Designing ‘Understanding Mechanisms’ to Fight Online Harassment

Shagun Jhaver
Georgia Institute of Technology
Atlanta, GA 30308, USA
sjhaver3@gatech.edu

Abstract
Online harassment is a growing and significant social problem. Most of the efforts to address online harassment have focused on creating moderation tools. A persistent complaint with current moderation mechanisms is that they often can’t distinguish passionate disagreements from online harassment. In this position paper, I argue that researchers and designers should focus on creating tools that can make these nuanced distinctions. I also call for designing different solutions for addressing online harassment and sincere misunderstandings. I propose that tools that influence individuals to find common ground despite having different views can be used to help manage the problem of online abuse.

Author Keywords
Online harassment; Content moderation; Blocklists.

ACM Classification Keywords
H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous

> *“[The dialog] is ultimately derailing ’cause the entire thing has been, we’re trying to have one conversation, and the other people just call us sexists and harassers, and it’s like, that’s not a response!" (P2) [7]*

Online Harassment

The problems of online harassment and digital hate crimes have grown increasingly more salient in recent years. According to a 2017 Pew research study, 41% of American adults have suffered online harassment and 66% of adults have witnessed at least one harassing behavior online [4]. Online harassment can have a deeply negative impact on its victims. They can suffer from trauma, anxiety, depression and other emotional problems. In some cases, the victims may even commit suicide [1]. The victims' injuries are exacerbated by the fact that search engines often index abusive posts that other users may access years after those posts first appear [2].

There has been an increasing demand from the users of social media sites like Facebook and Twitter that the sites should do more to protect their users [8, 13]. It is also in the business interests of these sites to ensure that their communities are not abusive. Prior research has found that many users leave online communities if they become too toxic [8]. Therefore, it is critical that effective steps be taken to control the problem of online harassment.

Limitations of Moderation Mechanisms

One key approach to addressing the problem of online harassment has been to develop moderation and blocking mechanisms [3, 5, 10]. Many platforms have taken steps such as designing anti-abuse policy and implementing centralized and distributed moderation mechanisms [10] and tools to report abusive behavior [3]. However, these tools often fall short of addressing the needs of the harassment victims. Many users complain that they are not adequately protected from online harassment and at the same time, a number of users feel they are censored unfairly [8].

This highlights the need for researchers to delve deeper into the mechanisms associated with online harassment and develop more effective anti-abuse tools. Online harassment is not just a technical but a social problem. In their work on identifying women's experiences with online harassment, Vitak et al. have called for researchers to work closely with the platforms where abuse is most prevalent and study the perspectives of the attackers as well as the victims [12].

Over the past two years, my advisors (Amy Bruckman and Eric Gilbert) and I have conducted studies to understand the perspectives and experiences of both users who have suffered online harassment and those who have been accused of harassing others [7, 8]. We have found that the tradeoffs between online harassment and freedom of speech can often be complex and subjective. Everyone agrees that death threats and rape threats are abusive behaviors and should be censored but beyond that, where do we draw the line? If we consider too broad a spectrum of online dispute under the umbrella of online harassment, it can provoke reactions that are problematic.

In our interviews with users accused of online harassment, we found that when users with passionate political views feel that their postings are censored unfairly and that their legitimate complaints are dismissed, it adds to their anger and promotes more aggressive behavior [7]. This suggests that measured responses to activities of such users may result in better outcomes. This does not mean that it is ethical to appease the true harassers. However, distinguishing harassers from those with passionate views is important to prevent users from becoming more abusive. Both human moderators and automated tools should strive to understand the context and local social norms when making these distinctions during moderation.

Designing 'Understanding Mechanisms' as Complimentary Solutions

I suggest that researchers and designers should consider different solutions to address the problem of harassment and instances of sincere misunderstanding. We need more effective and nuanced content moderation solutions that detect abusive activities online and take actions to stop them. At the same time, I argue that a complimentary way to address these issues is to build tools that help individuals with opposing ideologies understand one another. This task is challenging - in our research, I have noticed that individuals on opposite sides of political debates often have deeply negative views of one another [7, 8]. Such views are exacerbated by the presence of filter bubbles online [11]. However, I have also observed that many users frequently share the same moral values as their opponents. This suggests that there are opportunities for mitigating potential abuse by creating good dialog between users with opposing views.

Design solutions that focus on creating civil conversations can influence individuals to not stereotype others they don't agree with [9]. It can also help mitigate situations where differences in cultural and social norms of different groups and misunderstandings create conditions for hostility. Tools that help bridge across different norms of politeness by disentangling the mode of address from content can help facilitate more civil conversations [6].

In conclusion, I call for researchers to study online sites where individuals find common ground despite their different values [9] and develop tools that help create good dialog. I believe that this line of work along with efforts to improve content moderation will help address the problem of online harassment and foster fairer, safer and more tolerant online communities.

REFERENCES

1. Zahra Ashktorab and Jessica Vitak. 2016. Designing Cyberbullying Mitigation and Prevention Solutions through Participatory Design With Teenagers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*. ACM Press, New York, New York, USA, 3895–3905.
2. Danielle Keats Citron. 2014. *Hate crimes in cyberspace*. Harvard University Press.
3. Kate Crawford and Tarleton Gillespie. 2016. What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society* 18, 3 (mar 2016), 410–428. DOI: http://dx.doi.org/10.1177/1461444814543163
4. Maeve Duggan. 2017. Online Harassment. *Pew Internet Project* (2017).
5. R Stuart Geiger. 2016. Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space. *Information, Communication & Society* 19, 6 (2016).
6. Catherine Grevet. 2016. *Being nice on the internet: Designing for the coexistence of diverse opinions online*. Ph.D. Dissertation. Georgia Institute of Technology.
7. Shagun Jhaver, Larry Chan, and Amy Bruckman. 2018a. The view from the other side: The border between controversial speech and harassment on Kotaku in action. *First Monday* (2018).
8. Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018b. Online Harassment and Content Moderation: The Case of Blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)* (2018).

9. Shagun Jhaver, Pranil Vora, and Amy Bruckman. 2017. *Designing for Civil Conversations: Lessons Learned from ChangeMyView*. Technical Report. Georgia Institute of Technology.

10. 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*.

11. Eli Pariser. 2011. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin.

12. 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 ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '17*. DOI:http://dx.doi.org/10.1145/2998181.2998337

13. Charlie Warzel. 2016. "A Honeypot For Assholes": Inside Twitter's 10-Year Failure To Stop Harassment. (2016). https://www.buzzfeed.com/charliewarzel/a-honeypot-for-assholes-inside-twitters-10-year-failure-to-s