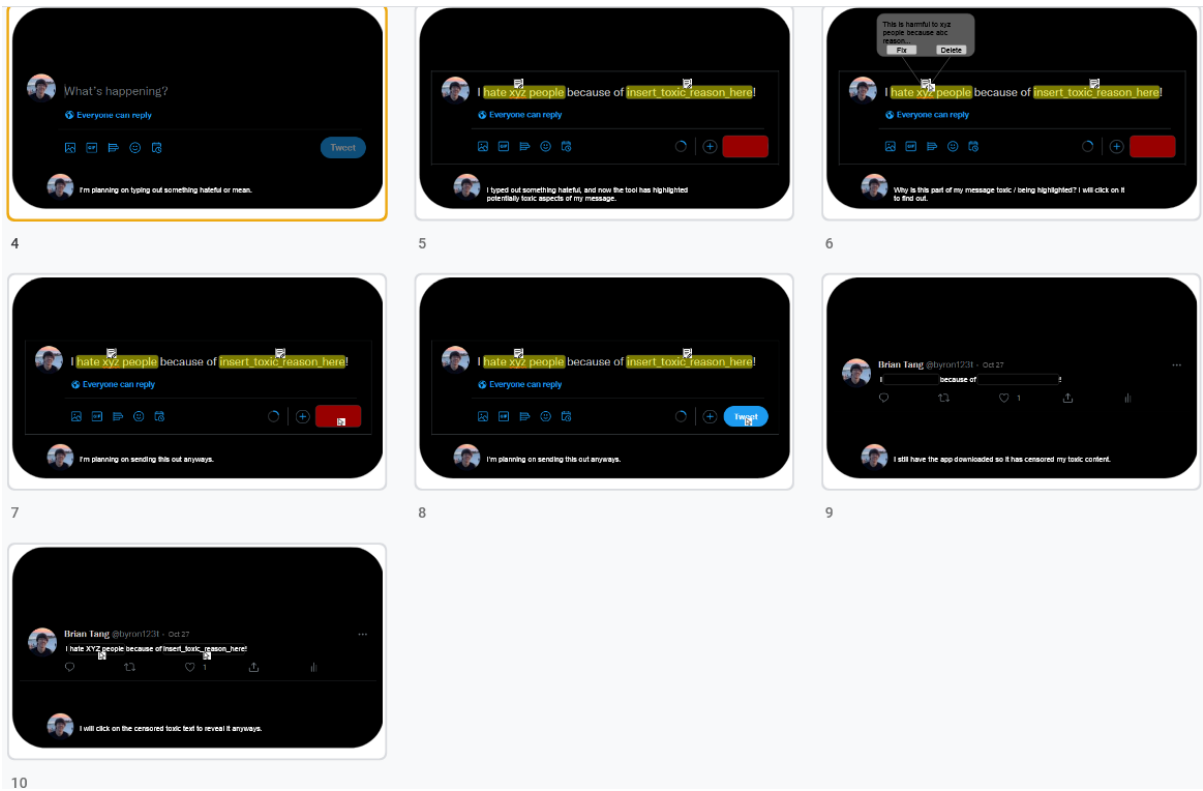


Fig. 29. Storyboard 2



Fig. 30. Storyboard 3

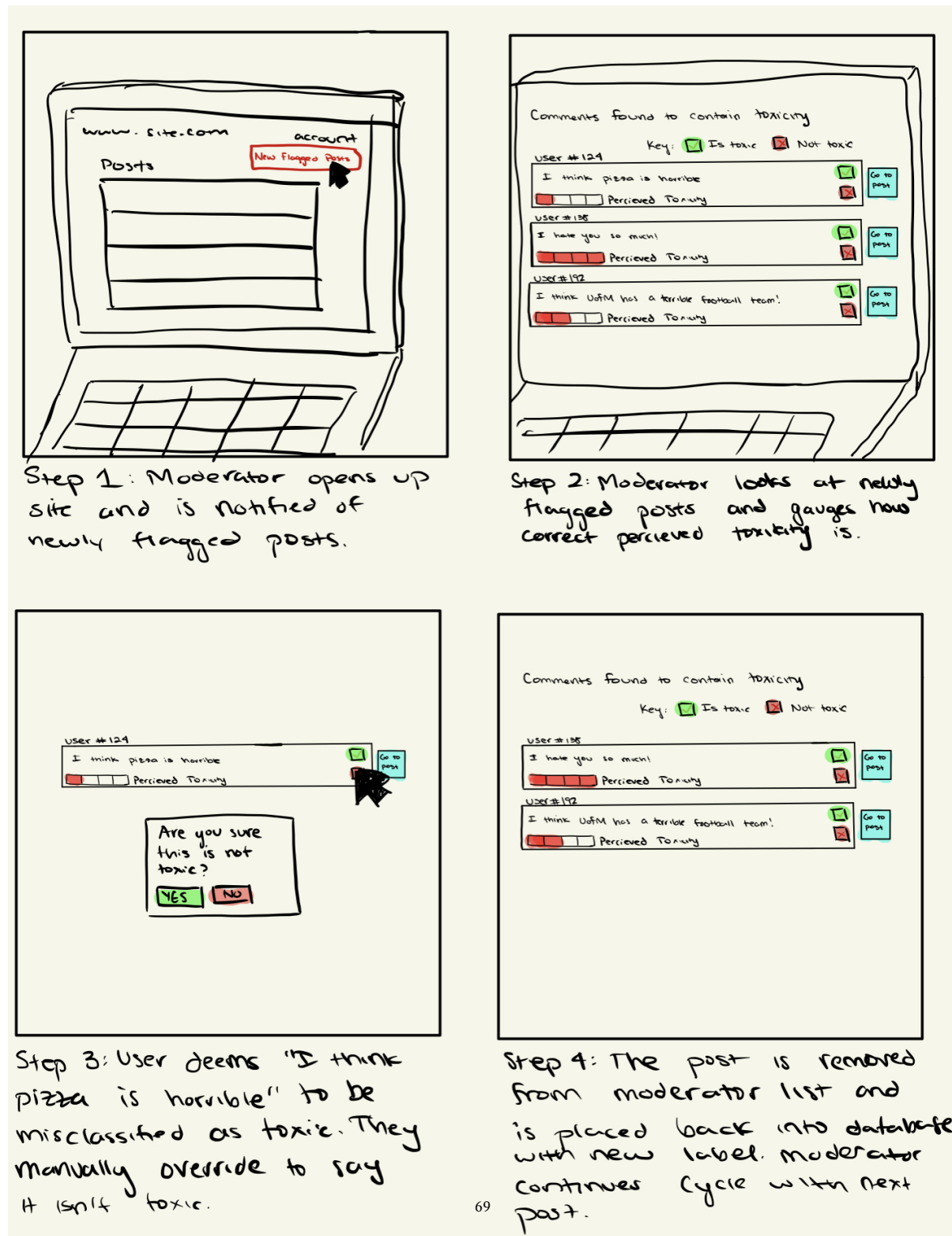


Fig. 31. Storyboard 4

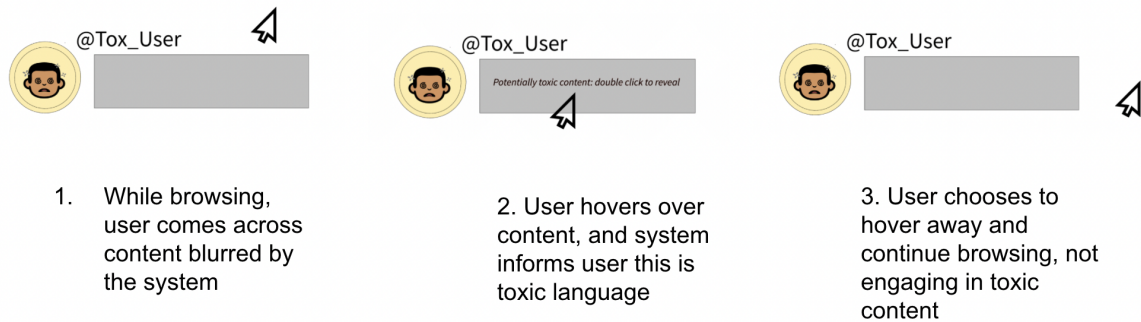


Fig. 32. Storyboard 5

D.4 Final Personas

The persona is a general user of a social media platform which frequently contains toxic language. This user both reads and writes content on the platform. This user at times will choose to read toxic content (individual persona 3) but will at other times avoid it to save themselves the emotional energy of processing it (individual persona 5). This user sometimes will also write toxic content and at times change the toxic content as a response to reading the toxicity detection system feedback (individual personas 1 and 2). The second persona would be the moderator of the system. They are a volunteer or paid worker of the platform and engages in moderating flagged content. The moderator personal determines if content should be considered toxic with assistance from the toxicity detection system.

D.5 Final Sketches

The final sketches can be found at <https://docs.google.com/presentation/d/1dpRZyaPxpCCJ2WfMibzgKmqEthemNspqeLBEM2Nf-U8/edit?usp=sharing> where each different action is on a different slide.

D.6 Final Storyboards

