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 [8] with potentially dire consequences [6, 3], computer-mediated communication (CMC) and computer science (CS) researchers have begun working toward technological interventions, often focusing on machine-learning approaches [12, 2, 4, 7, 11]. 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 [10] 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) [5].

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 [9] 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

CYBER BULLYING

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, 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 MÃľtais, and Hans Wortmann (Eds.). Springer Berlin Heidelberg, 277–283.

5. 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

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

7. 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

8. Maeve Duggan. 2014. *Online Harassment*. Technical Report. Pew Research Center.

9. Joshua Guberman and Libby Hemphill. 2017. Challenges in Modifying Existing Scales for Detecting Harassment in Individual Tweets. In *50th Annual Hawaii International Conference on System Sciences (HICSS-50).*

10. 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

11. 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

12. 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