CSCW 2020

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

Yan Xia
Fudan University

Haiyi Zhu
Carnegie Mellon University

Tun Lu
Fudan University

Peng Zhang
Fudan University

Ning Gu
Fudan University

Motivation

Parent – 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.

Target – Holy s**t you're dense, aren't you? ?

Child – ? ? ?

He literally did. It's *Statute* you numbskull.

Antecedents?

Toxicity elicitation

Language toxicity

Toxicity

Consequences?

Icons made by Freepik from Flaticon.com

Related Work & Hypotheses

- Antecedents of language toxicity
  - H1: **Author's propensity toward toxicity** will increase the language toxicity of his/her text.
  - H2: **Author's experience in the community** will reduce the language toxicity of his/her text.
  - H3: **Toxicity in discussion context** will increase the language toxicity of text.
  - H4: **Polarity in discussion context** will increase the language toxicity of text.

- Consequences of language toxicity
  - Q1: How will the language toxicity of text influence the **volume of discussion**?
  - Q2: How will the language toxicity of text influence the **evaluation of discussion**?

- Antecedents and consequences of toxicity elicitation
  - Q3: How will the antecedents and consequences of **toxicity elicitation** differ from those of language toxicity?

Method

- Reddit comments
  - r/announcements, r/worldnews, r/politics, r/todayilearned, r/AskReddit
  - 19,682/152,632/200,635/139,353/265,893 comments => preprocessed

- Quantifying comment features: NLP tools
  - Toxicity: Perspective API (* with human validation)
  - Polarity: TextBlob library

- Regression analyses
  - Study I, Antecedents of Toxicity: What author/parent comment features predict target comment toxicity?
  - Study II, Consequences of Toxicity: What child comment features are predicted by target comment toxicity?

Study I: Antecedents of Toxicity

		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

H1: **Author's propensity toward toxicity** will increase the language toxicity of his/her text. **Supported**

H2: **Author's experience in the community** will reduce the language toxicity of his/her text.

H3: **Toxicity in discussion context** will increase the language toxicity of text. **Supported**

H4: **Polarity in discussion context** will increase the language toxicity of text.

Study I: Antecedents of Toxicity

		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 ***

Q3: How will the antecedents (and consequences) of **toxicity elicitation** differ from those of language toxicity?

Study II: Consequences of Toxicity

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 *

Q1: How will the language toxicity of text influence the **volume of discussion**?

Study II: Consequences of Toxicity

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 *

Q3: How will the (antecedents and) consequences of **toxicity elicitation** differ from those of language toxicity?

Study II: Consequences of Toxicity

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 ***

Q2: How will the language toxicity of text influence the **evaluation of discussion**?

Study II: Consequences of Toxicity

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 ***

Q3: How will the (antecedents and) consequences of **toxicity elicitation** differ from those of language toxicity?

Discussion & Design Implications

- From within and without: **Triggers of toxicity**
  - Toxicity in discussion context
  - Design implication: To interfere with this toxicity generation process
- Flames in disguise: **Toxicity elicitation**
  - Strong-toned / Sarcastic / Digressive / Against common sense or values
  - Design implication: To regulate toxicity-eliciting comments
- The multi-faced devil: **Complexity of toxicity**
  - Different target / emotion / intention => Different consequence
  - Design implication: To distinguish different types of toxic comments in detection and regulation

Limitations & Future Work

- Bias of NLP tools
- Specific time, community and regulation settings
- Limited modeling of toxicity dynamics
- Further questions:
  - Other ways to vitalize a discussion without toxicity?
  - How will the resulted discussions differ?
  - How to stop toxicity from offending people while retaining the “edge” of discussion?

Thanks to the anonymous reviewers and Ge Gao, Diyi Yang, Dakuo Wang, Beisi Zhou, Xiaofeng Zhao for helping us with the study.

Thank you for listening!

2020.09.23