Statistical Modeling of Harassment against Reddit Moderators

Hind Almerekhi
Hamad Bin Khalifa University
Doha, Qatar
hialmerekhi@mail.hbku.edu.qa

Haewoon Kwak
Qatar Computing Research Institute, HBKU
Doha, Qatar
haewoon@acm.org

Bernard J. Jansen
Qatar Computing Research Institute, HBKU
Doha, Qatar
jjansen@acm.org

ABSTRACT

Despite the dedication that some volunteer moderators of online communities display when performing their moderation duties, they become targets of hate and harassment by other users. To understand what causes the change in moderator role from heroes to victims, we analyze the responses of 1,818 moderators on Reddit to an online survey about moderation practices and harassment. We built a statistical model and found 6 significant independent variables that affect harassment on moderators, such as the knowledge of community norms, which increases harassment on moderators the most. Our findings imply that vulnerable moderators in toxic communities need countermeasures against harassment.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in collaborative and social computing**.

KEYWORDS

Online communities, Reddit, moderation, toxicity, survey

ACM Reference Format:
Hind Almerekhi, Haewoon Kwak, and Bernard J. Jansen. 2020. Statistical Modeling of Harassment against Reddit Moderators. In *Companion Proceedings of the Web Conference 2020 (WWW '20 Companion), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3366424.3382729

1 INTRODUCTION

Discussion platforms like Reddit offer millions of users with a platform that allows them to express their opinions around various topics, making them rich sources of data for a variety of social insights [1]. Reddit, which is famous for being "the front page of the internet" enables users to build and engage in communities (i.e., subreddits) to exchange different types of content. While Reddit, amongst other platforms, promises users the complete freedom to express their interests through communities in a safe manner, users often get exposed to toxic posts that include hate and incivility [5]. Hence, online discussion platforms recruit paid or volunteer moderators to manage the communities and monitor the activities of users.

The main job of a content moderator is to screen the postings of users to discard malicious content and penalize users that post such content [3]. The duties of moderators evolved with the increase in communities and their size to include ensuring the growth of communities, revising guidelines, and recruiting more moderators if needed. To encourage positive interactions between users in communities, moderators should protect users by preventing the spread of toxicity and hatefulness. Moderators of popular communities find that achieving this task is extremely difficult. As different communities have different norms and values, moderators' practices must be different across communities, which means that the user's response to moderators is different as well.

Some users respond aggressively to moderators, especially when they delete user's posts or comments, leading to targeted harassment by users towards moderators. Since users behave differently across communities, this implies that the level of harassment against the moderators can also be different. In addition to community-level differences, moderation style largely influences the user perception of moderation transparency [6]. We thus hypothesize that the moderation practices of a moderator influence the level of the harassment against the moderator. We build an Ordinary Least Squares (OLS) regression model to explain the level of the harassment using multiple variables of moderation practices.

2 RELATED WORK

One of the main problems that moderators struggle with is the prevention of hateful and abusive behavior in online communities [2]. The struggle often arises when the values of moderators and users clash. Moderation style and practices always govern how users perceive moderator actions. The goal with any moderation task is to support the user's freedom of speech without limiting their expressiveness. This goal correlates with the moderation style and practices of moderators [8]. Generally, applying stricter moderation rules negatively affects the user experience and their ability to enjoy engaging in discussions online positively. As a solution, some prior studies suggest that providing feedback to users in online platforms [4] and designing systems that support high-quality comments [7] can benefit volunteer moderators and users alike in fostering a positive online experience. In this work, we stir from the norm and focus on identifying the moderation practices that lead moderators to be harassed by users. By focusing on this problem, we aim to identify the point where moderators transform from heroes to victims (or perhaps villains) from the perspective of users.

3 SURVEY DESIGN

To create the survey, we focused on the aim of the study and designed questions that targeted getting responses from moderators regarding their moderation practices on Reddit and how it relates to toxicity and harassment. We organized our survey into three

WWW '20 Companion, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7024-0/20/04.
https://doi.org/10.1145/3366424.3382729

main sections: (A) general moderation questions, (B) moderation challenges questions, and (C) moderation and toxicity questions. In section (A), we asked the moderators if they are using any automatic moderation tools (Tools Usage) and how helpful the tools are with their work (Tools Helpfulness). In section (B), we asked moderators if they ever been criticized or harassed based on moderation choices and the importance of preventing toxicity (importance Toxicity). We also asked moderators if harassment and hate on moderators can be prevented (harassment Mods Stop). As for moderation rules, we asked moderators how well they know the rules (Knowledge Of Rules), the fairness of the rules (Rules Fairness), and if they apply them strictly or loosely (Strictness Of Rules). As for section (C), we asked the moderators about the frequency of users violating moderation rules (Users Violation) and if they are moderating multiple subreddits (Mod Sub Count). Additionally, we asked respondents a couple of demographic questions about their gender and age group to gauge the diversity of responses. Lastly, we randomly sampled about 7,000 potential survey respondents and ensured that the potential participants had varying experiences on Reddit. The response rate of the 7,000 moderators was 26.7%.

