A Study of Cyberbullying Detection and Mitigation on Instagram

Zahra Ashktorab
College of Information Studies
University of Maryland, College Park
parnia@umd.edu

CSCW '16 Companion, February 27 - March 02, 2016, San Francisco, CA, USA
ACM 978-1-4503-3950-6/16/02.
http://dx.doi.org/10.1145/2818052.2874346

Abstract
My dissertation addresses developing applications to mitigate anxiety and depression resulting from cyberbullying. Through the “Continuum of Harm” framework, the technological solutions resulting from participatory design sessions I conducted with adolescents can be categorized through a three-pronged approach: 1) **Primary Prevention**, in which the cyberbullying incident is prevented before it starts; 2) **Secondary Prevention**, where the goal is to decrease the problem after it has been identified, and 3) **Tertiary Prevention**, when intervention occurs after a problem has already caused harm. **I investigate the design and effectiveness of technological mechanisms to mitigate cyberbullying through Tertiary prevention on the popular social networking platform Instagram.**

Author Keywords
cyberbullying detection; cyberbullying mitigation

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

Introduction
Cyberbullying is an umbrella term that captures instances of bullying, harassment, and intimidation through mediated platforms. With the growing popularity of social media and

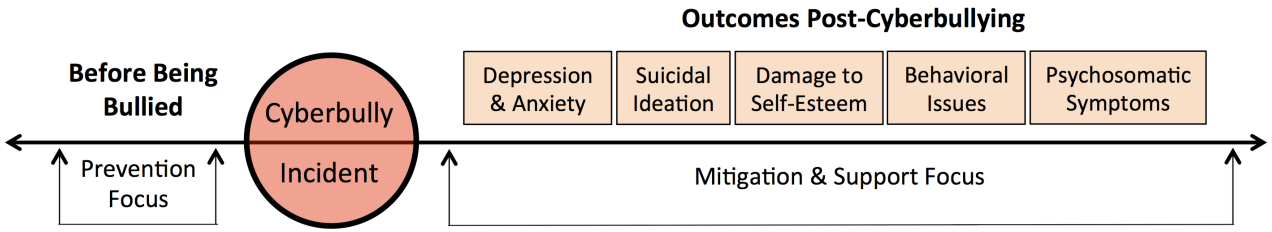

Figure 1: The Cyberbullying Continuum of Harm describes the different types of emotional distress may follow cyberbullying.

other forms of computer-mediated communication technologies, incidences of cyberbullying have significantly increased [15]. At least 42% of teens in the United States have experienced cyberbullying [13]. Victims of cyberbullying experience emotional problems like anxiety and depression[10, 8]. Teens who are bullied have a higher risk of suicide, which is currently the third leading cause of death among young people [10, 11, 8].

The advent of social media platforms like YouTube, Facebook, Instagram, and Ask.fm provide bullies with a larger and more widely visible platform through which they can harass their victims regardless of temporal or spatial constraints. The affordances of these platforms, such as the high visibility and persistence of posted content make it more difficult for victims to seek refuge from their tormentors [3]. The anonymity/pseudonymity that many sites offer further enables cyberbullying because bullies can hide their identity from their victims. For example, on Twitter, users create a "handle" that need not be connected to their real identity. Likewise, the high rate of abusive comments on question-asking sites Ask.fm and spring.me are attributed (in part) to the ability for users to post anonymously, i.e., without any identifying information showing [12]. Even in identified spaces such as Facebook or messaging services (texts, instant message), we are seeing an increase in the frequency and severity of cyberbullying messages [19]. The negative effects of cyberbullying, which include depression, anxiety and even suicide[7, 18, 16, 9] highlights the critical need for interventions to protect adolescents from the negative emotional effects that such harassing activities cause.

