"Is This Cyberbullying or Not?": Intertwining Computational Detection with Human Perception (A Case Study)

Edward Dillon, Jamie Macbeth, Robin Kowalski, Elizabeth Whittaker and Juan E. Gilbert

Abstract Cyberbullying refers to bullying that occurs through the Internet or text messaging. Understanding the nature of cyberbullying and its implications has become an important issue in society. In an attempt to assist with intervention and prevention efforts, the development of computational systems for detecting acts of cyberbullying has become a common trend. However, prior research notes that such systems are typically vulnerable to inaccurate detections, in particular false-positives. Given the prevalence of cyberbullying across age demographics, understanding how humans identify such activity is important for informing and improving such prevention/intervention efforts and reducing system vulnerability. A study was conducted that asked 180 participants to evaluate three excerpts taken from the social media site Formspring. Participants indicated that the use of profane words, and the determination that someone was harmed by the content of the social media post were the most likely determinants that cyberbullying occurred in the post.

Keywords Cyberbullying · Detection · Human perception

1 Introduction

Youth today have been described as the "always on" generation or as "forever tied to their technology." A cursory look at statistics showing prevalence rates of Internet and cellular phone use confirm that youth today are growing up in a wired

E. Dillon (✉) · J.E. Gilbert
Department of Computer and Information Science and Engineering,
University of Florida, Gainesville, FL 32611, USA
e-mail: ecdillon@cise.ufl.edu

J. Macbeth
Department of Electrical and Computer Systems Engineering,
Fairfield University, Fairfield, CT 06824, USA

R. Kowalski · E. Whittaker
Department of Psychology, Clemson University, Clemson, SC 29634, USA

D. Nicholson (ed.), *Advances in Human Factors in Cybersecurity*,
Advances in Intelligent Systems and Computing 501,
DOI 10.1007/978-3-319-41932-9_28

culture [1]. While technology allows individuals to bridge global divides almost instantaneously among other advantages [2–4], it creates digital divides between those who are adept at technology use (e.g., adolescents) and those who are less so (e.g., parents; educators). Furthermore, in spite of the fact that technology builds social relationships easily with the click of a button, these same relationships can be quickly ended through behaviors such as cyberbullying.

Understanding cyberbullying has been complicated by a failure of researchers to reach a consensus regarding exactly how cyberbullying should be defined and/or measured [5, 6]. Some researchers define cyberbullying broadly as bullying that occurs through the use of technology, specifically the Internet and cellular phones. Others adopt a more narrow approach to identifying cyberbullying by the venue by which the cyberbullying occurs (e.g., chat rooms, online gaming, webpages). These differing conceptualizations influence the ways in which cyberbullying is measured which, in turn, affects reported prevalence rates.

Prevalence rates of cyberbullying typically range between 1 and 40 % [1, 7–9]. Reported prevalence rates depend not only on how cyberbullying is defined, but also on demographic characteristics of the sample (e.g., age, sex, race), the venue used to determine whether cyberbullying occurred, the time frame used for determining cyberbullying (e.g., one month, six months, lifetime); and, the liberal versus conservative criteria used to determine cyberbullying (i.e., must have occurred at least once versus two to three times a month or more).

Cyberbullying has often been compared to traditional bullying in that both behaviors involve acts of aggression that are intended to cause harm or distress, and that occur among individuals whose relationship is characterized by a power imbalance [5, 10–12]. While sharing certain features in common, cyberbullying and traditional bullying differ in important ways. Perpetrators of cyberbullying often hide beneath an umbrella of anonymity. People will say and do things anonymously that they would not say to someone's face, opening up the pool of individuals who might be willing to perpetrate cyberbullying [13]. In addition, whereas most traditional bullying occurs at school during the school day [14], cyberbullying can occur at any time. Thus, victims are accessible to perpetrators any time of the day or night. Punitive fears surrounding the two types of bullying often differ as well. Though few youth, in particular, report victimization, the reasons for not reporting differ between traditional bullying and cyberbullying. Victims of traditional bullying fear retaliation by the perpetrator. Victims of cyberbullying fear that adults will remove the technological means by which the victimization is occurring. In addition, youth victims of cyberbullying are often reluctant to tell adults because they assume, probably correctly, that parents or educators will not be well-versed in the technology by which victimization occurred.