4 MODELING HARASSMENT AGAINST MODERATORS

After cleaning the responses for analysis, we found a total of 1,818 usable responses for the study (88.77% males, 12.43% females, and 3.80% other genders). We also found that 77.72% of the respondents were between 18 and 34 years old. First, we converted the textual responses (e.g. gender) to numerical values for the regression analysis. Then, we build an OLS model to explain the influences of the practices of a moderator, which are represented as multiple independent variables, on the level of harassment against the moderator as the dependent variable. The results of the fitted regression model are in Table 1.

Table 1: Effect of independent variables on the harassment of moderators.

	coef.	std. err.	*p*-value
importance Toxicity	0.0824	0.024	***
harassment Mods Stop	-0.1416	0.022	***
Knowledge Of Rules	0.2441	0.034	***
Rules Fairness	-0.0037	0.049	
Strictness Of Rules	-0.0266	0.028	
Users Violation	0.1827	0.029	***
Rate Of Sub Toxicity	0.2365	0.029	
Mod Sub Count	0.0266	0.023	
Toxicity Similarity	-0.0132	0.021	
Tools Usage	0.0801	0.051	
Tools Helpfulness	0.1104	0.038	**
Gender	-0.0474	0.027	*
Age	0.1257	0.032	***

R-squared	0.741

Note: *** : p < 0.01, ** : p < 0.05, *: p < 0.10.

We find six significant variables ($p < .05$) in Table 1. While only one variable, (harassment Mods Stop), decreases the level of harassment against a moderator, other variables increase the harassment. Several interesting trends emerge in the model. First, if moderators work for more toxic communities (Users Violation), they are likely to be harassed. While this is not unexpected, considering that moderators are mainly volunteers, this calls for special attention to the moderators of toxic communities that are more vulnerable to harassment. This finding is particularly important because moderators who think that harassment can be prevented less (harassment Mods Stop) are more likely to be harassed. It might be a sign of repetitive exposure to harassment of unprotected moderators. Second, if moderators are well aware of rules (knowledge Of Rules) and realize the usefulness of the automatic tools (Tools Helpfulness), they are likely to be harassed. In other words, the level of involvement in the community exposes moderators to more harassment. This involvement includes knowing the rules and perhaps overestimating the usefulness of moderation tools. Third, if moderators think that preventing toxicity is important (importance Toxicity), moderators are likely to be harassed. This finding is reasonable because if moderators value preventing toxicity, they are most likely victims of toxicity and harassment. Last but not least, older moderators (Age) are more susceptible to harassment. Perhaps surprisingly, (Gender) was only a weak predictor of moderator harassment.

5 CONCLUDING REMARKS

In this work, we built a regression model based on the responses of moderators to study the correlation between moderation practices and moderator harassment. Our findings show that moderators get harassed when they think that automation tools help them with their work and when the number of users that violate community rules increases. The results show that the starting point of targeting the harassment towards moderators starts with the increase in rule violation and knowledge of community rules.

REFERENCES

[1] Hind Almerekhi, Haewoon Kwak, Joni Salminen, and Bernard J. Jansen. 2020. Are These Comments Triggering? Predicting Triggers of Toxicity in Online Discussions. In *Proceedings of the 29th International Conference on World Wide Web*.

[2] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 3175–3187.

[3] Bryan Dosono 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*. ACM, Article 142, 13 pages.

[4] Bernd Huber, Katharina Reinecke, and Krzysztof Z. Gajos. 2017. The Effect of Performance Feedback on Social Media Sharing at Volunteer-Based Online Experiment Platforms. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 1882–1886.

[5] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Trans. Comput.-Hum. Interact.* 26, 5, Article 31 (July 2019), 35 pages.

[6] Prerna Juneja, Deepika Rama Subramanian, and Tanushree Mitra. 2020. Through the Looking Glass: Study of Transparency in Reddit's Moderation Practices. *Proc. ACM Hum.-Comput. Interact.* 4, Article Article 17 (Jan. 2020), 35 pages.

[7] Deokgun Park, Simranjit Sachar, Nicholas Diakopoulos, and Niklas Elmqvist. 2016. Supporting Comment Moderators in Identifying High Quality Online News Comments. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 1114–1125.

[8] Simon T. Perrault and Weiyu Zhang. 2019. Effects of Moderation and Opinion Heterogeneity on Attitude Towards the Online Deliberation Experience. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM.