Mitigating Harassment in Online Communities with Human-Bot Moderation: Insights from Reddit Communities

An Nguyen[*] Arun Rai[†] Likoebe Maruping[‡]

This paper sets out to explore the duality of human-bot moderation in mitigating harassment. We examine communities that use a block-list type of bot to prevent harassment from the source of harassment. We expect that the employment of the bot alongside human moderation will create a shielding effect - a declining trend of harassment towards community members, followed by an emboldening effect - an uptick in harassment by community members towards their perceived outgroup members, and, finally, a spillover effect - an increase in harassment in neighbor communities who share the same topic of discussion but not the same moderators. We use Detoxify, a BERT-based model, to determine the probability that a comment represents harassment. . We then use Bayesian Structure Time Series to examine the three types of effects of human-bot moderation in the communities under study

Introduction

The need to converse emerged since the dawn of human civilization. In this Internet age, that need for face-to-face conversation has evolved into the need for online conversation, which ranges from decade-old Internet Relay Chat and forums to today's newsgroup and social networks. Not all conversations are civilized conversations. In online settings, because of online anonymity, people may engage in even more harmful activities Duggan (2021b), such as harassment and cyberbullying. Harassment does not only negatively affects individuals who participate in online communities but also harms said communities as a whole.

[*] Georgia State University, anguyen192@gsu.edu

[†] Georgia State University, arunrai@gsu.edu

[‡] Georgia State University, lmaruping@gsu.edu

In a recent survey by Pew Research (Vogels 2021) , 41% of American states that they have personally experienced some forms of harassment, while 25% of those surveyed stated that they even received threats, stalking, sexual and sustained harassment. Although the consequences of online harassment on individuals are more prominent in the short term, a few people, especially women who experienced it, reported that they suffered negative long-term impact (Duggan 2021a). More than that, harassment also poses great threats for the community as a whole, ranging from the quality erosion of content to exclusion of certain groups of contributors. A prominent example of this is 8chan, a loosely moderated image-board site where users can post images annonymously about anime, popular culture, to politics and sports (Wikipedia n.d.). From its inception in 2013, the site was gradually dominated by a group of users with extreme ideologies, who drove away other genuninely interested contributors along with a variety of other topics. The site then turned into a home for antisemitism, misogyny, and anti-immigration ideologies. It was associated with a series of mass shooting in 2019: the New Zealand mass shooting at two mosques (Regan and Sidhu 2019), El Paso mass shooting at Walmart (Mezzofiore and OŚullivan 2021), and Dayton, Ohio mass shooting (Paul P. Murphy and Levenson 2019). Consequently, the site was shut down by its network infrastructure provider (Cloudfare 2019), web service (Robertson 2019). Such incidents along with other related concerns pressure platforms to mainly focus on addressing harassment.

Most platforms used human moderation as the immediate solution when they first encountered online toxicity. However, moderation by just human ran into several issues for platforms owner. *First,* as platform expands, governing with just human moderation as traditionally done is costly to scale up, considering how vast platforms have become over the last decades. Gillespie (2018). *Second,* a number of research and reports show that human moderators, no matter paid or volunteer, experience burnout and emotional distress Snyder (2020). Along with that, human moderators cannot be solely relied on in case of emergency. During the Covid-19 pandemic, platforms sent their human moderators home while starting to automate the moderating process Lapowsky (2020). . *Last but not least,* although human judgement is always used as a standard, it known to have biases, especially towards delicate matter such as political ideology Diakopoulos and Naaman (2011). Thus, the future of moderation cannot depend on human actions alone.

