Exploring Antecedents and Consequences of Toxicity in Online Discussions: A Case Study on Reddit

YAN XIA, Fudan University, China
HAIYI ZHU, Carnegie Mellon University, USA
TUN LU, Fudan University, China
PENG ZHANG[*], Fudan University, China
NING GU, Fudan University, China

Toxicity in online discussions has been an intriguing phenomenon and an important problem. In this paper, we seek to better understand toxicity dynamics in online discussions via a case study on Reddit that explores the antecedents and consequences of toxicity in text. We inspected two dimensions of toxicity: *language toxicity*, i.e. how toxic the text itself is; and *toxicity elicitation*, i.e. how much toxicity it elicits in its response. Through regression analyses on Reddit comments, we found that both author propensity and toxicity in discussion context were strong positive antecedents of language toxicity; meanwhile, language toxicity significantly increased the volume and user evaluation of the discussion in some sub-communities, while toxicity elicitation showed mixed effects. We then discuss how our results help understand and regulate toxicity in online discussions by interpreting the complicated triggers and outcomes of toxicity.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: Toxicity; online discussions; quantitative analysis; Reddit

ACM Reference Format:
Yan Xia, Haiyi Zhu, Tun Lu, Peng Zhang, and Ning Gu. 2020. Exploring Antecedents and Consequences of Toxicity in Online Discussions: A Case Study on Reddit. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 108 (October 2020), 23 pages. https://doi.org/10.1145/3415179

1 INTRODUCTION

Toxic behaviors, including harassment [22, 73], personal attack [76], hate speech [26, 67], etc., have been prevalent in online discussions, and have had detrimental impacts on the experience of users. Most previous studies viewed toxic behaviors as anomalous, ascribed them to certain anomalous individuals [5, 11, 61, 69], and attempted to eradicate such content and users by proposing detection [14, 19, 27, 55, 59, 74, 77, 78] and regulation [46, 49, 58] strategies.

Recent studies suggested that toxic behaviors are much more complicated. For example, a study on trolling [13] showed that situational factors may cause ordinary people to perform toxic behavior.

[*] Corresponding author.

Authors' addresses: Yan Xia, kellyxia26@gmail.com, Fudan University, Shanghai, China; Haiyi Zhu, haiyiz@cs.cmu.edu, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA; Tun Lu, lutun@fudan.edu.cn, Fudan University, Shanghai, China; Peng Zhang, pzhang14@fudan.edu.cn, Fudan University, Shanghai, China; Ning Gu, ninggu@fudan.edu.cn, Fudan University, Shanghai, China.

2573-0142/2020/10-ART108 $15.00
https://doi.org/10.1145/3415179

Table 1. Examples of toxicity-eliciting text.

He didn’t have to appoint him and that’s the point. So, now that you’re aware of how that works, please don’t spread this oversimplification that misinforms people anymore.	**No, I just won the argument. Relationships is built around a business contract, not love. Or else women wouldn’t initiate between 70 to 80 percent of divorce.**
- He literally did. It’s *Statute* you numbskull. - Holy s**t you’re dense, aren’t you? - People are spelling it out for you, and you’re just arguing past them. You *cannot* be this dense, and still be arguing in good faith.	- [removed] - Well, I can guarantee 100% that you just pulled that entire sentence and "facts" out of your a**.

Yet we lack a holistic understanding of the dynamics of toxic behaviors, such as what factors can predict toxic behaviors, and what consequences toxic behaviors might lead to.

In this paper, we contribute to the understanding of toxic behaviors by quantitatively exploring a wide range of antecedents and consequences of toxic text in online discussions. We inspected two dimensions of toxicity: *language toxicity* and *toxicity elicitation*. Language toxicity is defined as how much the text itself contains content that is rude, disrespectful or unreasonable, which has been studied extensively in natural language processing (NLP) literature [19, 27, 55, 59, 74, 77, 78]. By contrast, text with high toxicity elicitation is what elicits high toxicity in its response, but not necessarily contains toxic language itself: Table 1 shows two examples of such text pieces, along with their replies[1]. Interestingly, many of those text pieces are usually able to get away with toxicity regulation and remain in the discussion as a source of toxicity. Therefore, we also considered toxicity elicitation – the extent to which a piece of text elicits toxicity in its response – as another aspect of toxicity in our analysis.

We conducted quantitative analysis on Reddit[2] comments, after collecting and quantifying the features of each comment with the help of NLP tools, including the *Perspective API*[3] for quantifying the language toxicity of text pieces. Specifically, we conducted regression analyses to find out correlations among the extracted features of comment-reply pairs as well as the structural features of comment trees, in order to answer: i) what factors can predict the language toxicity and toxicity elicitation of a comment, and ii) how can the language toxicity and toxicity elicitation of a comment influence the subsequent discussion. It is worth mentioning that we modeled toxicity as a continuous value, as we actually perceive various intensity of toxicity in real comments. The combination of continuous modeling and regression analysis revealed how variation in toxicity intensity correlated with variation in other factors, leading to a more nuanced understanding of toxicity dynamics.

Considering the disparate discussion styles and community norms among different subreddits of Reddit, we conducted our analyses separately on five of the most popular subreddits: r/announcements, r/worldnews, r/politics, r/todayilearned, and r/AskReddit. Our results show that: i) author propensity was a strong positive antecedent of language toxicity in all subreddits; ii) controlling for that effect, toxicity in discussion context was a strong positive antecedent of both language toxicity and toxicity elicitation in all subreddits; iii) language toxicity significantly increased the volume and user evaluation of discussion in some of the subreddits; iv) the effect of toxicity elicitation varied by subreddit.

[1] "[removed]" indicates a reply removed by administrators.

[2] https://www.reddit.com/

[3] https://perspectiveapi.com/#/

We further discuss with qualitative examples how our results help understand toxicity dynamics in online discussions. Specifically, we interpret the complicated triggers and outcomes of toxicity, based on which we propose design implications for toxicity regulation, including mitigating the influence of contextual toxicity, detecting and regulating toxicity-eliciting comments, distinguishing different types of toxicity and tailoring regulation measures accordingly, etc. With quantitative analysis and complementing qualitative examples, our study reveals a more fine-grained and comprehensive picture of toxicity dynamics, and also provides insights for understanding and regulating toxicity in online discussions.

2 RELATED WORK AND HYPOTHESES

2.1 From Inappropriate Behavior to Toxicity

Various inappropriate behaviors were observed and studied in online discussions, yet the "inappropriate" feature of such behavior was described in disparate ways as antisocial [14], uncivil [16, 39], flaming [41, 51, 56], abusive [55, 59], offensive [19, 77] or malicious [52], hardly endowed a widely recognized name or definition. In this work, we focused on a certain type of inappropriateness – *toxicity*. Following the Perspective API, we defined toxicity, specifically *language toxicity*, as the "*(explicit) rudeness, disrespect or unreasonableness of a comment that is likely to make one leave the discussion*", resonating with the concepts of toxic disinhibition [50, 70] and toxic behavior [47, 48] in previous studies.

2.2 Toxicity Dynamics in Online Discussions

We synthesized prior literature that explored toxicity-related factors, and formed hypotheses in cases where prior literature suggested specific directions of the effects. When prior work did not provide specific directions, we formed open research questions.

2.2.1 Antecedents of Language Toxicity.

- *Author's propensity toward toxicity.* For a certain comment, the propensity of its author should largely influence the toxicity within. With respect to trolling, an act closely related to toxicity, there were a number of studies suggesting that trolling should be mainly ascribed to a small number of individuals who are predisposed to start trolling by nature [5, 11, 61]. Although a recent study implied that anyone can become a troll under certain circumstances [13], it is almost beyond doubt that the toxicity of a comment would have strong positive correlation with the author's propensity toward toxic posting.
  Hypothesis 1: *Author's propensity toward toxicity will increase the language toxicity of his/her text.*
- *Author's experience in community.* Since newcomers of a community may often act non-normatively due to lack of commitment or ignorance of community norms [45], we inferred that users who are less engaged in the community tend to demonstrate less conformity to the regulations and rules of the community, which normally disapprove of toxicity [62].
  Hypothesis 2: *Author's experience in the community will reduce the language toxicity of his/her text.*
- *Toxicity in context.* We expected toxicity in discussion context to be a positive contributing factor of subsequent toxicity. A rich body of literature showed that all kinds of emotions [6, 17, 23, 29, 30, 43, 44] and behaviors [12, 13, 24] are contagious in online communities, while revenge and conflict escalation is prevalent in offline settings [3, 8, 66]. Studies on incivility also showed that uncivil comments in context lead to hostile cognition [65] and communicate tolerance of incivility [64], likely triggering more incivility in response. Also, among user-reported reasons of posting malicious comments, "supporting other malicious

comments" along with "following the context norm" were on the list, while two of the most important reasons "resolving a feeling of inferiority or frustration" and "hostility toward others" might also suggest exposure to other malicious comments [52].
Hypothesis 3: *Toxicity in discussion context will increase the language toxicity of text.*