Much research remains to be conducted on cyberbullying. A number of key research issues surrounding cyberbullying include: (1) the extent of co-occurrence of involvement in cyberbullying and traditional bullying; (2) how people, particularly bystanders, react to cyberbullying situations; and, of relevance to the present study, (3) how people identify cyberbullying. From a victim's or a bystander's perspective, many decision points are involved in making the attribution that

someone has, indeed, behaved aggressively, that the other individual is, in fact, unequal to them in power, and that the behavior has the potential to be repeated over time.

Understanding cyberbullying and its potential implications is important. Furthermore, examining acts of cyberbullying through human perception could enhance prior and current attempts to address such behavior. For instance, there has been a trend of employing computational practices for detecting acts of cyberbullying in real-time. One practice has emphasized the creation of natural language processors for interpreting textual syntax from real-time conversations in the form of online messages or blogs that can be associated with acts of cyberbullying [15–18]. Another practice has focused on the development of interactive software that provides a method of intervention based on its detection of cyberbullying acts via textual analysis [19, 20]. However, both practices have been found to be vulnerable to the potential for false-positives [15, 21] and other inaccuracies [18, 20] during detection. Examining ways to reduce false-positives and related inaccuracies is critical to improving the detection of cyberbullying activities. Employing computational techniques/methods that are derived from human perception could help reduce such vulnerability.

In light of these limitations in previous research, the current study was designed to examine human perception of cyberbullying. In other words, the study investigated the cues that participants use to determine whether or not cyberbullying has occurred. Previous research by Kontostathis et al. [16] noted that "bad" words increased the likelihood that posts from Formspring would be labeled as cyberbullying (see also [20]). Prevalence rates of cyberbullying posts in their research ranged between 7 and 14 %. While also examining word choice in posts, the current study included a broader range of variables involved in determining whether cyberbullying has or has not occurred.

2 Method

A total of 180 participants (68.9 % female; 30.0 % male; 1.1 % did not report a gender) were recruited to review, interpret, and analyze real-world conversations that occurred online. The participants' ages ranged from 18 to 29 (M = 20.2; SD = 10.39). With regard to ethnicity, 86.0 % of the participants were Caucasian, 7.9 % African American, 2.8 % Asian or Asian American, 2.2 % Hispanic, 0.6 % American Indian, and 0.6 % of another ethnic background.

Conversations for this study were extracted from interactions that occurred on a social networking site called Formspring (currently renamed to Spring.me), which were pre-coded by three raters as either (a) definitely cyberbullying, (b) definitely not cyberbullying, or (c) ambiguous regarding cyberbullying. One conversation from each of the three categories was selected for inclusion in this study. Section 2.1 provides a background synopsis of each of the three selected conversations.

After completing a survey of questions regarding their personal experiences with cyberbullying, participants reviewed, analyzed, and interpreted three conversations in counterbalanced order. For each conversation, a list of targeted questions was used to gather explicit feedback about the participants' interpretation. These included:

- *Did acts of cyberbullying occur in the conversation?*
- *Why or what factors led you to believe cyberbullying have occurred?*
- *How hurtful did you find this conversation?*
- *Is this conversation representative to what you see online?*

To gather quantitative feedback during this assessment, many of the questions were asked in the form of yes/no responses, multiple choice, or Likert scales. However, one of these questions did allow for the participants to provide an open-ended response. This open-ended question was used to gather qualitative feedback during this assessment. For data analysis, one-way ANOVAs and T-tests were used to assess the participants' responses to these questions.

2.1 Conversation Synopsis

As previously mentioned, the participants were exposed to three conversations during this study. These conversations were composed of topics ranging from imposed threats of harm to derogatory remarks about a celebrity. The length of these respective conversations varied. Furthermore, the role of the individuals differed amongst these conversations.

Conversation #1 This conversation was based on a dialogue between multiple individuals. The individuals are engaged in a heated verbal exchange that includes derogatory language and threats of harm.

Conversation #2 This conversation was a brief dialogue between two individuals. Individual A asks a question regarding self-image while mentioning the possibility of losing weight. Individual B responds to Individual A's question with specific advice.

Conversation #3 This conversation was a mini-dialogue between two individuals. Individual A began the conversation using derogatory language about a well-known celebrity. Individual B confronts Individual A for this particular behavior while also using derogatory language during the process.