Preparing themselves for the future of moderation, platform, and community owners shifted from using mainly humans to relying mostly on machines for moderation tasks, reasoning that machine is faster at scale Gorwa, Binns, and Katzenbach (2020). For example, Wikipedia implemented a wide range of bots to automate tasks on each Wiki page, one of which is the ClueBot NG bot. This bot claims to detect whether an edit is an act of vandalism (Wikipedia 2021). Facebook also used AI intensively in its online moderation ranging from hate speech to misinformation detection (Schroepfer 2021). Consistency is also a good virtues that machine brings to moderation tasks. In the case of Covid-19 pandemic, technology companies boasted their innitial results of switching to machine moderation. In a report, Facebook stated its independence of human moderators stating that 95% of the hate speech they have taken down was performed by Artificial Intelligence (Schroepfer 2021) despite several criticisms for dismissing its human moderators Stokel-Walker (2020). Despite its rising popularity in recent years, moderation relying solely on machines ignites a whole new level of concern given the inherently complicated landscape of moderation. *First*, as with many other tasks being automated, there is a burning question about whether machines can totally replace humans in such delicate matter. In this vein of discussion, Gillespie (2020) argues that although machine-based moderation is inevitable, humans must remain in the loop. The bias could come from the training datasets as proven in Binns et al. (2017). Gorwa, Binns, and Katzenbach (2020) also raised the same concern by arguing that algorithmic moderation could create injustice in large-scale socio-technological systems. *Secondly*, in a similar vein with Gorwa, Binns, and Katzenbach (2020), Gillespie (2020) concerns about transparency and accountability should there be no human in the loop. *Moreover*, the platform's point of view about the trade between free speech and safety, Gorwa, Binns, and Katzenbach (2020) pointed out that purely algorithmic moderation would undermine the political nature of speech. *Last but not least,* Mark Zuckerberg admitted himself that machine is not sensitive to "nuances" in languages or the intent behind the comment yet, which inevitably can lead to misclassification (Canales).

Past discussions on platforms' governance suggest that neither relying on solely humans nor machines works effectively for online moderation against harassment. Thus, human-machine moderation is a viable solution for the future of platform governance. More recent discussions, research included, on online moderation shifted their attention to human-machine moderation (Chandrasekharan, Gandhi, Mustelier, & Gilbert, 2019; Jhaver, Birman, Gilbert, & Bruckman, 2019; Kiene & Hill, 2020; Kiene, Jiang, & Hill, 2019). Through-out these studies, Reddit, Discord, and Twitch stand out as the most studied platforms for human-machine moderation. These platforms allow customized moderation at the community level: each community is run by a team of moderators, human and machine included. As a result, this decentralized governance invites a variety of governing modes across different communities. Kiene, Jiang, and Hill (2019) and Kiene and Hill (2020) studied the successful use of bot moderators when human moderators faced an exploded amount of content on Discord and Reddit. Seering et al. (2018) discovered that the moderation bot also played a social role in facilitating discussion on Twitch. Not only does the literature explore the overall effect of human-bot moderation, but it also dives into specific bots' effects. Chandrasekharan et al. (2019) claimed to study the first "open source, AI-backed socio-technological moderation systems" - the Crossmod. Although the study confirmed the bot's superior performance judging from human moderators' positive feedback, in reality, it is not used as much as other bots on Reddit. As of July 2022, the bot seems to cease operation, judging from the two communities it is monitoring. The mixed results of human-machine's moderation in research could be due to the fact that not all human-bot moderation is executed the exact same way.

As the literature suggests (Gillespie 2020), human-in-the-loop moderation is a better fit for moderation at community level as opposed to human-out-of-the-loop moderation. One prominent example of the two mechanism is the Crossmod bot (Chandrasekharan et al. 2019) and the AutoModerator bot on Reddit. Automod was developed independently in 2012, then was officially adopted as Reddit's official tool in 2015. From its grassroots popularity, Automod rose to become the only platform-incorporated machine moderator as well as the most adopted bot across all communities. The authors highly believe that the stark contrast between the performance of Crossmod and Automod is due to the configuration of human-machine collaboration. While Auto¬mod includes human in many of its actions, human-in-the-loop mechanism, Crossmod only involves human's judgement at the beginning and the end of the process, human-out-of-the-loop mechanism.

Second paragraph should focus on the human-in-the-loop versus human-out-of-the-loop distinction. It should also make clear that we focus our interest on the human-in-the-loop approach.

Consistent with the example and what have been theorized in the literature, we argue that human-in-the-loop mechanism works best for collaboration, particularly in the case of community moderation. Adding on to this growing human-bot moderation literature, we aim to explore one type of bot that has grown in popularity on such platforms. Specifically, we studied a block-list anti-harassment bot with human-in-the-loop mechanism. We expect that this human-in-the-loop mechanism will help ease the collaboration between bot and human, which eventually strengthens the good sides of solely human moderation. However, the machine's block-list feature would work so effectively that the bot will lead the community to radicalism, with little to no space for civilized discussions.