- *Polarity in context.* We also expected the polarity of text in discussion context to be a positive antecedent of toxicity, where polarity refers to the semantic orientation of text on a scale from negative to neutral to positive, as in NLP literature and tools [54, 57], and higher polarity of text corresponds to stronger opinion (either positive or negative) within. Looking into the emotion theory, we found that a number of researchers held the view that emotions are not merely internal feeling states, but also social relationships [21], or affective responses to the environment [25]. Specifically, anger is expressed to repulse or restrain others, or to correct others' behavior [25]. We thus expected higher polarity (i.e. stronger opinion) to be eliciting more anger, into more toxicity externalized.
Hypothesis 4: *Polarity in discussion context will increase the language toxicity of text.*

2.2.2 *Consequences of Language Toxicity.*

- *Volume of discussion.* Negative emotions were shown to boost user response on both BBC discussion forum [15] and Twitter [42], and more specifically, incivility was shown to increase polarization [1, 2, 39] and facilitate interaction [10]. Nonetheless, studies also showed that most people do not respond to explicit aggression [81], and toxic behaviors such as harassment tend to decrease user participation and retention [72]. Therefore, we wished to explore how toxicity actually correlates with the volume of discussion.
Question 1: *How will the language toxicity of text influence the volume of discussion?*
- *Evaluation of discussion.* On the one hand, prior literature suggested that toxicity might devalue a discussion. Assuming that the toxicity in a comment embodies the emotion of anger, which simplifies cognitive processing [53], leading people toward impulsive, ill-considered, stereotypical judgement and action [9], we expected comments of more toxicity to involve less rational thinking, thus be of less intrinsic value. Studies also showed that personal conflicts impede work quality on Wikipedia [4], and incivility in comments hurts the perceived quality of an article [60]. On the other hand, the evaluation of a discussion in a discussion community is often operationalized by other users' response to it – how much attention they pay, how much they empathize, and how much they think it is worth further discussing – and it is very possible that comments of higher toxicity may attract higher level of user attention, and obtain more recognition in the end. We therefore wished to explore how toxicity actually correlates with the evaluation of discussion.
Question 2: *How will the language toxicity of text influence the evaluation of discussion?*

2.2.3 Toxicity Elicitation. Apart from the comments detected of high toxicity, we were aware that a large number of innocuous-looking comments also contributed to the thriving of toxicity by eliciting highly toxic responses. Thus, we considered a novel feature of *toxicity elicitation* – the extent to which a text piece elicits toxicity in its replies. Although such concept was not elaborated in literature, we found studies that implied the existence of toxicity elicitation: for example, posts that involve covert strategies of trolling [31, 32] may not display explicit toxicity, yet will very likely elicit toxic responses; and ostensibly civil conversations showing certain warning signs might derail into toxic ones [79]. Therefore, we also intended to study how this new feature correlates with prior and posterior factors of a discussion.

Question 3: *How will the antecedents and consequences of toxicity elicitation differ from those of language toxicity?*

Table 2. Introduction of subreddits.

Subreddit	#Subscribers	#Comments (2017.12)	Introduction
r/announcements	37.4m	19,346	Where administrators post about latest features of Reddit and users discuss in comments.
r/worldnews	20.9m	765,709	Where users discuss major news from around the world, excluding US-internal news.
r/politics	4.9m	1,753,120	Where users discuss US politics.
r/todayilearned	20.6m	472,650	Where users share what they learn today.
r/AskReddit	22.2m	5,245,881	Where users ask and answer thought-provoking questions.

3 METHOD

3.1 Dataset

We chose Reddit as our platform of study. As a popular news sharing and discussion website, Reddit was ranked as the #3 most visited website in the U.S. and #6 in the world by Alexa Internet[4]. Users of Reddit are allowed to share links, create posts and leave comments, all of which then get to be voted up or down by others, and the authors earn *post karma* or *comment karma* when their posts or comments get upvoted. Items (links or posts) on Reddit are organized into subreddits, each dedicated to a specific topic.

Our reasons for choosing Reddit include: i) it is an online discussion community with great popularity and offers a huge amount of online discussion data, ii) it provides a user voted score for every comment as an explicit measure of user evaluation, and iii) comments on Reddit follow a clear nested parent-child (comment-reply) structure with presumed influence flows from parent comments to child comments, making the data appropriate for our analysis.

Considering the various discussion styles and community norms among different subreddits, we planned to conduct our study on separate subreddits, and explore how results hold or vary. We chose five of the most popular text-based discussion communities for study: r/announcements, r/worldnews, r/politics, r/todayilearned, and r/AskReddit[5]. A brief introduction of these subreddits is given in Table 2. The subreddits were chosen based on an overall consideration of the number of subscribers (i.e. the size of the potentially affected audience of the discussion), the number of comments (i.e. the amount of discussion), as well as the style of discussion in each subreddit. For example, r/announcements was selected despite the limited amount of discussion within, considering its large number of subscribers and the unique discussion style it offers, where administrators are actively involved. Meanwhile, r/politics was selected for the large amount of discussion that happened within, despite its limited number of subscribers.

We acquired data from the public Reddit dataset on pushshift.io [7]. We first selected respectively 100,000 comments posted in December 2017 from the five subreddits, except for r/announcements that only had 19,346 comments that month. Then we augmented the dataset by collecting all later comments posted between December 2017 and June 2018[6] that belong to the same posts, to obtain complete comment trees. Our augmented dataset contained 19,682/152,632/200,635/139,353/265,893 comments for each subreddit. Using the Reddit API[7], we then collected features of each user in the dataset, including his/her account creation time, post/comment karma, etc.

[4] https://www.alexa.com/

[5] All subreddit statistics discussed in the rest of this paper will be presented in this order if not specified.

[6] All reddit posts are archived (unable to comment) after 6 months.

[7] https://www.reddit.com/dev/api/

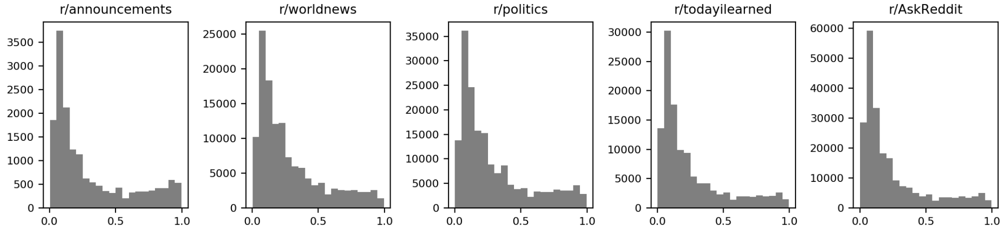

Fig. 1. Toxicity distributions of valid comments in the five subreddits.

3.2 NLP Tools: Validating the Perspective API

We used existing NLP tools to generate the linguistic features of each comment, including the Perspective API for toxicity, and the TextBlob [54] library for polarity.

As an automatic toxicity detection tool, the Perspective API models the toxicity of each input text as a continuous score from 0 to 1 (1 being most toxic). The API was trained with millions of comments rated by human workers on a scale from "very healthy" to "very toxic" [38] on a 5-point Likert scale, and the output toxicity score of the API reflects the intensity of toxicity in the input text, as shown in the interactive examples on its webpage. Similarly, the TextBlob library generates a polarity score from -1 to 1 for each piece of input text, with -1 being most negative and 1 being most positive.

The TextBlob Python library had been widely used as a sentiment analysis tool in various applications and studies [28, 33, 34, 80]. While the Perspective API had also been extensively applied in fields including machine-aided content moderation, it had been shown to be vulnerable to adversarial examples [38] (i.e. it is possible to subtly modify a piece of highly toxic text to get a significantly lower toxicity score from the API), and thus it was not clear how reliable the Perspective-generated scores would be if directly used as a measurement of text toxicity. Therefore, we conducted a validation study to explore how well Perspective-generated toxicity scores aligned with human perception on our dataset.