3 Results

3.1 Did Cyberbullying Occur?

The responses to this particular question were primarily composed of a yes/no format. Based on the responses given by the participants, it was determined that two

Table 1 Occurrence of cyberbullying activity—descriptive data

Conversation	Occurrence of cyberbullying acts		Male (N = 54[a])		Female (N = 124[a])	
	Yes (%)	No (%)	Yes (%)	No (%)	Yes (%)	No (%)
Conversation #1	91	9	98	2	90	10
Conversation #2	6	94	0	100	2	98
Conversation #3	82	18	81	19	88	12

One participant did not provide a response for Conversations #1 and #2
Three participants did not provide a response for Conversation #3
[a]One participant did not provide gender; this participant's data were excluded from the male/female columns

of the three conversations exhibited acts of cyberbullying. In particular, 91 % of the participants indicated such activity to be true in Conversation 1, while 82 % mentioned the same for Conversation 3. Only 6 % of the participants believed acts of cyberbullying occurred in Conversation 2. Gender-wise, 98 and 90 % of the male and female participants, respectively, determined that cyberbullying occurred in Conversation 1, while 81 and 88 % respectively said the same for Conversation 3. Table 1 provides descriptive data for these respective findings.

After conducting statistical analysis on these response rates, a one-way ANOVA revealed a statistical significance: $F(2, 532) = 414.78$, $p < 0.01$. T-tests were employed to conduct direct comparisons between the response rates for each respective conversation. These tests revealed a statistical significance for all three paired comparisons: *Conversation #1 versus Conversation #2* ($p < 0.01$), *Conversation #1 versus Conversation #3* ($p = 0.03$), and *Conversation #2 versus Conversation #3* ($p < 0.01$). These results suggest that the participants reported acts of cyberbullying at significantly higher rates for Conversation #1 than for Conversations #2 and #3, respectively. Significantly higher rates were also reported for Conversation #3 when compared to Conversation #2.

3.2 Why or What Factors Led to This Decision?

For each of the Formspring conversations, we performed a qualitative analysis of respondents' open-ended textual explanations of their judgments about whether the conversation contained cyberbullying or not. Threats of physical violence, personal ridicule/harassment, and profanity/derogatory terms were found to be trending responses given by the participants for these respective conversations when cyberbullying was determined. They directly quoted certain words and phrases that appeared in the conversation, and described the language using a variety of terms such as "derogatory", "vulgar", "abusive", "hateful", "vicious" and "demeaning". In relation to this trend, the participants' responses were categorized using the following classifications:

Table 2 Trends of cyberbullying activity by classification—descriptive data

Conversation #1	Conversation #2	Conversation #3
Classification 1: 19 %	Classification 1: 0 %	Classification 1: 6 %
Classification 2: 4 %	Classification 2: 1 %	Classification 2: 8 %
Classification 3: 50 %	Classification 3: 1 %	Classification 3: 40 %
Classification 4: 18 %	Classification 4: 0 %	Classification 4: 28 %
Classification 5: 2 %	Classification 5: 4 %	Classification 5: 4 %
Classification 6: 7 %	Classification 6: 94 %	Classification 6: 14 %

- *Classification 1: Threats of physical violence*
- *Classification 2: Making fun or belittling someone's intelligence*
- *Classification 3: Profanity, derogatory terms, name calling, or insults*
- *Classification 4: Other, general, or more ambiguous feedback*
- *Classification 5: N/A or No Response*
- *Classification 6: No Cyberbullying Occurred*

Table 2 depicts the categorized responses given by the participants for each respective conversation.

After conducting statistical analysis on these response rates, T-tests revealed that Conversation #1 received the highest rate for cyberbullying activity on this basis of *profanity*, *derogatory terms*, *name calling*, or *insults*. Specifically, responses for this particular classification were detected in Conversation #1 at a significantly higher rate than Conversation #2 ($p < 0.01$) and Conversation #3 ($p < 0.01$), respectively. In a similar manner, Conversation #3 received significantly higher rates for cyberbullying activity on the basis of *other*, *general*, or *more ambiguous* responses than Conversations #1 ($p = 0.05$) and Conversation #2 ($p < 0.01$) respectively. In general, Conversation #2 was found to exhibit significantly less tendencies for cyberbullying behavior overall than Conversation #1 ($p < 0.01$) and Conversation #3 ($p < 0.01$) respectively.

