Toxicity Online: Conceptual Devices for Understanding and Explaining Cyberbullying, Online Harassment, and Cyber Aggression

Leave Authors Anonymous
for Submission
City, Country
e-mail address

Leave Authors Anonymous
for Submission
City, Country
e-mail address

Leave Authors Anonymous
for Submission
City, Country
e-mail address

ABSTRACT

This paper reviews various terms used by Internet researchers: “cyberbullying”, “online harassment”, “cyber aggression”, and “toxicity”. We examine how scholars use them in both overlapping and disparate ways to refer to myriad hurtful and antisocial online behaviors. These behaviors can have dire and long-lasting effects on victims and online communities and demand researchers’ attention. However, without clarity, the terms operate as jargon rather than as devices that help scholars from multiple fields communicate with one another and with the public. We recognize that one problem for the terms is the translation from offline to online spaces, and discuss the inadequacies of the analogies between online and offline behaviors that fall under them. We also illustrate the problems this lack of specificity presents using example incidents from Instagram and Twitter. We then propose a new taxonomy of toxicity as a useful perspective for dealing with these behaviors. Our goal is to clarify terms to provide conceptual tools for scholars to theorize about, model, and reduce toxic Internet behaviors.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous; See http://acm.org/about/class/1998/ for the full list of ACM classifiers. This section is required.

Author Keywords

Authors’ choice; of terms; separated; by semicolons; include commas, within terms only; required.

INTRODUCTION

In recent years, toxic online behaviors have garnered increased attention from the media and have earned a permanent place in public discourse. In response to what is clearly a widespread problem [9] with potentially dire consequences [7, 3], computer-mediated communication (CMC) and computer science (CS) researchers have begun working toward technological interventions, often focusing on machine-learning approaches [20, 2, 5, 8, 18]. While many people across a multitude of fields in both research and industry are working toward curbing this problem, the precise nature of the problem remains unclear. Laypeople and reserachrs alike appear to hold a multitude of beliefs about what is and is not harassment, cyber bullying, or hurtful behavior, and whether harassment and cyber bullying are unique categories. Pater and colleagues [17] describe how, in industry, this lack of consensus is exemplified by the great differences between social-media platfroms in what is and is not considered appropriate behavior . Among the general public, the myriad of opinions is typified by the difference between what groups (e.g. teenagers and adults) consider to be cyber bullying [1] as differnces of opinion within groups (e.g. parents) [6].

It is probably fair to say that most of us have, in broad terms, a good idea of of what is or is not online harassment, cyber bullying, and/or cyber-aggression. But, what would you say if asked to provide detailed definitions for these terms? Being able to differentiate between malicious and benign online content depends not only on our own sense of where the lines are, but also on agreeing with eachother about when these lines are crossed. The most common method of training machine-learning classifiers to detect inappropriate online behavior relies on people like us to manually differentiate malicious content from benign content. The reliability of machine-learning classifiers is tied to the inter-rater reliability of the humans who provide the training data. Guberman and Hemphil [10] found that, even when provided with quite specific definitions for labelling content, human raters still differed in their interpretations for a variety reasons.

In this paper, we focus primarily on the varied definitions of the terms, as used by internet researchers. We highlight the usage of three main terms: “online harassment”, “cyber bullying”, and “cyber aggression”. Beyond the methodological implications of the strenght of our definitions, it is important to the research community as a whole that we understand what our peers are referring to when they use these terms. Often, these terms used without operational definitons, such that it is unclear whether researchers are talking about the same phenomena. In other cases, one of these terms will be used as an umbrella term encompassing the other two terms. There seems to be disagreement about which of these terms is the superset under which the other terms fit. Additionally, some researchers go to great lengths to justify the adaptation of traditional definitons of bullying to the various antisocial behaviors observed online, even when evidence suggests that these adaptated definitions do not adequately describe the online phenomena. We believe it is necessery to reconcile these differing perspectives and definitons in order to combat the problem most effectively. As a community, it is important that we agree about the nature of this problem, and that our definitions conform not only to the observed phenomena, but also to the types of behavior members of the public / users of various CMC platforms actaully find to be concerning. After a discussion of the existing usages of these terms, we propose a new taxonomy in which the superset is less value-laden than terms like cyber bullying. We suggest researchers focus on specific behaviors, which may vary in severity and required action depending on the contexts in which they occur.