We randomly sampled 500 comments from our dataset (100 comments per subreddit), and for each comment asked 3 workers on Figure Eight[8] to rate the toxicity of it, on a scale of 1 to 5 (5 being most toxic). Since the discrepancy among workers seemed acceptable (Krippendorff's α=.54, average rating standard deviation=.53), we took the average of the 3 workers' ratings as the final human-labeled toxicity score of each comment. The final human-labeled toxicity scores showed high positive correlation with Perspective-generated scores[9] (Pearson's r=.82, p<.001).

3.3 Data Preprocessing

We first generated the tree structure of the comments using a top-down approach, excluding comments whose root nodes did not exist in the dataset. Then we used the Python library *langdetect* to detect the language of each comment, and used NLP tools to generate the linguistic features of those written in English. During the process, we marked invalid comments including deleted comments (comments deleted by the author), removed comments (comments removed by administrators), and comments not detected as English. Statistics of the preprocessed dataset are listed in Table 3, and the toxicity distributions of valid comments in the five subreddits are shown in Figure 1.

[8] https://www.figure-eight.com/

[9] Note that we used the default TOXICITY model of the Perspective API. We also experimented with the SEVERE_TOXICITY model, yet the TOXICITY scores better matched the human-labeled scores on our sampled dataset, as shown in Appendix A.

Table 3. Dataset statistics.

	r/annc	**r/wn**	**r/pol**	**r/til**	**r/ar**
#all comments	19,682	152,632	200,635	139,353	265,893
#comments not in tree	301	6,277	7,171	4,658	39,186
#invalid comments	3,282	23,338	27,307	19,255	41,334
#removed comments	367	8,024	7,197	2,588	3,217
#deleted comments	1,245	8,286	9,494	7,051	12,306
#users	7,989	44,012	40,211	53,709	96,876
Avg. #comments per user	2.03	2.90	4.26	2.22	2.30
Avg. #children per comment	0.73	0.83	0.78	0.77	0.58
Avg. #descendants per comment	2.84	3.71	2.39	3.23	1.67

3.4 Study I: Antecedents of Toxicity

We designed our first study to explore the antecedents of toxicity – respectively, language toxicity and toxicity elicitation – in discussions. Specifically, we conducted regression analysis to find out correlations between the language toxicity/toxicity elicitation of a target comment and features of its parent comment. We used the direct parent comment to represent the precedent discussion because i) it should have had the most influence on the target comment among others, as the target comment author chose to reply directly to it, and ii) by doing so we avoided adding up cascaded effects along the nested parent-child structure. Corresponding to our review of related works, we designed the study as follows.

3.4.1 Variables. In the variable names below, a prefix of *Target* indicates a feature of the target comment, and *Parent* indicates a feature of the parent comment. We include the details of how we operationalized the variables in Table 4.

- Independent variables.
  We included one variable for author's propensity toward toxicity: *Target-author-toxicity*;
  Two variables for author's experience in community: *Target-author-age*, and *Target-author-karma*;
  One variable for toxicity in context: *Parent-language-toxicity*;
  Two variables for polarity in context: *Parent-positivity*, and *Parent-negativity*.
- Dependent variables.
  Target-language-toxicity, and *Target-toxicity-elicitation.*
- Control variables.
  Parent-length, *Parent-%-caps*, and *Parent-%-punctuation.* Also, we had *Target-language-toxicity* as a control variable when having *Target-toxicity-elicitation* as the dependent variable.

The platform-specific variables were chosen in consideration of both how they were used in related works and whether they were available in our collected data. Specifically, we measured the user's experience in the community by his/her age on Reddit as in [36], as well as the user's comment karma, which reflects his/her contribution to the Reddit community via posting comments [63]. The control variables were first selected in reference to previous studies that explored the linguistic features of Reddit comments [18, 20, 37, 71], and then some of them (e.g. politeness, readability) were excluded after correlation and multicollinearity tests to assure low correlation (-.3<Pearson's r<.3) and multicollinearity (VIF<5) between variables.

3.4.2 Method. We used linear regression because most of our independent and dependent variables were continuous. Separate models were trained for all dependent variables. All variables were standardized for easier interpretation of the regression coefficients.

Table 4. Variable descriptions.

	Toxicity Features
-language-toxicity	The toxicity score of the comment, generated using the Perspective API.
-toxicity-elicitation	The toxicity elicitation of the comment, calculated as the toxicity score of the chronologically first child comment (direct reply) of the comment. *(Note that we were not calculating the toxicity elicitation of a comment as the average toxicity score of all its child comments, to avoid cascading effects among the child comments from earlier ones to later ones – yet theoretically the concept should be referring to the overall toxicity level in the child comments.)*
	Polarity Features
-positivity	The positivity score of the comment, calculated as $\max(0, pol)$, pol being the polarity score of the comment, generated using TextBlob. *(We split polarity into two different variables to model positivity and negativity on independent axes, so that the strength of the opinion would be separated from the stance of it.)*
-negativity	The negativity score of the comment, calculated as $\max(0, -pol)$, pol being the polarity score of the comment, generated using TextBlob.
	Other Linguistic Features
-length	The length (number of characters) of the comment.
-%-caps	The percentage of capitalized characters in the comment. *(We used percentage instead of number to avoid strong correlation between linguistic variables.)*
-%-punctuation	The percentage of punctuation marks in the comment.
	Author's Propensity
-author-toxicity	The average toxicity score of the comment author's history comments in the dataset. *(We excluded comments whose authors had no history comment in the dataset.)*
	Author's Experience
-author-age	The comment author's age on Reddit, i.e. the time elapsed between creating his/her Reddit account and posting the comment.
-author-karma	The comment author's comment karma on Reddit, i.e. his/her contribution to Reddit by posting comments.
	Volume of Discussion
-#children	The number of child comments (direct replies) of the comment.
-#descendants	The number of descendants of the comment.
-height	The distance from the comment to the deepest leaf of the comment sub-tree (i.e. its descendant with the greatest depth), which equals the levels of descendants of the comment.
	Evaluation of Discussion
-score	The user-voted score of the comment, calculated as the number of user upvotes minus the number of user downvotes.
-score-max	The maximum user-voted score of the comments.

3.4.3 Preprocessing. For Study I, we had to filter out all invalid comments, for which toxicity scores could not be generated or author information was missing. We acknowledge the limitation of doing so, because invalid comments (especially removed ones) might have embodied different toxicity dynamics from the valid ones. However, in this work we value the retained comments above the removed ones, since we believe the content of those retained comments should have had larger and longer-term effect on the community, and thus should be of higher value to both the community and our research.

To exclude the possible effect of earlier sibling comments, we conducted the regression only between a parent comment and its first child comment (as the target comment). Besides, to ensure consistency within our two studies, we took only comments from the *first* level in tree (i.e. comments replying directly to Reddit posts) as the parent comments, and comments from the *second* level as target comments. We also excluded comments with no author posting history. After preprocessing, we obtained 242/1,654/3,715/1,590/2,391 parent-target comment pairs for Study I.

3.5 Study II: Consequences of Toxicity

We designed our second study to explore the consequences of toxicity – respectively, language toxicity and toxicity elicitation – in discussions. Specifically, we conducted regression analysis to discover correlations between the language toxicity/toxicity elicitation of a target comment and the volume or user evaluation of the discussion it leads. We designed the study as follows.

3.5.1 Variables. In the variable names below, a prefix of *Target* indicates a feature of the target comment, and *Children* indicates a feature of the child comments. The details of how we operationalized the variables are shown in Table 4.

- Independent variables.
  Target-language-toxicity, and *Target-toxicity-elicitation.*
- Dependent variables.
  We included three variables for discussion volume: *Target-#children*, *Target-#descendants*, and *Target-height*;
  Two variables for discussion evaluation: *Target-score*, and *Children-score-max*.
- Control variables.
  Target-positivity, *Target-negativity*, *Target-length*, *Target-%-caps*, and *Target-%-punctuation.*

Considering the hierarchical thread structure on Reddit [75], we selected the number of direct replies, the number of total descendants and the height of the comment tree to be the platform-specific metrics of discussion volume. Besides, we measured the user evaluation of a comment by the user-voted score[10] of it, as in previous Reddit studies [20, 36, 40].

3.5.2 Method. We conducted negative binomial regression for all dependent variables because all of them were over-dispersed count variables. Separate models were trained for all dependent variables. All independent variables and control variables were standardized for easier interpretation of the regression coefficients.

3.5.3 Preprocessing. For Study II we had to exclude most invalid comments when calculating language toxicity and toxicity elicitation, yet we included removed comments among child comments and assumed they had a toxicity score of 1 (most toxic), since comments removed by administrators usually demonstrate high toxicity. Besides, all invalid comments still remained in the tree structure and contributed to features including #children, #descendants, height, etc.