The final storyboards can be found at <https://docs.google.com/presentation/d/1BanGhATjsOudpVSjq6MCbiHpySU6UwHtdz2AzAJe0og/edit?usp=sharing> where each different action is on a different set of slides.

D.7 Final Paper Prototype

The final paper prototype slides can be found at https://docs.google.com/presentation/d/1585WCiW4UCQPFTI_d5ISpeSH07Dao4LNCgn-5dIcD60/edit?usp=sharing where each prototype for a particular action is on a different set of slides. Currently, they are in the order we expect them to be interacted with, but will have the final exact interaction set up during our user evaluation as specific by in our wizard-of-oz explanation.

E USABILITY EVALUATION

E.1 Individual Heuristic Evaluation Notes

- - Goal: avoid writing toxic context -> 1. Avoid reading toxic content 2. Avoid writing toxic content
- Plugin Settings page: mildly -> extremely toxic - blur / notify toggles sliders - explain why toggle
- blur appears overlaid on, e.g., Twitter - hover to get "show" btn (...) - show "hide" button to right
- - highlight orange when I write toxic content - click to get explanation pop-up - no explanation if toggle is off - can "apply" / delete from pop-up - how to choose what change(s) to suggest?
- 1 - where to go for option screen? - how to start the app? == 0 (missing context for presentation) - browser plugin
- 2 - no obvious problems
- 3 - cancel btn hidden by highlight click == 2 (see 4) -> could mark w/ branded icon
- 4 - hover vs. click == 2 (not a serious issue unless it's not clear what highlighting / blurring means)
- 5 - can users undo acceptance of bad suggested change? (Yes, if implemented as text edits) == 0
- 6 - no obvious issues
- 7 - accelerators: none would be useful - personalization: how do we know what is considered toxic? == 3 - customization: basics already included
- 8 - no obvious issues - impressively minimalistic
- 9 - aside from issues in 4, works as expected
- 10 - same as 7 (personalization) == 1 (but presentation of facts / tutorials)

E.2 Individual Simplified User Study Notes

• Participant 1:

Setting up tool

- (1) User felt that settings page was unclear, for example, the wording on the messages was not straightforward.
- (2) User tried to reference documentation, but was unable to find any help.
- (3) User indicated that rectangles were unclear and inverted (should be from left to right).
- (4) User indicated they weren't sure how much actual content would be censored based on the setting they selected.

Reading toxic content

- (1) User mentioned that they got frustrated with single clicking not working.
- (2) User suggested to show a small icon or button instead of double clicking.
- (3) User indicated there was no point in having a hide button.