While Facebook has held the majority of the public and researchers' interest over the last decade, adolescents are increasingly flocking to other platforms, including Instagram and Snapchat. Teens are seeking privacy from their superiors on social networking platforms on which their parents are not active [6]. For my dissertation, I am 1) studying technological mechanisms for counteracting adolescent cyberbullying on Instagram, a popular social networking platform; 2) Developing an automatic detection algorithm for identifying cases of cyberbullying on Instagram; 3) Developing a Tertiary Prevention Cyberbullying Mitigation tool that sends positive messages to victims of cyberbullying; and 4) Evaluate the effectiveness of developed system. The primary focus will be on Instagram due to the increased prevalence of cyberbullying among adolescents; however, I will also consider mechanisms that can be applied to other social platforms. The project is guided by the following re-

search questions; my work specifically addresses the sub-questions:

1. **How can technical tools mitigate the instances of cyberbullying among adolescents on photo-sharing websites such as Instagram?**
   (a) **Does the effectiveness of mitigation vary based on the person from whom it is coming from?**
   (b) **Does the effectiveness of mitigation vary based on the publicness or privateness of the mitigation?**
   (c) **Based on the affordances of a platform such as Instagram, what is the most effective way to deliver cyberbullying mitigation?**

To address these questions, I am in the process recruiting Instagram users over Mechanical Turk who are over 18 years old, asking them preliminary questions about their social media use through a survey. I am developing an Instagram-specific cyberbullying detection classification method to accurately detect cyberbullying on Instagram, monitoring feeds of participants and automating the sending over various types of cyberbullying mitigation messages. I will be sending a survey after the support has been administered to evaluate the effectiveness of these messages.

Work in Progress

Until this point, I have conducted two pilot studies on cyberbullying discourse and mitigation that frame my current research questions. The first study explores the nature of cyberbullying on ask.fm through topic modeling. The second study involves conducting participatory design with school-aged children in order to co-design mitigation technologies.

Pilot Study I: Topic Modeling of Cyberbullying Data
As an initial exploratory study, I ran topic modeling on cyberbullying data on ask.fm, a social media platform that is known for it's proclivity for cyberbullying. In order to discover the different types of interactions on Ask.fm, we ran Latent Dirichlet Allocation topic modeling (LDA) on two samples of data for 10,20,30,40,50,60,70,80,90 and 100 topics [1] and found 11 types of discourse on ask.fm [1]

Pilot Study II: Participatory Design
In order to understand how youth would design technologies for cyberbullying mitigation, I conducted a total of ten participatory design sessions, five each with two classes at a local high school. The ninth graders were ages 14 and 15, while the twelfth graders were aged 17 and 18. Sessions were constrained to the times during which students had "free periods," which typically lasted 45 minutes. Since cyberbullying is common in high school settings [17], I chose two grades the youngest and older groups; this allowed me to gain insights into how perspectives vary between younger and older adolescents, who have different degrees of access to technology and social media [14]. In the design sessions, students were encouraged to think out of the box regarding what would be technologically possible. I analyzed the resulting solutions from our sessions through a framework that considers the different stages of cyberbullying symptoms and is based on preventative measures aimed at mitigating the "Continuum of Harm" in domestic violence [20] (see Figure 1). Through this framework, the technological solutions resulting from our design

[1] Additional categories included (1) Things that Annoy you/you Hate; (2) "listing all your followers"; (3) Picture/Video Request; (4) Thoughts and Opinions Discourse; (5) Like Solicitation and Rating Discourse; and (6) Preference Questions

sessions can be categorized through a three-pronged approach: 1) **Primary Prevention**, in which the cyberbullying incident is prevented before it starts; 2) **Secondary Prevention**, where the goal is to decrease the problem after it has been identified, and 3) **Tertiary Prevention**, when intervention occurs after a problem has already caused harm[20]. Two researchers who were involved in designing the participatory design sessions coded each resulting solution based on this framework individually.

Expected Contributions

In my dissertation project, I investigate the effectiveness of tertiary cyberbullying mitigation support on Instagram. My participatory design studies as well as recent literature reveal that Instagram is a popular social networking platform on which cyberbullying occurs [6]. Using the affordances of interaction on Instagram, a social networking platform that has risen in popularity among youth as reflected by the Pilot Work I have conducted, I am building and evaluating a system that consists of an automatic detection algorithm to flag instances of cyberbullying and then send support to victims of cyberbullying as an attempt to mitigate the potential emotional damage that the cyberbullying has caused. While there has been research conducted on co-designing with youth to create effective tools to counter cyberbullying [2], and attempts to automatically identify cyberbullying through classification algorithms [5, 4, 12],there has been zero attempts to evaluate the effectiveness of cyberbullying mitigation tools. This study is the first of its kind that proposes an actionable method of evaluating a cyberbullying mitigation tool.

