Characterizing Community Guidelines on Social Media Platforms

Jialun "Aaron" Jiang
University of Colorado Boulder
Boulder, CO
aaron.jiang@colorado.edu

Skyler Middler
University of Colorado Boulder
Boulder, CO
skyler.middler@colorado.edu

Jed R. Brubaker
University of Colorado Boulder
Boulder, CO
jed.brubaker@colorado.edu

Casey Fiesler
University of Colorado Boulder
Boulder, CO
casey.fiesler@colorado.edu

ABSTRACT

Social media platforms use community guidelines to enact governance and moderate content, but the limitation in their moderation capacity forces them to choose the types of misbehavior they focus more on. In this work, we analyze these choices through a content analysis of the community guidelines of 11 major social media platforms. We find 66 different types of rules across their community guidelines, with great variability in the coverage of these rules across different platforms. Our research reveals the types of misbehavior that platforms chose to focus on, and motivates further inquiries into policymaking and content moderation in specific problem areas such as inciting violence and voter suppression.

CSCW'20 Companion, October 17–21, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-8059-1/20/10. https://doi.org/10.1145/3406865.3418312

KEYWORDS

Community guidelines; content moderation; platform governance; social media

Adult Non-Consensual Intimate Imagery (11)
Adult Non-Sexual Nudity (6)
Celebrating Own Crime (2)
Child Exploitation Imagery (11)
Child Nudity (9)
Coordinating harm (9)
Creep Shots (5)
Criminal Group Coordination (7)
Criminal Group Propaganda (7)
Cruel and Insensitive (3)
Digital Nudity (6)
Distribution of Virus (4)
Eating Disorder Depiction (9)
Eating Disorder Promotion (9)
Engagement Abuse (6)
False News and Misinformation (4)
Fraud and Financial Scams (6)
Graphic Violence: Animal Abuse (10)
Graphic Violence: Child Abuse (10)
Graphic Violence: Mutilated Humans (10)
Harassment and Bullying (11)
Hate Group Coordination (6)
Hate Group Propaganda (6)
Hate Speech: Dehumanization (9)
Hate Speech: Exclusion/Segregation (9)

INTRODUCTION

Content moderation is essential for online platforms [6]. While content moderation work has largely been invisible [6], it has been increasingly at the center of the public's attention, with stories like the controversy around the different actions taken on the U.S. President's violence-inciting comments by social media platforms [7], or algorithms removing harmless content by mistake as human moderators were sent home for social distancing due to the COVID-19 pandemic [3]. Even though users often do not read community guidelines or similar sitewide policies [9], they remain the authoritative source for social media platforms to implement content moderation [4]. However, the inevitable limitation of moderation capacity means that platforms may have to face a trade-off in the types of abuse they choose to focus on, but what are the different platforms' choices? In this paper we aim to uncover these choices through the lens community guidelines, and answer the research question: How are rules similar or different across platforms in their community guidelines? Answers to this question will shed light into the differences between social media platforms in the kinds of misbehavior that they deem as important or worthy of moderation.

RELATED WORK

Policies and governance are important for community health and play a central role in translating community values to user interaction [1]. Social media platforms are often governed by multiple layers of rules [4], from higher-level terms of service and community guidelines, to lower-level rules created by individual smaller communities, such as subreddits on Reddit and Groups on Facebook. On the level of smaller subcommunities, Fiesler et al. characterized rules of 100,000 subreddits and found high variability in the types of rules created by communities, noting that the existence of certain types of rules was highly context dependent [4]. On the level of platforms, prior research has investigated policies on different problem areas, such as copyright [5] and harassment [10], and also found that they could be highly variable across websites. Our research extends prior work by similarly examining variability, but across all problem areas identified by the community guidelines of major social media platforms.

METHODS

To gain a comprehensive understanding of the community guidelines on social media platforms, we chose the 15 social media platforms with the most monthly active users based on published statistics [2]. After excluding platforms that do not have published community standards or guidelines in English (WeChat, Qzone, Sina Weibo, and Douban), we generated the final list of social media platforms for analysis, shwn in Table 1. The first two authors read the community guidelines on these social media platforms in November 2019, and independently coded for emergent rule types and then came together to adjudicate differences and iterate on codes. We did not find any new codes after coding for the rules on Facebook and YouTube, the two