To ensure consistency within the two studies, we only used comments from the *second* level in tree as target comments, and comments from the *third* level as child comments. After preprocessing, we obtained 427/2,981/5,597/3,247/4,758 target-children comment pairs for Study II.

4 RESULTS

4.1 Study I: Antecedents of Toxicity

The results of Study I are shown in Table 5, and the main findings are summarized in Table 6. The regression coefficients of the control variables are shown in the additional results in Appendix B.

We found that as expected in H1, target comment author's propensity toward toxicity, measured by the average toxicity of his/her previous comments, showed very strong positive correlation with

[10] There has been a "vote fuzzing" mechanism on Reddit that fuzzes the actual number of votes to prevent vote manipulation. However, it has been claimed that after July 2015, the voting data provided (including the net vote score we used) "can actually be trusted" – see https://www.reddit.com/r/modnews/comments/3cglvp/introducing_rmodsupport_semiama_with_me_the/csvkc56/.

Table 5. Regression results of Study I: Antecedents of Toxicity. ***: p<.001, **: p<.01, *: p<.05. A positive/negative regression coefficient indicates a positive/negative correlation between the independent variable and the dependent variable, and a smaller p-value indicates a higher significance of the correlation. The absolute value of a coefficient indicates the effect size of the correlation.

		Dependent Variable: *Target-language-toxicity*				
Factor (Expected Correlation)	Independent Variable	Regression Coefficient in r/annc	 r/wn	 r/pol	 r/til	 r/ar
Author's propensity (+)	*Target-author-toxicity*	.207 **	.133 ***	.125 ***	.103 ***	.160 ***
Author's experience (-)	*Target-author-age*	-.029	-.055 *	-.029	-.035	.024
	Target-author-karma	-.031	.008	.004	.030	-.003
Toxicity in context (+)	*Parent-language-toxicity*	.210 **	.182 ***	.156 ***	.184 ***	.246 ***
Polarity in context (+)	*Parent-positivity*	.135 *	-.000	.005	.000	-.018
	Parent-negativity	.017	-.038	-.020	-.027	-.014
		Dependent Variable: *Target-toxicity-elicitation*				
Factor (Expected Correlation)	Independent Variable	Regression Coefficient in r/annc	 r/wn	 r/pol	 r/til	 r/ar
Author's propensity (?)	*Target-author-toxicity*	.011	.068 **	.024	.033	.041 *
Author's experience (?)	*Target-author-age*	.013	.010	.012	.005	.013
	Target-author-karma	-.032	.039	-.008	.006	-.019
Toxicity in context (?)	*Parent-language-toxicity*	.173 *	.118 ***	.082 ***	.151 ***	.136 ***
Polarity in context (?)	*Parent-positivity*	.052	.029	-.021	-.031	-.016
	Parent-negativity	.049	.025	.007	-.037	.007
	Control Variable					
/	*Target-language-toxicity*	.260 ***	.113 ***	.115 ***	.168 ***	.142 ***

Table 6. Main findings of Study I: Antecedents of Toxicity.

H1:	*Author's propensity toward toxicity will increase the language toxicity of his/her text.*
Result:	Author's propensity toward toxicity significantly increased the language toxicity of a comment in all subreddits. **(H1 is supported.)**
H2:	*Author's experience in the community will reduce the language toxicity of his/her text.*
Result:	Author's experience in the community was not significantly correlated with language toxicity in most subreddits. **(H2 is not supported.)**
H3:	*Toxicity in discussion context will increase the language toxicity of text.*
Result:	Toxicity in discussion context significantly increased the language toxicity of a comment in all subreddits. **(H3 is supported.)**
H4:	*Polarity in discussion context will increase the language toxicity of text.*
Result:	Polarity in discussion context was not significantly correlated with language toxicity in most subreddits. **(H4 is not supported.)**
Q3:	*How will the antecedents (and consequences) of toxicity elicitation differ from those of language toxicity?*
Result:	Only toxicity in discussion context significantly increased the toxicity elicitation of a comment in all subreddits.

the language toxicity of the target comment in all subreddits (Coef.>=.103, p<.01). This means that the increase of one standard deviation in the average toxicity of the author's previous comments led to the increase of over .103 standard deviations in the language toxicity of the current comment. Thus author propensity was shown to be a strong positive antecedent of language toxicity.

Controlling for that effect, strong positive correlation was also shown between the language toxicity of the target comment and that of its parent comment in all subreddits (Coef.>=.156, p<.01), as expected in H3. That is to say, the increase of one standard deviation in the language toxicity of the parent comment led to the increase of over .156 standard deviations in the language toxicity of the target comment. Thus, toxicity in discussion context was shown to be another strong positive antecedent of language toxicity.

Meanwhile, target comment author's experience, measured by the author's age in the community and his/her comment karma (i.e. contribution by posting comments), did not show significant negative correlation with language toxicity (H2) in most subreddits. The only exception that weakly supported the hypothesis was that in r/worldnews, target comment author's age showed negative correlation with the language toxicity of the target comment (p<.05). Similarly, polarity in the discussion context also did not show significant positive correlation with language toxicity (H4) in most subreddits, the only exception that weakly supported the hypothesis being that in r/announcements, the positivity of the parent comment showed positive correlation with the language toxicity of the target comment (p<.05). However in both cases, there was no coherent correlation observed throughout all metrics and subreddits, and thus no strong conclusion could be drawn from the results.

On the other hand, toxicity in discussion context seemed to be the only strong positive antecedent of toxicity elicitation. Among all the independent variables, only the language toxicity of the parent comment showed significant positive correlation with the toxicity elicitation of the target comment in all subreddits (Coef.>=.082, p<.05), controlling for the language toxicity of the target comment itself. More intuitively, with the language toxicity of the target comment fixed, the increase of one standard deviation in the language toxicity of the parent comment led to the increase of over .082 standard deviations in the toxicity elicitation of the target comment.

In summary, our results support the hypotheses that both the author's propensity toward toxicity (H1) and the level of toxicity in the discussion context (H3) will increase the language toxicity of text. However, the hypothesized negative correlation between language toxicity and the author's experience in community (H2) and the hypothesized positive correlation between language toxicity and the level of polarity in discussion context (H4) are not supported in general. Besides, we found that toxicity in discussion context also increased the toxicity elicitation of text (Q3).

We here propose potential explanations for the unexpected results. As for the unsupported negative correlation between language toxicity and author's experience in community, it is possible that as some users got more involved in and recognized by the community, they were equipped with more confidence to freely express their opinions and emotions, whereas newcomers tended to be more reserved. And regarding the unsupported positive correlation between language toxicity and polarity in discussion context, it is possible that online language patterns had evolved toward covert aggressiveness, and many of those seemingly neutral comments actually incorporated irritatingly strong opinions. Future work is needed to test these speculations and identify the underlying causes of the findings.

4.2 Study II: Consequences of Toxicity

The results of Study II are shown in Table 7, and the main findings are summarized in Table 8. The regression coefficients of the control variables are shown in the additional results in Appendix B.

Table 7. Regression results of Study II: Consequences of Toxicity. ***: p<.001, **: p<.01, *: p<.05.

Factor (Expected Correlation): Volume of Discussion (?)						
Dependent Variable	Independent Variable	Regression Coefficient in r/annc	r/wn	r/pol	r/til	r/ar
Target-#children	*Target-language-toxicity*	.055	.065 **	.056 **	.068 **	.052 **
	Target-toxicity-elicitation	-.128	-.027	-.008	-.009	.019
Target-#descendants	*Target-language-toxicity*	.005	.083 ***	.113 ***	.107 ***	.080 ***
	Target-toxicity-elicitation	-.144 **	-.055 **	.012	.021	.032 *
Target-height	*Target-language-toxicity*	-.040	.022	.026	.049 *	.009
	Target-toxicity-elicitation	.016	-.024	.001	.048 *	.044 *

Factor (Expected Correlation): Evaluation of Discussion (?)						
Dependent Variable	Independent Variable	Regression Coefficient in r/annc	r/wn	r/pol	r/til	r/ar
Target-score	*Target-language-toxicity*	.019	.011	.023	.025	.034 *
	Target-toxicity-elicitation	-.050	-.018	-.002	-.016	.014
Children-score-max	*Target-language-toxicity*	.015	.018	.028 *	.012	.049 **
	Target-toxicity-elicitation	-.029	-.021	.009	-.025	.062 ***

Table 8. Main findings of Study II: Consequences of Toxicity.