More than that, we do not only examine the effect of the human-bot moderation on other communties. Since most platforms host a series of communities (e.g., Facebook, Reddit, Discord, etc.), at least some communities within a platform may discuss the same issues. As the literature on moderation provide a strong patterns of spillover effect when moderation measurements are in place Jhaver et al. (2021), we suggest that the introduction of bot to one community's moderator team may affect other communities with similar topics. The literature also witnesses a strong pattern of spillover effect

My main concern with these research questions is that the preceding introduction material does not really set them up. Somewhere in this introduction we need to motivate the need to understand the different types of effects—within and outside of the harassed community—that can result from human-bot moderation. (our answer, of course, is shielding, emboldening and spillover). One possibility is that we devote the second paragraph of the intro to the following points: -There is increasing recognition that neither human moderation nor bot moderation alone are sufficient to combat harassment in online communities; -Given their complementary strengths, human-bot moderation is fast becoming the approach of choice; -As interest in the utilization of human-bot moderation is accelerating, there is an opportunity to learn what effects it is having inside and outside of the communities that are the target of harassment.

Theoretical Background

Toxicity and Harassmnet

Social Catergorization Lense

Source, Target, and Acts of Harassment

Focal Community and Neighbor Community

Human Machine Moderation

Human-out-of-the-loop versus Human-in-the-loop Moderation

AI-based Moderation

Rule-based Moderation

Moderation Strategy

Hypotheses

Shielding Effects of Human-bot Moderation

Emboldening Effects of Human-bot Moderation

Spillover Effects of Human-bot Moderation

Empirical Analysis

Context

Reddit and its moderation policies

The bot, *r/saferbot* and the community, *r/femaledatingstrategy*

Data Collection, Measurement, and Research Design

Analysis

Discussion and Future Research

Ali, Shiza, Mohammad Hammas Saeed, Esraa Aldreabi, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. 2021. "Understanding the Effect of Deplatforming on Social Networks." In *13th ACM Web Science Conference 2021*. ACM. https://doi.org/10.1145/3447535.3462637.

Binns, Reuben, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. "Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation." In *International Conference on Social Informatics*, 405–15. Springer.

Canales, Katie. "Mark Zuckerberg Said Content Moderation Requires 'Nuances' That Consider the Intent Behind a Post, but Also Highlighted Facebook's Reliance on AI to Do That Job." *Business Insider*. https://www.businessinsider.com/zuckerberg-nuances-content-moderation-ai-misinformation-hearing-2021-3.

Chandrasekharan, Eshwar, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. "Crossmod: A Cross-Community Learning-Based System to Assist Reddit Moderators." *Proceedings of the ACM on Human-Computer Interaction* 3 (CSCW): 1–30.

Chandrasekharan, Eshwar, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. "You Can't Stay Here." *Proceedings of the ACM on Human-Computer Interaction* 1 (CSCW): 1–22. https://doi.org/10.1145/3134666.

Cloudfare. 2019. "Terminating Service for 8Chan." https://blog.cloudflare.com/terminating-service-for-8chan/.

Diakopoulos, Nicholas, and Mor Naaman. 2011. "Towards Quality Discourse in Online News Comments." In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, 133–42.

Dosono, Bryan, and Bryan Semaan. 2019. "Moderation Practices as Emotional Labor in Sustaining Online Communities: The Case of AAPI Identity Work on Reddit." In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13.

Duggan, Maeve. 2021a. "Part 4: The Aftermath of Online Harassment." *Pew Research Center*. https://www.pewresearch.org/internet/2014/10/22/part-4-the-aftermath-of-online-harassment/.

———. 2021b. "The Broader Context of Online Harassment." *Pew Research Center*. https://www.pewresearch.org/internet/2017/07/11/the-broader-context-of-online-harassment/.

Dwoskin, Elizabeth, and Nitasha Tiku. 2020. "Facebook Sent Home Thousands of Human Moderators Due to the Coronavirus. Now the Algorithms Are in Charge." *Washington Post*. https://www.washingtonpost.com/technology/2020/03/23/facebook-moderators-coronavirus/.