Writing toxic content

- (1) User hovered over the highlighted text expecting something to happen based on their experience with the earlier tasks.
- (2) User mentioned there was no way to do nothing and that it seemed they had to either apply the suggested change or delete the content.
- (3) User suggested having an X mark in the corner of the window pop-up.
- (4) User indicated the desire to have enter, delete, and esc keys interact with the pop-up.

• Participant 2:

Setting up the tool

- (1) User is placed into settings after downloading the tool and asked to choose their preferences.
- (2) User didn't know how to initially interact with the interface when given the various options. They also weren't sure what they could interact with, specifically it was unclear if the toxicity bars could be changed.

- (3) After realizing they could interact with the toxicity bars, the user didn't understand what the bars were meant to do in the context of the content they would see.
- (4) The user felt that some of the terminology was ambiguous. For instance, if something is mildly toxic, should they expect to be able to see mildly toxic content, or is toxic content what is filtered out.

Reading toxic content

- (1) The user was placed onto the home page of twitter with a tweet on the screen.
- (2) The user was asked to reveal the toxic content. They proceeded to hover over the blurred content and immediately click, not recognizing that the action to interact with the content was just hovering.
- (3) Their initial click did not open the content given it requires a double click. Requiring a double click, or some form of restriction to access the content, was found to be helpful in case the user accidentally clicks on the box.
- (4) The user was able to hide the content after removing the blur, but was unsure exactly would happen when clicked, i.e. if it would reblur or hide the entire message.

Writing toxic content

- (1) The user is placed on a screen where they are writing something toxic.
- (2) The toxic portion of their text is highlighted, but the user unsure what to do next as they don't know how to determine why the text is highlighted.
- (3) They realize they can click on the highlight which produces a pop up. They see that the pop up provides an alternative, less toxic statement, as well as whether they want to apply or delete the content.
- (4) The user first tries the apply button which replaces their current text with a new set of text. They were frustrated that after application there was no undo button incase they didn't like the change.
- (5) When given the choice to delete, the user was confused about what the delete button would do. They were unsure if they were removing the highlight, or actually deleting the content.
- (6) The user wanted an option to go back to the settings and adjust the sensitivity of the toxicity detection, but didn't know how to get back to the setting menu.

• Participant 3:

Setting up the tool

- (1) User is facing the settings page after installing the tool.
- (2) User is not sure which parts of the interface can be interacted with and which parts are pre-set.
- (3) User is confused by the design of the toxicity bars. They are unsure if by moving the bar to towards the red will show them more extreme or less extreme content.
- (4) User believes the toxicity bars are quite subjective and is skeptical the level of toxicity implied is the same as what they desire to see or not see.

Reading toxic content

- (1) User is presented with Twitter page with the toxicity system blurring out text.
- (2) User is unsure what the blurred text represented initially.
- (3) User hovers over the text and clicks to reveal toxic text. User was frustrated the text would not reveal, and then double clicked.
- (4) User double clicks to hide text, admits double clicking takes getting used to.

Writing toxic content

- (1) User is unsure why text was highlighted, believes double clicking to reveal the text is too high of a barrier and hinders exploration.
- (2) It is not clear what the "apply" button does.
- (3) User is also confused if "delete" deletes the suggested text or the initial user-inputted text.

F USER EVALUATION

F.1 Apparatus Screenshots

Fig. 33 shows an example of the tool condition for deleting toxic content. Please refer to the Final Design and High-Fidelity prototype for complete demonstration of the tool.

For plain Twitter condition, participants are shown an original Twitter interface with toxic content pre-typed into the replying area (see fig. 34).

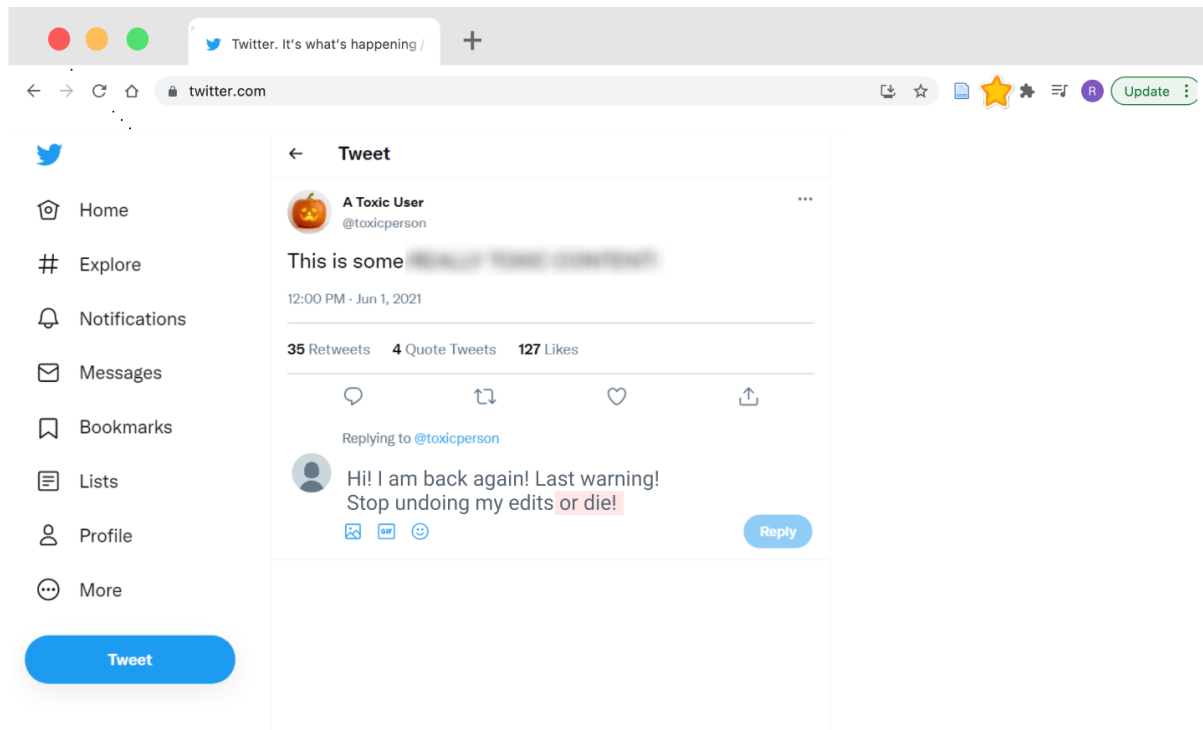


Fig. 33. Screenshot of the tool prototype condition for users.