ONLINE HARASSMENT

While reserach into human issues in computing, such as online harssment, are currently in vogue (and for good reason), academics have been investigating harassment in online spaces since the nascent era of modern CMC tools. In the later years of the internet-relay chat's popularity (IRC) , Herring [11] investigated the ways in which gender-based harassment manifested online. Herring operationalizes harassment using a definiton from *Black's Law Dictionary* as behaviors (either singularly or in repetition) "'which tended to annoy, alarm and verbally abuse"' individuals [p. 151-152]. Operationalized as such, Herring uses "harassment" to refer to a variety of behaviors falling under the umbrella of the definition provided. Among these behaviors are attempts to provoke, intimidate, and silence female IRC users.

Using "online harassment" as an umbrella category under which nearly all concivable negative online behaviors fit is a common occurence in the literature. Unlike Herring [11], however, who provides a clear definition for the operationalization of "online harassment" as a superset, others appear to define the superset by the behaviors it encompasses (or, viewed another way, do not define the superset at all). For example, in the Pew Research Center's 2014 report no online harrassment, Duggan does not explicityly define harassment itself [9]. Instead, Duggan refers to harassment as a spectrum that contains a range of (presumably negative) behaviors "from garden-variety name calling to more threatening behavior" [p. 1]. The specific harassing behaviors referred to in the report include offensive name-calling, physical threats, sustained harassment over a period of time, sexual harassment, and stalking. Reiterating the notian that these behaviors exist on a continuum of severity, Duggan explains that men endure more harrasment than women, but women are disproportianately more likely to be victimized by the more severe types of harassment.

Like Duggan [9], Lenhart and colleagues [15] also operationalize "online harassment" as both a superset and as a continuum. They provide one of the most explicitly inclusive definitions of online harrassment that we have seen:

> Harassment can encompass a wide range of unwanted contact that is used to create an intimidating, annoying, frightening, or even hostile environment for the victim. Online harassment is generally recognized as referring to this type of negative and unwanted contact using digital means. Online harassment can be a brief occurrence or a sustained campaign of abuse and attacks; the perpetrator (or perpetrators) might be intimately known to the victim, or a stranger in another state or country. Online harassment is defined less by the specific behavior than its intended effect on and the way it is experienced by its target.

They note that their operationalization differs from legal definitions, which require harrassment to be methodical and/or repetitive in nature. We would like to draw attention to the distinction at the end of their definition; to Lenhart and colleagues, whether or not a behavior constitutes online harassment is linked to the *inent* behind the behavior. By this rationale, it is unclear how one would classify behaviors by individuals who intend no harm to others, but which cause harm nonetheless. Turkle notes that, perhaps due to a stark decrease in empathy over several decades, there are instances in which adolescents may hurt their friends over CMC platforms with no recognition or understanding of how their behaviors were cruel [21].

In addition to providing several fairly specific harrassing behaviors (n=20, including the six behaviors reported on by [9]), Lenhart and colleagues [15] create a slightly more complex taxonomy by including three intermediate categories. Their 20 specifit types of harassment fit into the broad cateogories of *direct harassment*, *invasion of privacy*, and *denial of access*. While these three categories seem useful for discussing harassment, it remains to be seen whether they will have utility for machine-learning approaches to stemming the behaviors contained within.