Q1:	*How will the language toxicity of text influence the volume of discussion?*
Result:	The language toxicity of a comment significantly increased the number of its replies in 4/5 subreddits.
Q2:	*How will the language toxicity of text influence the evaluation of discussion?*
Result:	In a few of the subreddits, the language toxicity of a comment significantly increased the user-voted score of the comment and its replies.
Q3:	*How will the (antecedents and) consequences of toxicity elicitation differ from those of language toxicity?*
Result:	(Volume) The toxicity elicitation of a comment showed mixed correlations with the volume of discussion in different subreddits. (Evaluation) In one of the subreddits, the toxicity elicitation of a comment significantly increased the user-voted score of its replies.

Regarding the volume of discussion, the language toxicity of a comment showed positive correlation with the number of its direct replies (p<.01) and its total descendants (p<.001) in all subreddits except r/announcements. Meanwhile, only in r/todayilearned did comment toxicity show significant positive correlation with the height of the sub-discussion tree it led (i.e. the levels of its descendants) (p<.05). In general, there seemed a positive correlation between language toxicity and the volume of discussion, except in a more formal discussion community – r/announcements.

On the other hand, toxicity elicitation showed mixed correlations with discussion volume in different subreddits. Specifically, the toxicity elicitation of a comment showed negative correlation with the number of its descendants in r/announcements and r/worldnews (p<.01), while the correlation being positive in r/AskReddit (p<.05). Besides, there existed positive correlation between the toxicity elicitation of a comment and the levels of its descendants in r/todayilearned and r/AskReddit (p<.05). In r/politics, no significant correlation was observed between toxicity elicitation and discussion volume.

Regarding the user evaluation of discussion, the language toxicity of a comment showed positive correlation with the user-voted score of itself in r/AskReddit (p<.05), as well as with the maximum user-voted score of its direct replies in both r/politics (p<.05) and r/AskReddit (p<.01). Meanwhile, the toxicity elicitation of a comment showed positive correlation with the maximum score of its direct replies in r/AskReddit (p<.001), and no significant correlation was observed in other subreddits.

Although no correlation observed between the language toxicity or toxicity elicitation of a comment and the volume or evaluation of the discussion was coherent throughout all the metrics and subreddits, provided the expected negative effects of toxicity on the volume and evaluation of discussion (e.g. regulation), we would argue that our results indicate that toxicity probably embodied certain features that attracted response and recognition.

It should be noted that our results only indicate correlation instead of causality, and there might be multiple explanations underlying those quantitative results. However, we will try to complement them with qualitative examples and further interpret those results in the next section.

5 DISCUSSION

5.1 From within and without: Triggers of Toxicity

Our study first reveals with fine-grained quantitative evidence what the antecedents of toxicity were, and how much they mattered in triggering toxicity. Specifically, results of Study I first show that author propensity was a strong positive antecedent of language toxicity, which aligns with previous studies [5, 11, 61] that claimed certain individuals to be the source of toxic behaviors. Further, our study shows that controlling for that effect, toxicity in discussion context was also a strong positive antecedent of both language toxicity and toxicity elicitation, complementing previous works [13, 52, 64, 65] in further highlighting the significant effect of toxic discussion context in triggering subsequent toxicity.

Design implications: The notable effect of discussion context on toxicity generation implies that many users might be driven toward toxicity largely by contextual factors. One design implication that follows would be to mitigate the influence of contextual toxicity by somehow interfering with this process of toxicity generation, for example by providing users with real-time feedback of toxicity intensity in text as the users compose their comments: for users who are triggered toward toxic behavior mainly by contextual factors and initially chose to participate in the discussion out of good wills, a handful of hints should suffice to pull them back and remind them to reflect on the appropriateness of their comment.

5.2 Flames in Disguise: Toxicity Elicitation

In the paper, we examined an understudied aspect of toxicity – toxicity elicitation. In the way it was defined, toxicity-eliciting comments that do not necessarily seem toxic themselves might be another overlooked trigger of toxicity. We further inspected this phenomenon by identifying and analyzing examples of such toxicity-eliciting comments, among the top 500 comments in each subreddit that elicited more toxicity in their replies than they appeared to be themselves. We found that they mainly fell into four categories: i) strong-toned, condescending, ii) sarcastic, iii) digressive, and iv) against common sense/values, as shown in Table 9.

The first category consists of comments with strong or condescending tones. The left-column example showed a strong questioning tone ("If you guys ..., why do you ...?") on top of subjective

Table 9. Categories of inherently toxicity-eliciting comments.

Strong-toned, Condescending	**If you guys believe in a free an open internet, why do you silence conservative voices and run the whole website like a bunch of snowflakes?** - Did you write this comment? Is anyone silencing you? No, we just think you're an a**hole. We respect that you can write your opinion, provided it doesn't violate rules that have been established since when you opened the site. That doesn't mean that we have to *agree* with you, and people are free to argue with you. Now, in 8 months, when conservative viewpoints are bad for corporations and Verizon starts to throttle you, that's when you're going to be silenced.	**I'd tell you to grow up and deal with it like an adult.** *(A comment on the abortion of disabled fetuses.)* - I'd tell you to go f**k yourself it's not your decision! Abortion is absolutely legal in the US! :) - Choosing to abort a fetus you can't financially and/or emotionally take care of IS growing up and dealing with it like an adult.
Sarcastic	**So no matter what happens, no matter how they vote, you think they're bribed. I mean that's incredible, thank you for the insight into your world** - You really think they're not. I'm sorry for your stupidity.	**Very impressive! You've done well to skip this part <*quote*>** - I went out of my way to quote that part, you f**king maniac.
Digressive	**Give it to reddit to call for terrorism and a revolution because they will have to pay more for their Internet** *(A comment on the net neutrality repeal.)* - You are missing the point and trivializing the entire situation. I bet you wouldn't even lift a hand if Ajit Pai f**ked you in the a**. - [deleted]	**Perhaps you shouldn't get so angry because someone has a different viewpoint** - Everyone's entitled to their opinion, but your opinion is wrong and it makes you an a**hole. - Offensive opinions are offensive. - A viewpoint not based on facts is worthy of criticism.
Against Common Sense/Values	**Why is this a bad thing? This is awesome and it will be great for America.** *(A comment on the net neutrality repeal.)* - Your opinion is bulls**t and you should feel like bulls**t. - Lol ok guy. - How do you figure? - ACTIVE IN THESE COMMUNITIES The_donald What a surprise	**It should be enforced to be honest.** *(A comment on the abortion of disabled fetuses.)* - That's f**ked up. You should be ashamed. - [removed] - Holy s**t. You actually want to force people to abort disabled fetuses. You know that is what the Nazis did, right?

assumptions ("silence conservative voices", "run the whole website like a bunch of snowflakes"), while the right-column example showed condescension with an almost imperative tone ("I'd tell you to grow up ...").

The second category corresponds to sarcastic comments, that were usually intended to mock a certain target by saying the opposite of what was actually meant. It is clear that both examples shown in the table conveyed inherent disagreement and contempt under the disguise of approbation, consequently attracting irritated repliers.

The third category contains digressive comments, i.e. comments that missed the point of the discussion, possibly intentionally to get attention by leading toward a controversial topic (as in the left-column example), or for the purpose of defending oneself in an argument (as in the right-column example). It should be noted that this category connects to the "digress" strategy in covert trolling [31, 32].

The final category includes comments that went against common sense or common values. The ones against common sense would often be considered "dense", as shown in the left-column example in Table 1; and the ones against widely-accepted common values, e.g. those supporting the net neutrality repeal, or the enforcement of abortion of disabled fetuses (as shown in the last row of Table 9), would usually elicit very toxic replies that often involve swearing or personal attacks.

We believe those comments demonstrated inherent intentions, patterns or styles of language that naturally elicited toxicity, and it is very likely that those seemingly benign comments escaped regulation and led users toward impulsive toxic styles of speaking.

Design implications: The phenomenon of toxicity elicitation suggests opportunities to design algorithms that automatically detect and predict comments of high toxicity elicitation, beyond detecting comments of high language toxicity. The modeling process would be similar to the task of predicting whether a conversation goes awry [79] or controversial [35]. Moreover, online communities and forums seeking to regulate toxicity should develop policies and moderation procedures that also regulate those seemingly innocuous comments that actually stimulate toxicity, for example educating users with different categories of toxicity elicitation and encouraging them to report such comments – so that when facing toxicity elicitation, they could act in a way other than retorting with toxicity. Additionally, those user-labeled toxicity-eliciting comments potentially provide more accurate and fine-grained training data for developing the detection and prediction algorithms of toxicity elicitation.

5.3 The Multi-Faced Devil: Complexity of Toxicity

It is interesting that our results on the consequences of toxicity have shown the complexity of it. Specifically, in Study II, language toxicity showed positive correlation with discussion volume in most subreddits, despite being defined as something that likely draws people away from the discussion. Similarly, language toxicity showed positive correlation with the user recognition of the discussion (measured by the user-voted scores) in some of the subreddits, despite going against community norms.

To better understand why toxicity could sometimes increase the volume of the subsequent discussion and the user-voted scores, we examined in depth a few examples that obviously differed from the most typical toxic comments – those intended to attack a certain individual. For example, with regard to discussion volume, we found that toxicity might instead elicit response when it was intended to offend a large group of people, e.g. "Holy f**king s**t. I DIDN'T KNOW THAT HALF OF REDDIT ARE LITERAL COMMUNISTS". Although such comments usually got downvoted a lot, they would, at the same time, elicit a remarkable number of replies from the large group of people who felt offended.

Moreover, toxicity might as well elicit both response and recognition from other users. For example, when the toxicity in text reflected widespread anger toward some people or event, it would boost participation and recognition among users who shared the same feeling. Under a post in r/announcements about the F.C.C's repeal of the net neutrality rules, for instance, there existed a considerable number of highly toxic comments (toxicity score >.8) against the F.C.C. that earned both replies and user-voted scores (>100), possibly from users with similar viewpoints.

As another example, in r/AskReddit, where toxicity was shown to be positively correlated with both the volume and user evaluation of discussion in Study II, many highly toxic comments (toxicity score >.8) received a large number of replies (>10) and upvotes (>100) because they expressed strong feelings regarding a post or topic that were probably shared by many other people, for example "Our kneecaps are weak as f**k. Talk about a design flaw." and "what f**king world am i living in".

We can thus observe from the examples above that toxicity could actually vary in *target* (e.g. individual vs. group), *emotion* (e.g. hostility vs. anger) and *intention* (e.g. to insult a person vs. to express a feeling), and consequently different response would be elicited. This richer interpretation of toxicity should uncover part of the actual dynamics underlying our quantitative results, and provide insights for toxicity regulation in online discussions.

Design implications: Our analysis suggests that toxicity is a broad concept with nuanced subdimensions, which might lead to different outcomes. This first urges administrators to carefully examine the different types of toxicity present in the community as well as their corresponding outcomes, and then tailor the moderation policies to fit different types of toxic comments. For example, those intended to insult others should probably be fully regulated, while the moderators might have to consider the trade-off between user vitality and content quality when dealing with those toxic comments that express widely shared feelings and potentially elicit response and recognition. To aid human moderation, new computational methods could be developed to detect the nuanced emotions or intentions underlying the toxic text for distinguishing different types of toxicity, using newly built datasets of toxic text with richer labels and discussion context. Further, machine learning models could also be trained to predict the consequence of a toxic comment based on its text as well as preceding context, so that potential negative consequences could be avoided through early moderation.

6 LIMITATIONS & FUTURE WORK

We are aware that our work has been limited in multiple ways. First of all, we used existing NLP tools to assess the language toxicity of text. Although our verification analysis showed that the output of the Perspective API aligned well with human coders' judgement, it is still possible that our results have been affected by the limitations and biases of the tool. For example, prior research has indicated that the Perspective API might have overestimated the toxicity of African American English (AAE) comments [68]. As a result, the correlation effects we observed might have been boosted or reduced, depending on whether the existing biases were related to some other hidden factors that also correlated with the dependent variables or independent variables. For example in Study I, if our data included discussions within AAE users where the toxicity of comments were overestimated by the API, it is possible that the observed positive correlation between the language toxicity of the parent comment and the target comment would have been boosted by this bias. Similarly in Study II, if AAE users were inherently more (resp. less) likely to elicit response, the observed positive correlation between language toxicity and discussion volume would have been boosted (resp. reduced). We would love to see future studies that quantify the biases in the NLP tools we used, and reproduce our findings after correcting the biases.

Moreover, our dataset only covered about a month of time in a specific community (though in different sub-communities), and excluded regulated content. It is thus possible that our conclusions may not fully generalize to other time, community or regulation settings, and we expect similar studies to reveal how toxicity dynamics vary under different circumstances.

Also, our regression analyses were conducted over a limited range of metrics, and mainly focused on a single level of parent-child (comment-reply) structure, while the real flow of toxicity definitely involves influence and interaction along the entire cascade, and could be measured within more dimensions. We are therefore eager to see future works that delineate toxicity dynamics with more complicated models. It is also interesting how latent variables (e.g. intentions, emotions, opinions) come into play, but the modeling and inference of unexternalized factors should be rather challenging.

Last but not least, our work has called attention to, yet not probed into many interesting issues regarding toxicity and discussion dynamics. For example, in cases where toxicity elicits response and recognition, are there other ways to vitalize a discussion without toxicity? How do those resulted discussions differ in terms of popularity, fruitfulness, and the discussion experience they offer? How should we appropriately regulate toxicity, to still retain the "edge" of the discussion, but stop it from getting on people's nerves? We look to future studies to answer these questions.

7 CONCLUSION

We presented a study that explored the antecedents and consequences of toxicity – specifically language toxicity and toxicity elicitation – in online discussions, via quantitative analysis on Reddit comments. We found both author propensity and toxicity in discussion context to be strong antecedents of toxicity, and the volume and user evaluation of discussion to be positively correlated with language toxicity in some sub-communities. Our interpretation of the complicated triggers and outcomes of toxicity as well as the discussion of design implications should help understand and regulate toxicity in online discussions.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their insightful comments and Ge Gao, Diyi Yang, Dakuo Wang, Beisi Zhou and Xiaofeng Zhao for helpful discussions throughout the study. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 61932007 and Grant 61902075.

REFERENCES

[1] Ashley A Anderson, Dominique Brossard, Dietram A Scheufele, Michael A Xenos, and Peter Ladwig. 2014. The "nasty effect:" Online incivility and risk perceptions of emerging technologies. *Journal of Computer-Mediated Communication* 19, 3 (2014), 373–387.

[2] Ashley A Anderson, Sara K Yeo, Dominique Brossard, Dietram A Scheufele, and Michael A Xenos. 2016. Toxic talk: How online incivility can undermine perceptions of media. *International Journal of Public Opinion Research* 30, 1 (2016), 156–168.

[3] Lynne M Andersson and Christine M Pearson. 1999. Tit for tat? The spiraling effect of incivility in the workplace. *Academy of management review* 24, 3 (1999), 452–471.

[4] Ofer Arazy, Lisa Yeo, and Oded Nov. 2013. Stay on the Wikipedia task: When task-related disagreements slip into personal and procedural conflicts. *Journal of the American Society for Information Science and Technology* 64, 8 (2013), 1634–1648.

[5] Paul Baker. 2001. Moral panic and alternative identity construction in Usenet. *Journal of Computer-Mediated Communication* 7, 1 (2001), JCMC711.

[6] Sigal G Barsade. 2002. The ripple effect: Emotional contagion and its influence on group behavior. *Administrative Science Quarterly* 47, 4 (2002), 644–675.

[7] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. *arXiv preprint arXiv:2001.08435* (2020).

[8] Robert J Bies and Thomas M Tripp. 1996. Beyond distrust. *Trust in organizations* (1996), 246–260.

[9] Galen V Bodenhausen, Lori A Sheppard, and Geoffrey P Kramer. 1994. Negative affect and social judgment: The differential impact of anger and sadness. *European Journal of social psychology* 24, 1 (1994), 45–62.

[10] Porismita Borah. 2014. Does it matter where you read the news story? Interaction of incivility and news frames in the political blogosphere. *Communication Research* 41, 6 (2014), 809–827.

[11] Erin E Buckels, Paul D Trapnell, and Delroy L Paulhus. 2014. Trolls just want to have fun. *Personality and individual Differences* 67 (2014), 97–102.

[12] Damon Centola. 2010. The spread of behavior in an online social network experiment. *science* 329, 5996 (2010), 1194–1197.

[13] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 1217–1230.

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

[15] Anna Chmiel, Pawel Sobkowicz, Julian Sienkiewicz, Georgios Paltoglou, Kevan Buckley, Mike Thelwall, and Janusz A Hołyst. 2011. Negative emotions boost user activity at BBC forum. *Physica A: statistical mechanics and its applications* 390, 16 (2011), 2936–2944.

[16] Kevin Coe, Kate Kenski, and Stephen A Rains. 2014. Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication* 64, 4 (2014), 658–679.