Geigner, Timothy. 2021. "Facebook AI Moderation Continues to Suck Because Moderation at Scale Is Impossible." *Tech Dirt*. https://www.techdirt.com/2021/10/20/facebook-ai-moderation-continues-to-suck-because-moderation-scale-is-impossible/.

Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, Content Moderation, and* *the Hidden Decisions That Shape Social Media.* Yale University Press.

———. 2020. "Content Moderation, AI, and the Question of Scale." *Big Data & Society* 7 (2): 2053951720943234.

Gorwa, Robert, Reuben Binns, and Christian Katzenbach. 2020. "Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance." *Big Data & Society* 7 (1): 2053951719897945.

Jhaver, Shagun, Christian Boylston, Diyi Yang, and Amy Bruckman. 2021. "Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter." *Proceedings of the ACM on Human-Computer Interaction* 5 (CSCW2): 1–30. https://doi.org/10.1145/3479525.

Kiene, Charles, and Benjamin Mako Hill. 2020. "Who Uses Bots? A Statistical Analysis of Bot Usage in Moderation Teams." In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–8.

Kiene, Charles, Jialun Aaron Jiang, and Benjamin Mako Hill. 2019. "Technological Frames and User Innovation: Exploring Technological Change in Community Moderation Teams." *Proceedings of the ACM on Human-Computer Interaction* 3 (CSCW): 1–23.

Lapowsky, Isse. 2020. "After Sending Content Moderators Home, YouTube Doubled Its Video Removals." *Protocol*. https://www.protocol.com/youtube-content-moderation-covid-19.

Mezzofiore, Gianluca, and Donie OŚullivan. 2021. "Part 4: The Aftermath of Online Harassment." *Pew Research Center*. https://www.cnn.com/2019/08/04/business/el-paso-shooting-8chan-biz.

Paul P. Murphy, Drew Griffin, Konstantin Toropin, and Eric Levenson. 2019. "Dayton Shooter Had an Obsession with Violence and Mass Shootings, Police Say." https://www.theguardian.com/technology/2019/aug/04/mass-shootings-el-paso-texas-dayton-ohio-8chan-far-right-website.

Regan, Helen, and Sandi Sidhu. 2019. "49 Killed in Mass Shooting at Two Mosques in Christchurch, New Zealand." *CNN*. https://edition.cnn.com/2019/03/14/asia/christchurch-mosque-shooting-intl/index.html.

Robertson, Adi. 2019. "8chan Goes Dark After Hardware Provider Discontinues Service." https://www.theverge.com/2019/8/5/20754943/8chan-epik-offline-voxility-service-cutoff-hate-speech-ban.

Schroepfer, Mike. 2021. "Update on Our Progress on AI and Hate Speech Detection." https://about.fb.com/news/2021/02/update-on-our-progress-on-ai-and-hate-speech-detection/.

Seering, Joseph, Juan Pablo Flores, Saiph Savage, and Jessica Hammer. 2018. "The Social Roles of Bots: Evaluating Impact of Bots on Discussions in Online Communities." *Proceedings of the ACM on Human-Computer Interaction* 2 (CSCW): 1–29.

Snyder, Kristen. 2020. "TikTok Content Moderators Allege Emotional Distress." https://dot.la/tiktok-content-moderators-2657593810.html.

Stokel-Walker, Chris. 2020. "As Humans Go Home, Facebook and YouTube Face a Coronavirus Crisis." *WIRED*. https://www.wired.co.uk/article/coronavirus-facts-moderators-facebook-youtube.

Vogels, Emily. 2021. "The State of Online Harassment." *Pew Research Center*. https:

//www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/.

Wikipedia. n.d. "8chan," n.d. https://en.wikipedia.org/wiki/8chan.

———. 2021. "User:ClueBot NG." https://en.wikipedia.org/wiki/User:ClueBot_NG.

Yang, Yukun. 2019. "When Power Goes Wild Online: How Did a Voluntary Moderator's Abuse of Power Affect an Online Community?" *Proceedings of the Association for Information Science and Technology* 56.