Lenhart and colleagues' [15] definition of harassment is clearly stated and comprises specific behaviors which can be grouped together into convenient categories. However, it also includes some rather ambigous language. For example, early in the paper, Lenhart and colleagues define harassment and abuse as both consisting of "unwanted contact that is used to create an intimidating, annoying, frightening, or even hostile environment for the fictim and that uses digital means to reach the target" [p. 3]. Being that the two terms appear to be synonymous, their use in the report's title, *Online Harassment, Areigital Abuse, and Cyberstalking in America* is confusing. The title seems to indicate that online harassment, digital abuse, and cyberstalking are three different, but related, things. Throughout the paper, the authors refer to both "harassment and abuse" and "harassment or abuse," such that it is further unclear whether they intend for these terms to be used interchangably, or whether there is some subtle, yet unstated, difference. Additionally unclear is the rationale for the elevation of *cyberstalking* to a seperate entry in the paper's title, as it is referred to as a type of harassment (not something distinct from harassment and/or abuse) throughout the paper itself [1]. The comprehensiveness of Lenhart and colleagues' reported results, which are quite elucidating and build upon those of Duggan [9], is tampered by the ambiguity injected by the loose usage of these terms.

[1] In addition to being used in a somewhat confusing manner, there is a question of whether or not *cyberstalking*, a which seems to imply something set apart from traditional stalking, is a useful term. Some research shows that cyberstalking is much more of an outgrowth of traditional stalking, rather than a seperate entity , and referring to it as something seperate may obfuscate the severity of incidents by directing attention away from whatever stalkign may simultaneously be hapenning to a victim offline [19].

In contrast to Duggan [9], Lenhart and collagues [15], and Lenhart [14] who consider cyberbullying to be a type of online harrasment, as well as in contrast to Lenhart [13] and Jones and colleagues [12] who use the two terms interchangeably, Wolak and colleagues [22] ask whehter online harassment is a manifestation of cyberbullying. That is, whereas online harassment is typically operationalized as the superset into which cyberbullying fits, Wolak and colleagues wonder if the reverse might be true. We will delve into their results and into their usage of the term "cyberbullying" later. They defined online harassment as any threats or offensive behaviors, non-inclusive of sexual solicitation, sent to or posted publicly about a child or adolescent. Implied in this definiton, although not as clearly made explicit as in the previous examples, is a taxonomy of behaviors in which "harassment" is the umbrella term. This operationalization is quite inclusive, albeit not as clear as that of Lenhart and colleagues [15], and only inclusive of behaviors targeted towards minors. While the paper and its venue of publication are clearly focused on harassment as it pertains to minors, this particular definition ignores the fact that adults are also quite likely to be victims of online harassment [9, 15], making the definition potentially misleading.

The examples discussed above are indicative of the ways in which researchers of the nature and reach of hurtful online behaviors talk about these phenomena. Online harassment is generally a superset under which a myriad of behaviors fall. Sometimes the behaviors are clustered into groups, and sometimes they're ranked by severity. As researchers focus in on ever more specific types of harassment, definitions become more precise (and arguably, useful for discussing the behaviors themselves). Just as Pater has found that different social-media platforms consider different behaviors to count as harassment [17], so, too, do researchers. Some researchers focus on singular, specific behaviors that comprise harassment [16]. There is, perhaps, a need for more of such papers to tell us about the more granular details of what online harassment looks like. Interestingly, we found only a few instances in which self-harm was categorized as harassing behavior. One of these instances was Pater [17], who included self-harm in her investigation of social media platforms' harassment policies. The other was Boyd [1], who noted that some degree of online harassment is perpetrated by adolescents against themselves, perhaps for attention or validation.