Hate Speech: Inferiority (9)
Hate Speech: Slurs (9)
Hate Speech: Violence (10)
High Profile Impersonation (9)
Human Trafficking (4)
Inappropriate Interactions with Children (10)
Inauthentic Behavior (11)
Intellectual Property Infringement (8)
Interrupting Platform Services (4)
Mass Murder Coordination (7)
Mass Murder Support (7)
Minors sexualization (11)
Non-Consensual Intimate Imagery Threat (6)
Non-Consensual Sexual Touching (8)
Privacy Violation (8)
Private Impersonation (9)
Prostitution (7)
Regulated Goods: Alcohol and Tobacco Sale (4)
Regulated Goods: Endangered Species Sale(4)
Regulated Goods: Firearm Sales (8)
Regulated Goods: Human Organ Sale (1)
Regulated Goods: Live Animal Sale (2)
Regulated Goods: Marijuana Sales (8)
Regulated Goods: Non-medical Drug Sale (9)
Regulated Goods: Non-medical Drug Use (5)
Regulated Goods: Pharmaceutical Sales (9)
Sadism/Glorifying Violence (6)
Self-injury Depiction (10)
Self-injury Promotion (10)
Sexual Activity (9)

platforms that have the most extensive set of rules in our dataset. The final codebook revealed a total of 66 different types of rules across all platforms, shown in Table 2.

Then, using the codebook, the first two authors both independently coded the rest of the platforms, and checked for interrater reliability. We achieved a Cohen's Kappa of at least 0.7 for every platform, which is higher than the threshold of "substantial agreement." [8] The researchers also discussed coding disagreements to ensure that they were due to reasonable subjective judgments and not systematic misunderstandings, and eventually came to an agreement on all codes. Based on this coding, we then analyzed patterns across platforms and rule types.

RESULTS

Overall, we found significant variability in the coverage of infractions between the 11 social media platforms. Facebook's Community Standards were most comprehensive and covered all 66 rule types. YouTube's Community Guidelines came in second in terms of comprehensiveness, covering 56 out of 66 rule types. Discord's Community Guidelines, on the other hand, covered only 18 rule types, the least of the 11 platforms. Table 1 shows the full list of platforms and the number of rule types their community guidelines cover.[1]

We were also able to see some high-level patterns in the coverage of rule types. All platforms had rules against adult non-consensual intimate imagery (commonly known as "revenge porn"), child exploitation imagery (commonly known as "child porn"), and minor sexualization. These infractions are severe and punishable by law—child porn and minor sexualization, for example, are U.S. federal crimes [11,12], and 46 states in the U.S. already have established laws against revenge porn [14].

All 11 platforms also had rules against harassment and bullying, as well as inauthentic behavior, which generally refers to misrepresenting one's identity in order to mislead users or the platform. The widespread inclusion of harassment policies shows a heightened focus on the increasingly severe problem on social media; it is also a marked improvement from Pater et al.'s 2016 analysis [10] of platform harassment policies, in which they noted that Twitter and Pinterest did not have explicit harassment policies in their community guidelines.

Platform	Number of Rule Types Covered
Facebook	66
YouTube	56
LinkedIn	53
Pinterest	53
Twitter	52
Instagram	51
Tik Tok	48
Viber	30
Snapchat	29
Reddit	27
Discord	18

Table 1. The number of rule types covered by each platform's community guidelines.

[1] The full analysis of the appearance of rules for each platform is available at https://bit.ly/cscw20-community-guidelines.

Sexual Solicitation (7)
Sexually Explicit Language (6)
Spam (9)
Suicide Depiction (9)
Suicide Promotion (9)
Terrorism Coordination (8)
Terrorist Propaganda (8)
Theft (4)
Vandalism (3)
Violence and Incitement (10)
Voter Fraud and Suppression (3)

Table 2. 66 types of rules identified from community guidelines. Numbers in parentheses indicate the numbers of platforms that have rules regarding the specific abusive behavior.

The inclusion of Inauthentic Behavior was the result of the increasingly common fake accounts that aim to spread false information or propaganda. Facebook, for example, regularly tracks and takes down multiple inauthentic accounts working in concert to mislead people and cause harm, a kind of abuse that Facebook names "coordinated inauthentic behavior" [13].

On the other hand, we also saw some rule types that are less common. For example, only Facebook had rules against human organ sale, and only two platforms (Facebook and Instagram) had rules against live animal sale. While also under the category of regulated goods, their coverage was significantly lower than non-medical drug sale and pharmaceutical drug sale (i.e., prescription drug sale), two arguably more common types of regulated goods on social media which 9 platforms had rules for. Also, only two platforms, Facebook and LinkedIn, specifically prohibits the celebration and promotion of one's own crime, but it is also possible that other platforms did not have rules in such granularity and conflated it with other high-frequency rules such as inciting violence, for which 10 platforms had rules.