[17] Lorenzo Coviello, Yunkyu Sohn, Adam DI Kramer, Cameron Marlow, Massimo Franceschetti, Nicholas A Christakis, and James H Fowler. 2014. Detecting emotional contagion in massive social networks. *PloS one* 9, 3 (2014).

[18] Yogesh Dahiya, Partha Talukdar, et al. 2016. Discovering response-eliciting factors in social question answering: A reddit inspired study. In *Tenth International AAAI Conference on Web and Social Media*.

[19] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

[20] Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Eighth International AAAI Conference on Weblogs and Social Media*.

[21] Joseph de Rivera and Carmen Grinkis. 1986. Emotions as social relationships. *Motivation and emotion* 10, 4 (1986), 351–369.

[22] Maeve Duggan. 2014. *Online harassment*. Pew Research Center.

[23] James H Fowler and Nicholas A Christakis. 2008. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study. *Bmj* 337 (2008).

[24] James H Fowler and Nicholas A Christakis. 2010. Cooperative behavior cascades in human social networks. *Proceedings of the National Academy of Sciences* 107, 12 (2010), 5334–5338.

[25] Nico H Frijda and Batja Mesquita. 1994. The social roles and functions of emotions. (1994).

[26] Iginio Gagliardone, Danit Gal, Thiago Alves, and Gabriela Martinez. 2015. *Countering online hate speech*. Unesco Publishing.

[27] Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*. 85–90.

[28] Himaanshu Gauba, Pradeep Kumar, Partha Pratim Roy, Priyanka Singh, Debi Prosad Dogra, and Balasubramanian Raman. 2017. Prediction of advertisement preference by fusing EEG response and sentiment analysis. *Neural Networks* 92 (2017), 77–88.

[29] Jamie Guillory, Jason Spiegel, Molly Drislane, Benjamin Weiss, Walter Donner, and Jeffrey Hancock. 2011. Upset now?: emotion contagion in distributed groups. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 745–748.

[30] Jeffrey T Hancock, Kailyn Gee, Kevin Ciaccio, and Jennifer Mae-Hwah Lin. 2008. I'm sad you're sad: emotional contagion in CMC. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. 295–298.

[31] Claire Hardaker. 2010. Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. *Journal of politeness research* 6, 2 (2010), 215–242.

[32] Claire Hardaker. 2013. "Uh....not to be nitpicky,,,,,but... the past tense of drag is dragged, not drug.": An overview of trolling strategies. *Journal of Language Aggression and Conflict* 1, 1 (2013), 58–86.

[33] Ali Hasan, Sana Moin, Ahmad Karim, and Shahaboddin Shamshirband. 2018. Machine learning-based sentiment analysis for twitter accounts. *Mathematical and Computational Applications* 23, 1 (2018), 11.

[34] Jared B Hawkins, John S Brownstein, Gaurav Tuli, Tessa Runels, Katherine Broecker, Elaine O Nsoesie, David J McIver, Ronen Rozenblum, Adam Wright, Florence T Bourgeois, et al. 2016. Measuring patient-perceived quality of care in US hospitals using Twitter. *BMJ Qual Saf* 25, 6 (2016), 404–413.

[35] Jack Hessel and Lillian Lee. 2019. Something's Brewing! Early Prediction of Controversy-causing Posts from Discussion Features. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 1648–1659.

[36] Jack Hessel, Lillian Lee, and David Mimno. 2017. Cats and captions vs. creators and the clock: Comparing multimodal content to context in predicting relative popularity. In *Proceedings of the 26th International Conference on World Wide Web*. 927–936.

[37] Benjamin D Horne, Sibel Adali, and Sujoy Sikdar. 2017. Identifying the social signals that drive online discussions: A case study of reddit communities. In *2017 26th International Conference on Computer Communication and Networks (ICCCN)*. IEEE, 1–9.

[38] Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving Google's Perspective API Built for Detecting Toxic Comments. *arXiv preprint arXiv:1702.08138* (2017).

[39] Hyunseo Hwang, Youngju Kim, and Catherine U Huh. 2014. Seeing is believing: Effects of uncivil online debate on political polarization and expectations of deliberation. *Journal of Broadcasting & Electronic Media* 58, 4 (2014), 621–633.

[40] Aaron Jaech, Victoria Zayats, Hao Fang, Mari Ostendorf, and Hannaneh Hajishirzi. 2015. Talking to the crowd: What do people react to in online discussions?. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2026–2031.

[41] Joseph M Kayany. 1998. Contexts of uninhibited online behavior: Flaming in social newsgroups on Usenet. *Journal of the American Society for Information Science* 49, 12 (1998), 1135–1141.

[42] Jihie Kim and Jaebong Yoo. 2012. Role of sentiment in message propagation: Reply vs. retweet behavior in political communication. In *Social Informatics (SocialInformatics), 2012 International Conference on*. IEEE, 131–136.

[43] Adam DI Kramer. 2012. The spread of emotion via Facebook. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 767–770.

[44] Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 111, 24 (2014), 8788–8790.

[45] Robert E Kraut and Paul Resnick. 2012. *Building successful online communities: Evidence-based social design*. Mit Press.

[46] Travis Kriplean, Jonathan Morgan, Deen Freelon, Alan Borning, and Lance Bennett. 2012. Supporting reflective public thought with considerit. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. 265–274.

[47] Haewoon Kwak and Jeremy Blackburn. 2014. Linguistic analysis of toxic behavior in an online video game. In *International Conference on Social Informatics*. Springer, 209–217.

[48] Haewoon Kwak, Jeremy Blackburn, and Seungyeop Han. 2015. Exploring cyberbullying and other toxic behavior in team competition online games. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 3739–3748.

[49] Cliff Lampe and Paul Resnick. 2004. Slash (dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 543–550.

[50] Noam Lapidot-Lefler and Azy Barak. 2012. Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in human behavior* 28, 2 (2012), 434–443.

[51] Hangwoo Lee. 2005. Behavioral strategies for dealing with flaming in an online forum. *The Sociological Quarterly* 46, 2 (2005), 385–403.

[52] So-Hyun Lee and Hee-Woong Kim. 2015. Why people post benevolent and malicious comments online. *Commun. ACM* 58, 11 (2015), 74–79.

[53] Jennifer S Lerner, Julie H Goldberg, and Philip E Tetlock. 1998. Sober second thought: The effects of accountability, anger, and authoritarianism on attributions of responsibility. *Personality and Social Psychology Bulletin* 24, 6 (1998), 563–574.

[54] Steven Loria, P Keen, M Honnibal, R Yankovsky, D Karesh, E Dempsey, et al. 2014. Textblob: simplified text processing. *Secondary TextBlob: Simplified Text Processing* (2014).

[55] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*. 145–153.

[56] Patrick B O'sullivan and Andrew J Flanagin. 2003. Reconceptualizing 'flaming' and other problematic messages. *New Media & Society* 5, 1 (2003), 69–94.

[57] Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval* 2, 1-2 (2008), 1–135.

[58] Deokgun Park, Simranjit Sachar, Nicholas Diakopoulos, and Niklas Elmqvist. 2016. Supporting comment moderators in identifying high quality online news comments. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 1114–1125.

[59] John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deeper attention to abusive user content moderation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 1125–1135.

[60] Fabian Prochazka, Patrick Weber, and Wolfgang Schweiger. 2018. Effects of civility and reasoning in user comments on perceived journalistic quality. *Journalism Studies* 19, 1 (2018), 62–78.

[61] Adrian Raine. 2002. Annotation: The role of prefrontal deficits, low autonomic arousal, and early health factors in the development of antisocial and aggressive behavior in children. *Journal of Child Psychology and Psychiatry* 43, 4 (2002), 417–434.

[62] reddit contributors. 2020. Reddiquette | Reddit Help. https://www.reddithelp.com/en/categories/reddit-101/reddit-basics/reddiquette [Online; accessed 16-May-2020].

[63] reddit contributors. 2020. What is karma? | Reddit Help. https://www.reddithelp.com/en/categories/reddit-101/reddit-basics/what-karma [Online; accessed 16-May-2020].

[64] Leonie Rösner and Nicole C Krämer. 2016. Verbal venting in the social web: Effects of anonymity and group norms on aggressive language use in online comments. *Social Media+ Society* 2, 3 (2016).

[65] Leonie Rösner, Stephan Winter, and Nicole C Krämer. 2016. Dangerous minds? Effects of uncivil online comments on aggressive cognitions, emotions, and behavior. *Computers in Human Behavior* 58 (2016), 461–470.