For some respondents, the anonymity of one of the conversants in the conversation, as well as anonymity provided by the Internet and social media generally, played a major part in the identifiability of cyberbullying. Finally, some respondents picked up on the evidence in the conversation that the hurtful posts were reoccurring over a period of time, and that they continued in spite of appeals by one of the parties to stop them.

3.3 How Hurtful Were These Conversations?

The responses to this particular question were based on a 3-point Likert scale, where 1 = *not at all hurtful*, 2 = *moderately hurtful*, and 3 = *very hurtful*. Table 3 details the calculated means and standard deviations for each conversation. A one-way ANOVA revealed statistical significance: $F(2, 534) = 359.00$, $p < 0.01$.

Table 3 Level of hurt portrayed by conversations—descriptive data

Conversation	Mean	Standard deviation
Conversation #1	2.41	0.63
Conversation #2	1.03	0.18
Conversation #3	1.84	0.55

T-tests revealed that Conversation #1 was the most hurtful, followed by Conversation #3 and Conversation #2 (least hurtful).

3.4 How Representative Are These Conversations to Real Online Conversations?

The responses to this particular question were based on a 3-point Likert scale, where 1 = *not at all representative*, 2 = *moderately representative*, and 3 = *very representative*. Table 4 details the calculated means and standard deviations for each conversation. When conducting statistical analysis on these results, a one-way ANOVA revealed a statistical significance: $F(2, 531) = 12.38$, $p < 0.01$. T-tests revealed that Conversation #3 portrayed the greatest representation to real conversations that occur online, Conversation #2 displayed the second greatest portrayal, and Conversation #1 was portrayed to be the least representative.

4 Discussion

The results of the current study show that participants are able to differentiate instances in which cyberbullying does and does not occur. Additionally, the study highlights the cues participants used to determine whether cyberbullying has occurred.

Our results may inform researchers in computing and information science disciplines who are building natural language processing systems with the intention of detecting cyberbullying in real time by analyzing the texts of social media posts or text messages. In particular, our findings regarding how people made their determinations about whether cyberbullying was present may have direct implications for the likely success of different architectures for these systems. In many cases, our respondents found particular words or terms in the language of the conversations that made the cyberbullying easy to identify. This confirms Kontostathis, Reynolds,

Table 4 Representation of conversations—descriptive data

Conversation	Mean	Standard deviation
Conversation #1	1.77	0.65
Conversation #2	1.98	0.61
Conversation #3	2.09	0.63

Garron, and Edwards' conclusion, who also found that language used was also a factor [16]. One simple method for automatically detecting cyberbullying is to develop a dictionary of particular words, terms, or phrases that indicate bullying, and program the system to search messages for these words, flagging messages or posts as cyberbullying when these words are found. Our results suggest that these methods may be successful in some cases.

However, many other respondents, rather than pointing out specific words or phrases that indicated cyberbullying, appeared to take more of a wholistic approach where they focused on whether anyone involved in the conversation or reading the conversation would have been harmed, or whether feelings may have been hurt, independently of the particular words which were used. Some research has shown that many acts of cyberbullying, including slights, insults, and mocking, are highly personalized and contextual and may not use profanity [22]. As one of our respondents stated: "I think any language used with an intent to harm is cyberbullying—it doesn't matter if they're swearing or not." Building software to detect acts of cyberbullying in ways that correspond to these kinds of human judgments is more challenging because it requires systems to understand the meaning of a post or message at a far deeper level than keyword search.

5 Conclusion and Future Work

The purpose of this study was to determine the impact that human perception could contribute to the behavior of cyberbullying. As the results revealed, human perception can provide further insights and enhanced sentimental perspectives about the nature of cyberbullying activities. For this current study, conversations with varying levels of cyberbullying activity were intentionally employed. For future work, studies that utilize conversations of similar levels of cyberbullying behavior could be employed. Another future work will be to examine how the attributes of human perception can be transferred into intelligent systems. Developing computational systems that are able to employ such perceptions could improve the overall impact of cyberbullying detection.

Acknowledgments We would like to acknowledge and thank April Kontostathis from Ursinus College for granting us permission to use her FormSpring dataset to conduct this study.