While research geared toward understanding harassment uses the vocabulary described above, research geared toward technical solutions to the problem of harassment has a different lexicon. These studies often use terms other than online harassment. When online harassment *is* used, it is rarely operationalized at all. Instead, it refers to a set of features of online content presumed to be undesirable. For example, Dadvar and Jong [4] talk only about detecting "harrassing sentences." What are harassing sentences? In this case, a sentence is harassing depending on whehter it contains profanity, the pronouns employed, and the gender of the person that posted the content. They use cyberbullying interchangeably. This approach, in contrast to the definitions described earlier, is highly exclusionary and likely to lead to a lot of mis-labeling of content [10]. In another early case of applying machine-learning to online harassment, Yin and colleagues [23] first acknowledge the ambiguity surrounding the term "online harassment" before defining it as any intentional action meant to annoy another user in the community. Beyond that, they zero-in on a specific class of harassing behavior "in which a user systematically deprecates the contributions of another user" [p. 2]. This definition is inclusive, and the specific behavior is on which Herring [11] also identified. While the definition they use is reasonable, the way it is operationalized in their classifiers is potentially problematic. While they include contextual and sentiment features, the reliance on profanity is an unreliable indicator of harassment (as they themselves reported).

CYBER BULLYING

There appears to be a good deal of overlap between the ways in which researchers employ the terms "online harassment" and "cyberbullying." In fact, there are researchers who create the exact types of taxonomies described in the previous section referring to the superset as "cyberbullying" rather than as "online harassment" (or use the two terms completely interchangably) [13]. More frequently, however, at least in the space in which researchers are focusing on the nature, reach, and impact of cyberbullying, definitions of the term tend to be less inclusive than those used in similar situations to describe online harassment. The definitons of cyberbullying are more numerous and more varied. In one case, in introducing the idea of cyberbullying, a single paper draws upon more than five seperate definitions. They consider bullying to be the use of ICTS to maliciously and repeatedly threaten people, threats sent via ICTs that cause psychological and social problems for victims, a type of psychological bullying occuring via ICTs, a form of social aggression, *and* an extension of traditional bullying with several notable exceptions [24]. Unfortunately, not all of the definitions the authors combined are compatable with one another. In this section, we will discuss some of the myriad definitions of cyberbullying, definitions of cyberbullying based on traditional bullying and the problems thereof, and the ways in which different research communities are talking about cyberbullying in quite different ways.

Among those trying to figure out just what cyberbullying use, not including the definitions that are synonymous with the taxonomies of online harassment, two types definitions of cyberbullying prevail.

CYBER AGGRESSION

TOXICITY

MOVING FORWARD

REFERENCES

1. Danah Boyd. 2014. Bullying: Is Social Media Amplifying Meanness and Cruelty?. In *It's complicated: the social lives of networked teens*. Yale University Press, New Haven, 128–152.

2. Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing*

(SocialCom). IEEE, 71–80. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6406271

3. Noah Cohen and MaryAnn Spoto. 2015. Transgender N.J. Game Developer Jumps from GWB After Online Bullying. *NJ.com* (2015). http://www.nj.com/monmouth/index.ssf/2015/04/post_18.html

4. Maral Dadvar and Franciska de Jong. 2012. Cyberbullying Detection: A Step Toward a Safer Internet Yard. In *Proceedings of the 21st International Conference on World Wide Web (WWW '12 Companion)*. ACM, New York, NY, USA, 121–126. DOI: http://dx.doi.org/10.1145/2187980.2187995

5. Maral Dadvar, Roeland Ordelman, Franciska de Jong, and Dolf Trieschnigg. 2012. Towards User Modelling in the Combat against Cyberbullying. In *Natural Language Processing and Information Systems (Lecture Notes in Computer Science)*, Gosse Bouma, Ashwin Ittoo, Elisabeth MÃl'tais, and Hans Wortmann (Eds.). Springer Berlin Heidelberg, 277–283.