Here we would like to note that these community guidelines are not static; instead, they evolve over time and go through frequent revisions. Just like how the harassment policies have changed since Pater et al.'s analysis [10], it is likely that the community guidelines have become more comprehensive since our analysis in November 2019, by covering either additional rules that we identified, or completely new rules not listed in the current content analysis.

DISCUSSION & FUTURE WORK

Our analysis provides a comprehensive taxonomy of rules on social media platforms based on their published community guidelines. While there is clear variability in the coverage of rules on different platform, it is unclear why such a variability exists. But we can speculate that platforms may have chosen to focus on and made rules regarding the types of misbehavior that is most rampant on their platform, or made explicit the rules that are most reflective of their values. Overall, our analysis shows that platforms indeed make different choices in terms of the acceptable and unacceptable behaviors to make explicit. While no moderation system is perfect, these choices and trade-offs in rule making may contribute to the challenges and problems that platforms face.

While the coverage of rule types for specific platforms may change and update frequently, our findings provide a resource for other researchers to do more in-depth investigations about policymaking and governance in specific problem areas. For example, while prior research has examined harassment and copyright, what about inciting violence or voter suppression?

Furthermore, while rules often reflect the values and norms of a community, it is unclear how much these sitewide rules are perceived by users, from whom the implicit values and norms often derive. Therefore, for future work, we plan to conduct a large-scale survey to understand social media users' perceptions of the violation of these rules. We hope the result will inform policy makers and platform designers to enact governance with a careful consideration of their global users' differing values and backgrounds.

REFERENCES

[1] Alissa Centivany. 2016. Values, ethics and participatory policymaking in online communities. *Proceedings of the Association for Information Science and Technology* 53, 1: 1–10. https://doi.org/10.1002/pra2.2016.14505301058

[2] J. Clement. 2019. Facebook users by country 2019. *Statista*. Retrieved November 20, 2019 from https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/

[3] Elizabeth Dwoskin and Nitasha Tiku. Facebook sent home thousands of human moderators due to the coronavirus. Now the algorithms are in charge. *Washington Post*. Retrieved June 25, 2020 from https://www.washingtonpost.com/technology/2020/03/23/facebook-moderators-coronavirus/

[4] Casey Fiesler, Jialun "Aaron" Jiang, Joshua McCann, Kyle Frye, and Jed R. Brubaker. 2018. Reddit Rules! Characterizing an Ecosystem of Governance. In *Twelfth International AAAI Conference on Web and Social Media*. Retrieved August 30, 2018 from https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17898

[5] Casey Fiesler, Cliff Lampe, and Amy S. Bruckman. 2016. Reality and Perception of Copyright Terms of Service for Online Content Creation. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW)*, 1450–1461.

[6] Tarleton Gillespie. 2018. Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media. Yale University Press.

[7] Mike Isaac and Cecilia Kang. 2020. While Twitter Confronts Trump, Zuckerberg Keeps Facebook Out of It. *The New York Times*. Retrieved June 25, 2020 from https://www.nytimes.com/2020/05/29/technology/twitter-facebook-zuckerberg-trump.html

[8] J R Landis and G G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1: 159–174. https://doi.org/10.2307/2529310

[9] Jonathan A. Obar and Anne Oeldorf-Hirsch. 2020. The biggest lie on the Internet: ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society* 23, 1: 128–147. https://doi.org/10.1080/1369118X.2018.1486870

[10] Jessica A. Pater, Moon K. Kim, Elizabeth D. Mynatt, and Casey Fiesler. 2016. Characterizations of Online Harassment: Comparing Policies Across Social Media Platforms. In *Proceedings of the 19th International Conference on Supporting Group Work* (GROUP '16), 369–374. https://doi.org/10.1145/2957276.2957297

[11] U.S. Department of Justice. 2015. Citizen's Guide To U.S. Federal Child Exploitation And Obscenity Laws. Retrieved November 20, 2019 from https://www.justice.gov/criminal-ceos/citizens-guide-us-federal-child-exploitation-and-obscenity-laws

[12] 2015. Child Pornography. Retrieved June 23, 2020 from https://www.justice.gov/criminal-ceos/child-pornography

[13] 2019. Removing Coordinated Inauthentic Behavior From Georgia, Vietnam and the US. *About Facebook*. Retrieved June 23, 2020 from https://about.fb.com/news/2019/12/removing-coordinated-inauthentic-behavior-from-georgia-vietnam-and-the-us/

[14] 46 States + DC + One Territory NOW have Revenge Porn Laws | Cyber Civil Rights Initiative. Retrieved June 23, 2020 from https://www.cybercivilrights.org/revenge-porn-laws/