References

1. Lenhart, A.: Cyberbullying: what the research is telling us. Retrieved 16 Apr 2011 from http://www.pewinternet.org/Presentations/2010/May/Cyberbullying-2010.aspx (2010)
2. Ling, R., Stald, G.: Mobile communities: are we talking about a village, a clan, or a small group? Am. Behav. Sci. **53**, 1133–1147 (2010)
3. Magoc, D., Tomaka, J., Bridges-Arzaga, A.: Using the web to increase physical activity in college students. Am. J. Health Behav. **35**, 142–154 (2011)

4. Zhan, Z., Xu, F., Ye, H.: Effects of an online learning community on active and reflective learners' learning performance and attitudes in a face-to-face undergraduate course. Comput. Educ. **56**, 961–968 (2011)
5. Kowalski, R.M., Giumetti, G., Schroeder, A., Lattanner, M.: Bullying in the digital age: a critical review and meta-analysis of cyberbullying research among youth. Psychol. Bull. **140**, 1073–1137 (2014)
6. Smith, P.K., del Barrio, C., Tokunaga, R.: Definitions of bullying and cyberbullying: how useful are the terms? In: Bauman, S., Cross, D., Walker, J. (eds.) Principles of Cyberbullying Research: Definition, Measures, and Methods, pp. 29–40. Routledge, Philadelphia (2012)
7. O'Brennan, L.M., Bradshaw, C.P., Sawyer, A.L.: Examining developmental differences in the social-emotional problems among frequent bullies, victims, and bully/victims. Psychol. Schools **46**, 100–115 (2009)
8. Pontzer, D.: A theoretical test of bullying behavior: parenting, personality, and the bully/victim relationship. J. Family Violence **25**, 259–273 (2010)
9. Tokunaga, R.S.: Following you home from school: a critical review and synthesis of research on cyber bullying victimization. Comput. Hum. Behav. **26**, 277–287 (2010)
10. Kowalski, R.M., Limber, S., Agatston, P.W.: Cyberbullying: bullying in the digital age, 2nd edn. Wiley, Malden (2012)
11. Olweus, D.: School bullying: development and some important challenges. Ann. Rev. Clin. Psychol. **9**, 751–780 (2013)
12. Patchin, J.W., Hinduja, S.: Cyberbullying: an update and synthesis of the research. In: Patchin, J.W., Hinduja, S. (eds.) Cyberbullying Prevention and Response: Expert Perspectives, pp. 13–36. Routledge, New York (2012)
13. Diener, E.: The Psychology of Group Influence. Erlbaum, New York (1980)
14. Nansel, T.R., Overpeck, M., Pilla, R.S., Ruan, W., Simons-Morton, B., Scheidt, P.: Bullying behaviors among U.S. youth: prevalence and association with psychosocial adjustment. J. Am. Med. Assoc. **285**, 2094–2100 (2001)
15. Bayzick, J., Kontostathis, A., Edwards, L.: Detecting the presence of cyberbullying using computer software. In: 3rd Annual ACM Web Science Conference (WebSci '11), pp. 1–2 (2011)
16. Kontostathis, A., Reynolds, K., Garron, A., Edwards, L.: Detecting cyberbullying: query terms and techniques. In: 5th Annual ACM Web Science Conference (WebSci '13), pp. 195–204 (2013)
17. Xu, J.M., Jun, K.S., Zhu, X., Bellmore, A.: Learning from bullying traces in social media. In: The 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 656–666 (2012)
18. Yin, D., Xue, Z., Hong, L., Davison, B.D., Kontostathis, A., Edwards, L.: Detection of harassment on web 2.0. In: Proceedings of the Content Analysis in the WEB, vol. 2, pp. 1–7 (2009)
19. Bosse, T., Stam, S.: A normative agent system to prevent cyberbullying. In: Web Intelligence and Intelligent Agent Technology (WI-IAT), IEEE/WIC/ACM International Conference, vol. 2, pp. 425–430 (2011)
20. Dinakar, K., Jones, B., Havasi, C., Lieberman, H., Picard, R.: Common sense reasoning for detection, prevention, and mitigation of cyberbullying. In: ACM Transactions on Interactive Intelligent Systems (TiiS), vol. 2, no. 3, p. 18 (2012)
21. Kontostathis, A., Edwards, L., Leatherman, A.: Text mining and cybercrime. In: Text Mining: Applications and Theory. Wiley, Chichester (2010)
22. Mishna, F., Cook, C., Gadalla, T., Daciuk, J., Solomon, S.: Cyber bullying behaviors among middle and high school students. Am. J. Orthopsychiatry **80**(3), 362–374 (2010)