6. Matthew Davis, Sarah Clark, Dianne Singer, Amilcar Matos-Moreno, and Anna Daly Kauffman. 2015. Parents conflicted about how to label, punish cyberbullying. *C.S. Mott Children's Hostpital National Poll on Children's Health* 24, 4 (2015), Online. http://mottnpch.org/reports-surveys/parents-conflicted-about-how-label-punish-cyberbullying

7. Michelle Dean. 2012. The Story of Amanda Todd. *The New Yorker* (2012). http://www.newyorker.com/culture/culture-desk/the-story-of-amanda-todd

8. Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of Textual Cyberbullying.. In *The Social Mobile Web*. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/download/3841Karthik/4384

9. Maeve Duggan, Lee Rainie, Aaron Smith, Cary Funk, Amanda Lenhart, and Mary Madden. 2014. *Online Harassment*. Technical Report. Pew Research Center. 1–64 pages. http://www.pewinternet.org/2014/10/22/online-harassment/

10. Joshua Guberman and Libby Hemphill. 2017. Challenges in Modifying Existing Scales for Detecting Harassment in Individual Tweets. DOI: http://dx.doi.org/10.24251/HICSS.2017.267

11. Susan C. Herring. 1999. The Rhetorical Dynamics of Gender Harassment On-Line. *The Information Society* 15 (1999), 151–167.

12. Lisa M. Jones, Kimberly J. Mitchell, and David Finkelhor. 2013. Online harassment in context: Trends from three Youth Internet Safety Surveys (2000, 2005, 2010). *Psychology of Violence* 3, 1 (2013), 53–69. DOI: http://dx.doi.org/10.1037/a0030309

13. Amanda Lenhart. 2007. *Cyberbullying*. Technical Report. http://www.pewinternet.org/2007/06/27/cyberbullying/

14. Amanda Lenhart. 2010. Cyberbullying 2010: What the Research Tells Us. (2010). http://www.pewinternet.org/2010/05/06/cyberbullying-2010-what-the-research-tells-us/

15. Amanda Lenhart, Michele Ybarra, Kathryn Zickuhr, and Myeshia Price-Feeney. 2016. *Online Harrassment, Digital Abuse, and Cyberstalking in America*. Technical Report. Data & Society Research Institute and the Center for Innovative Public Health Research. 1–59 pages. OnlineHarassment,DigitalAbuse, andCyberstalkinginAmerica

16. Peter J. Moor, Ard Heuvelman, and Ria Verleur. 2010. Flaming on YouTube. *Computers in Human Behavior* 26, 6 (2010), 1536–1546. DOI: http://dx.doi.org/10.1016/j.chb.2010.05.023

17. Jessica A. Pater, Moon K. Kim, Elizabeth D. Mynatt, and Casey Fiesler. 2016. Characterizations of Online Harassment: Comparing Policies Across Social Media Platforms. ACM Press, 369–374. DOI: http://dx.doi.org/10.1145/2957276.2957297

18. Kelly Reynolds, April Kontostathis, and Lynne Edwards. 2011. Using machine learning to detect cyberbullying. In *Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on*, Vol. 2. IEEE, 241–244. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6147681

19. L. P. Sheridan and T. Grant. 2007. Is cyberstalking different? *Psychology, Crime & Law* 13, 6 (2007), 627–640. DOI: http://dx.doi.org/10.1080/10683160701340528

20. Sara Owsley Sood, Elizabeth F. Churchill, and Judd Antin. 2012. Automatic identification of personal insults on social news sites. *Journal of the American Society for Information Science and Technology* 63, 2 (2012), 270–285. http://onlinelibrary.wiley.com/doi/10.1002/asi.21690/full

21. Sherry Turkle. 2015. *Reclaiming Conversation: The Power of Talk in a Digital Age* (1st edition ed.). Penguin Press, New York.

22. Janis Wolak, Kimberly J Mitchell, and David Finkelhor. 2007. Does online harassment constitute bullying? An exploration of online harassment by known peers and online-only contacts. *J. Adolesc. Health* 41, 6 Suppl 1 (2007), S51–8.

23. 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.

24. Bayram ÃĞetin, Erkan Yaman, and Adem Peker. 2011. Cyber victim and bullying scale: A study of validity and reliability. *Computers & Education* 57, 4 (2011), 2261–2271. DOI: http://dx.doi.org/10.1016/j.compedu.2011.06.014