[66] Jeffrey Z Rubin, Dean G Pruitt, and Sung Hee Kim. 1994. *Social conflict: Escalation, stalemate, and settlement.* Mcgraw-Hill Book Company.

[67] Haji Mohammad Saleem, Kelly P Dillon, Susan Benesch, and Derek Ruths. 2017. A web of hate: Tackling hateful speech in online social spaces. *arXiv preprint arXiv:1709.10159* (2017).

[68] 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*. 1668–1678.

[69] Pnina Shachaf and Noriko Hara. 2010. Beyond vandalism: Wikipedia trolls. *Journal of Information Science* 36, 3 (2010), 357–370.

[70] John Suler. 2004. The online disinhibition effect. *Cyberpsychology & behavior* 7, 3 (2004), 321–326.

[71] Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*. 613–624.

[72] Wikimedia Support & Safety Team. 2015. Harassment Survey 2015. https://upload.wikimedia.org/wikipedia/commons/5/52/Harassment_Survey_2015_-_Results_Report.pdf [Online; accessed 4-April-2019].

[73] Jessica Vitak, Kalyani Chadha, Linda Steiner, and Zahra Ashktorab. 2017. Identifying women's experiences with and strategies for mitigating negative effects of online harassment. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1231–1245.

[74] William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*. 19–26.

[75] Tim Weninger, Xihao Avi Zhu, and Jiawei Han. 2013. An exploration of discussion threads in social news sites: A case study of the reddit community. In *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*. IEEE, 579–583.

[76] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*. 1391–1399.

[77] Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. 2012. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. 1980–1984.

[78] Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D Davison, April Kontostathis, and Lynne Edwards. 2009. Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB* 2 (2009), 1–7.

[79] Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1350–1361.

[80] Yabing Zhao, Xun Xu, and Mingshu Wang. 2019. Predicting overall customer satisfaction: Big data evidence from hotel online textual reviews. *International Journal of Hospitality Management* 76 (2019), 111–121.

[81] Marc Ziegele, Timo Breiner, and Oliver Quiring. 2014. What creates interactivity in online news discussions? An exploratory analysis of discussion factors in user comments on news items. *Journal of Communication* 64, 6 (2014), 1111–1138.

A COMPARISON OF PERSPECTIVE API MODELS

We experimented with the two different models provided by the Perspective API: TOXICITY and SEVERE_TOXICITY. Our results showed that on our sampled dataset, the TOXICITY model provided a smoother spectrum of toxicity scores (Figure 2), which better matched the human-labeled scores (Figure 3). The correlation between TOXICITY scores and human-labeled scores

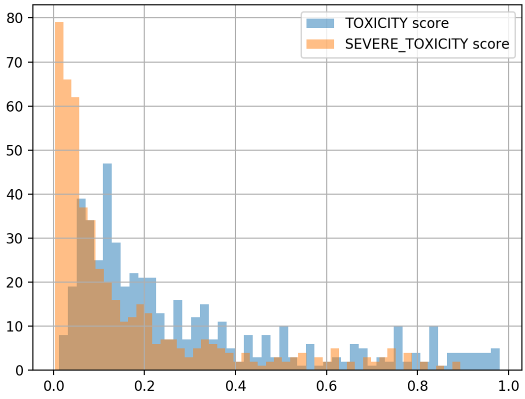

Fig. 2. Histograms of TOXICITY scores and SEVERE_TOXICITY scores of the sample.

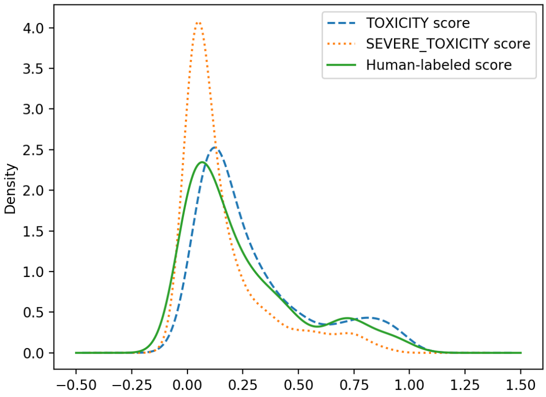

Fig. 3. Density plots of TOXICITY scores, SEVERE_TOXICITY scores and human-labeled scores of the sample.

(Pearson's r=.82, p<.001) was also higher than the correlation between SEVERE_TOXICITY scores and human-labeled scores (Pearson's r=.79, p<.001).

B ADDITIONAL RESULTS

The additional regression results (regression coefficients of control variables and corresponding significance levels) of Study I are shown in Table 10, and those of Study II are shown in Table 11.

Received January 2020; revised June 2020; accepted July 2020

Table 10. Additional regression results of Study I: Antecedents of Toxicity. ***: p<.001, **: p<.01, *: p<.05.

	Dependent Variable: ***Target-language-toxicity***				
Control Variable	Regression Coefficient in				
	r/annc	r/wn	r/pol	r/til	r/ar
Parent-length	-.050	.003	-.013	.043	-.026
Parent-%-caps	.046	.019	-.000	.054 *	.001
Parent-%-punctuation	.173 **	-.013	-.004	-.006	.019

	Dependent Variable: ***Target-toxicity-elicitation***				
Control Variable	Regression Coefficient in				
	r/annc	r/wn	r/pol	r/til	r/ar
Parent-length	-.041	.038	.013	-.001	-.021
Parent-%-caps	.067	.015	.010	-.007	.015
Parent-%-punctuation	-.115	.043	.010	-.014	.042 *

Table 11. Additional regression results of Study II: Consequences of Toxicity. ***: p<.001, **: p<.01, *: p<.05.

	Dependent Variable: *Target-#children*				
Control Variable	Regression Coefficient in				
	r/annc	r/wn	r/pol	r/til	r/ar
Target-positivity	-.105	-.001	-.015	.021	-.064 **
Target-negativity	-.091	-.057 *	.005	-.028	-.014
Target-length	-.084	.069 **	.043 **	.095 ***	.090 ***
Target-%-caps	.002	-.030	.027	.023	-.012
Target-%-punctuation	.089	-.030	-.011	-.018	.033
Intercept	.606 ***	.709 ***	.623 ***	.626 ***	.586 ***

	Dependent Variable: *Target-#descendants*				
Control Variable	Regression Coefficient in				
	r/annc	r/wn	r/pol	r/til	r/ar
Target-positivity	-.048	.077 ***	-.014	.082 ***	-.098 ***
Target-negativity	.000	-.066 **	.006	-.054 **	.008
Target-length	-.098	.110 ***	.128 ***	.188 ***	.052 **
Target-%-caps	-.072	-.075 ***	-.028	-.004	-.048 **
Target-%-punctuation	.256 ***	-.018	-.030 *	-.026	.007
Intercept	1.89 ***	2.32 ***	1.83 ***	2.19 ***	1.70 ***

	Dependent Variable: *Target-height*				
Control Variable	Regression Coefficient in				
	r/annc	r/wn	r/pol	r/til	r/ar
Target-positivity	.075	.050 *	.010	.006	-.018
Target-negativity	.044	-.010	.012	.010	.010
Target-length	.096	.024	.071 ***	.007	.024
Target-%-caps	-.080	-.020	-.010	-.027	-.004
Target-%-punctuation	.027	-.023	.004	-.009	-.016
Intercept	1.05 ***	1.29 ***	.975 ***	1.12 ***	.881 ***

	Dependent Variable: *Target-score*				
Control Variable	Regression Coefficient in				
	r/annc	r/wn	r/pol	r/til	r/ar
Target-positivity	-.068	-.004	-.001	.006	.004
Target-negativity	-.068	-.018	.002	-.021	-.006
Target-length	-.054	-.009	.003	-.002	-.013
Target-%-caps	.013	-.011	.008	.027	.002
Target-%-punctuation	.050	-.005	-.002	.007	.019
Intercept	6.03 ***	6.00 ***	5.97 ***	6.06 ***	6.07 ***

	Dependent Variable: *Children-score-max*				
Control Variable	Regression Coefficient in				
	r/annc	r/wn	r/pol	r/til	r/ar
Target-positivity	-.035	-.022	.001	-.000	-.031 *
Target-negativity	-.051	-.036	.012	-.036	-.009
Target-length	-.054	-.016	-.002	-.015	-.048 **
Target-%-caps	.005	-.023	.003	.036 *	-.026
Target-%-punctuation	.062	-.007	.002	-.006	.063 ***
Intercept	4.67 ***	4.75 ***	4.69 ***	4.86 ***	4.90 ***