The analysis and categorization of the different preventative types allows me to consider additional research questions, such as which preventative solution is most effective for cyberbullying prevention and how can we accurately measure this effectiveness? Until this point, technological cyberbullying prevention mechanisms have not been evaluated for effectiveness. The framework through which I look at potential solutions provides a straightforward way to begin to consider how one would compare different solutions. My work presents solutions to cyberbullying that were designed by the users most vulnerable to it: adolescents. Specific ways in which this study contributes to HCI are: 1) extending existing cyberbullying intervention design themes (specifically, Bowler et al. [2]) through the analysis of solutions designed with teenagers; and 2) Evaluating the effectiveness of cyberbullying mitigation techniques on Instagram.

REFERENCES

1. David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3 (2003), 993–1022.
2. Leanne Bowler, Eleanor Mattern, and Cory Knobel. 2014. Developing design interventions for cyberbullying: A narrative-based participatory approach. (2014).
3. Danah Boyd. 2009. Why youth (heart) social network sites: The role of networked publics in teenage social life. *Youth, identity, and digital media* (2009), 119âĂŞ–142.
4. Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2, 3 (2012), 18.
5. Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the Detection of Textual Cyberbullying.. In *The Social Mobile Web*.

6. Maeve Duggan. 2013. Photo and video sharing grow online. *Pew Research Internet Project* (2013).
7. Ryan Grenoble. 2012. Amanda Todd: Bullied Canadian Teen Commits Suicide After Prolonged Battle Online And In School. (october 2012). http://www.huffingtonpost.com/2012/10/11/amanda-todd-suicide-bullying_n_1959909.html
8. Sameer Hinduja and Justin W Patchin. 2010. Bullying, cyberbullying, and suicide. *Archives of Suicide Research* 14, 3 (2010), 206–221.
9. Elizabeth M Jaffe. 2010. Cyberbullies Beware: Reconsidering Vosburg v. Putney in the Internet Age. *Charleston L. Rev.* 5 (2010), 379.
10. Young Shin Kim and Bennett Leventhal. 2008. Bullying and suicide. A review. *International Journal of Adolescent Medicine and Health* 20, 2 (2008), 133–154.
11. Young Shin Kim, Bennett L Leventhal, Yun-Joo Koh, and W Thomas Boyce. 2009. Bullying increased suicide risk: prospective study of Korean adolescents. *Archives of suicide research* 13, 1 (2009), 15–30.
12. April Kontostathis, Kelly Reynolds, Andy Garron, and Lynne Edwards. 2013. Detecting cyberbullying: query terms and techniques. In *Proceedings of the 5th Annual ACM Web Science Conference*. ACM, 195–204.
13. Amanda Lenhart. 2007. Cyberbullying. *Pew Internet & American Life Project* (2007).
14. Sonia Livingstone and Ellen Helsper. 2007. Gradations in digital inclusion: children, young people and the digital divide. *New media & society* 9, 4 (2007), 671–696.
15. Gustavo S Mesch. 2009. Parental mediation, online activities, and cyberbullying. *CyberPsychology & Behavior* 12, 4 (2009), 387–393.
16. Matthew C Ruedy. 2007. Repercussions of a myspace teen suicide: Should anti-cyberbullying laws be created. *NCJL & Tech.* 9 (2007), 323.
17. Shari Kessel Schneider, Lydia O'Donnell, Ann Stueve, and Robert WS Coulter. 2012. Cyberbullying, school bullying, and psychological distress: A regional census of high school students. *American Journal of Public Health* 102, 1 (2012), 171–177.
18. Joe Shute. 2013. Cyberbullying suicides: What will it take to have Ask.fm shut down? (August 2013). http://www.telegraph.co.uk/health/children_shealth/10225846/Cyberbullying-suicides-What-will-it-take-to-have-Ask.fm-shut-down.html
19. PK Smith, Jess Mahdavi, Manuel Carvalho, and Neil Tippett. 2006. An investigation into cyberbullying, its forms, awareness and impact, and the relationship between age and gender in cyberbullying. *Research Brief No. RBX03-06. London: DfES* (2006).
20. David A Wolfe and Peter G Jaffe. 1999. Emerging strategies in the prevention of domestic violence. *The future of children* (1999), 133–144.