F.2 Anonymized and De-identified Participants Data

G PRESENTATION VIDEO

<https://www.youtube.com/watch?v=kshStX8833M>

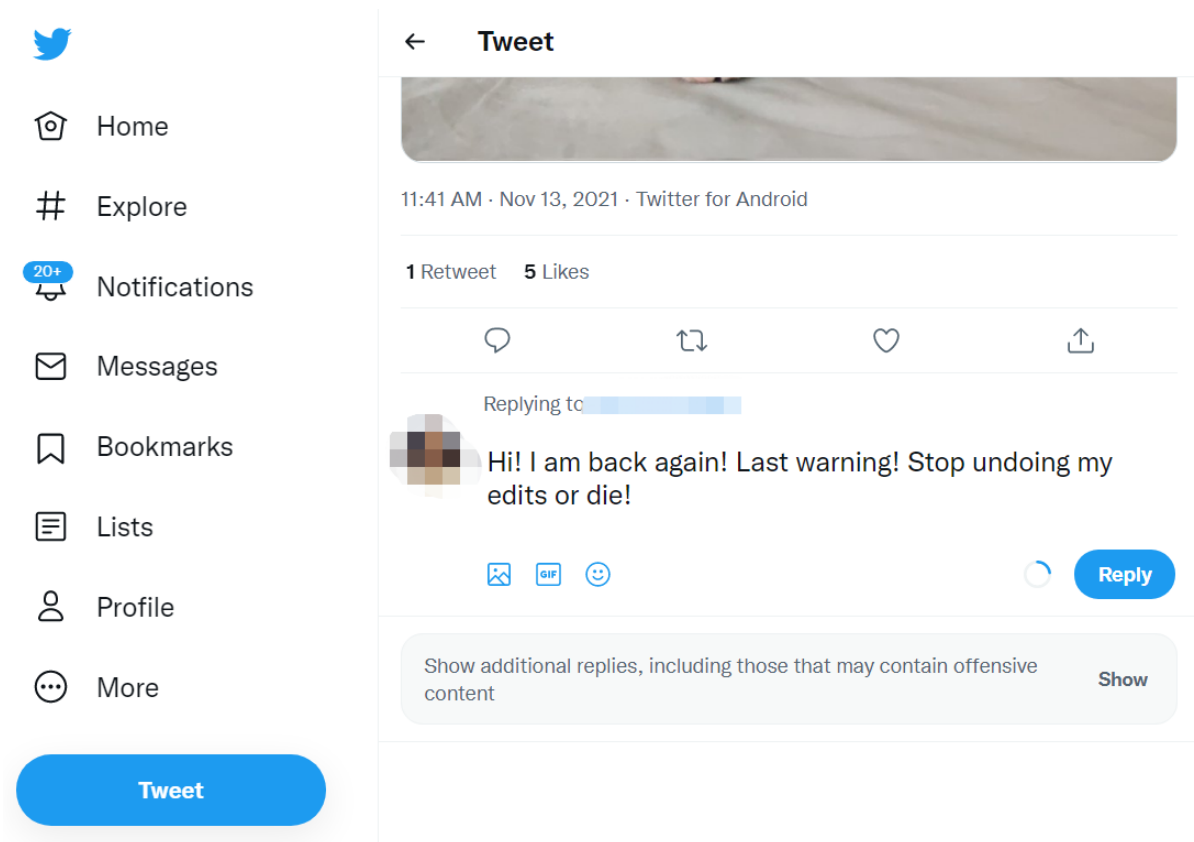


Fig. 34. Screenshot of the Twitter interface condition for users.

Tool	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10
deleting										
writing 1	7.93	2.53	11.22	1.55	9.75	9.54	4.93	2.89	14.2	5.87
writing 6	6.51	3.21	6.63	1.74	9.28	12.46	3.96	4.55	13.04	3.32
writing 2	4.58	1.23	15.76	2.03	5.95	8.12	3.36	3.6	10.2	6.06
fixing										
writing 3	9.88	2.17	5.84	1.15	9.07	9.1	5.06	3.98	12.4	8.16
writing 4	9.47	1.1	5.3	2.01	7.55	9.22	3.54	3.35	6.33	5.42
writing 5	5.04	1.9	4.16	2.51	8.55	8.24	4.63	3.21	9.32	5.36
Twitter										
deleting										
writing 1	10.76	7.25	3.28	6.37	16.48	13.5	5.54	4.84	9.65	8.9
writing 6	10.39	8.63	5.43	5.31	9.94	10	5.93	7.38	12.2	10.14
writing 2	5.87	6.13	8.7	3.02	16.2	8.29	7.62	6.81	11.8	8.71
fixing										
writing 3	21.45	8.52	9.63	11.46	14.6	9.17	9.13	8.64	10.5	12.4
writing 4	11.38	6.27	6.25	3.1	10	12.27	5.17	8.89	6.25	9.43
writing 5	17.69	7.33	9.61	6.1	14.8	9.25	5.85	7.98	8.2	9.67

Fig. 35. Anonymized and de-identified participants data for user evaluation, presenting the time taken for participants to complete the task in seconds.