Auditing Automated Content Moderation Tools

Motivation

As much of the world's discourse happens online, platforms that host and mediate online content play an increasingly influential role in moderating societal conversations. Platforms like Twitter, Threads, and TikTok have been compared to "global town squares" [12]. The practice of deciding whether to publish, remove, and flag content that is posted by third-party users is typically referred to as *content moderation* [8, 7]. This is what Goldman [3] refers to as "content regulation". Platforms that are faced with the prospect of moderating content face two primary challenges: (1) enforcing policies at scale; (2) ensuring that policies are applied consistently.

First, as the *scale* of data to be monitored and moderated has increased, consistent content moderation has become extremely challenging. In part due to the scale of the problem, alongside regulatory pressures, platforms have increasingly attempted to rely on algorithmic automated content moderation [4]. Yet, in reality, it is unclear just how much automation has taken over: much content moderation is implemented by underpaid contractors, largely in the Global South [1], hired by large platforms [2]. Second, content moderation needs to be *consistent*—across instances of content, users, geographies, and so forth. Greater consistency may lead to increased trust and improved discourse [8]. Unfortunately, public perception is that content moderation is inconsistently implemented [10, 9], and it is is often difficult to determine where, how and according to what rules content moderation is occurring [7]. The key question we then ask is:

How can we audit deployed content moderation systems for compliance with stated policies?

Research Questions

In this project, we aim to measure the consistency of deployed automated content moderation tools by designing a suite of tests. We look to measure both the effectiveness and reliability of these largely black-box tools. Based on our analysis of the content moderation policies and implementation approaches followed by the top platforms hosting user-generated content [11], we identified Twitch [1] as a promising platform for auditing. Twitch has several advantages for this study, the first being that it provides both users and content creators with direct access to automated tools to moderate the content they view. Second, streams on Twitch can be 'siloed', so we can set up controlled experiments where only experiment participants view the content to be tested. This prevents unintended harms as some of the content we test is offensive in nature. The main research questions we aim to answer are:

RQ1: What is the set of automated content moderation tools offered by Twitch, and how customizable are they? In particular, we look to determine the domains (image, text, audio etc.) and content types (hate speech, misinformation etc.) for which Twitch provides tools to users for content moderation. We also look to determine the use of machine learning in these tools.

RQ2: How effective are these tools at identifying content to be moderated, and how consistent with the platform's policies are they? Through the use of benchmark datasets, we aim to test the effectiveness of the automated tools provided across content types like hate speech, misinformation and abuse. We also want to determine if the moderation behavior is consistent with the platform's stated policies, and under different experimental conditions. The cross-cultural validity of the moderation behavior can also be tested.

RQ3: Can malicious users bypass the automated content moderation? Given instances where the automated moderation behaves as expected, it is critical to understand how robust the moderation is. We aim to check the resilience of the moderation algorithms to character substitutions [6] and adversarial examples [5].

Project Phases

The current state of the project is tracked in this GitHub wiki. The project will take place in 3 phases:

Phase 1: Platform Exploration [Completed] In this phase, we mainly tried to understand relevant platforms for auditing. In particular, we were looking for platforms that explicitly used machine learning-based tools for moderation, and not just rule-based detection. We identified Twitch and Discord as promising platforms, and focused on the moderation of text on Twitch as the first object of study. We determined Hate Speech and Harassment as the two categories of prohibited content we would focus on.

Phase 2: Moderation Effectiveness Evaluation [Ongoing] This phase has two components. The first, which is largely complete, involves engineering the pipeline required to automatically deliver the content to be tested to the platform. This required the creation

[1] We also identified Discord as another potential platform to test. However, currently Discord only filters images automatically, not text.

of several bots to send and receive data at appropriate rates. Additional bot functionality in terms of interleaving and trigger-based data acquisition is ongoing. The second component involves creating relevant datasets for evaluating the moderation effectiveness and its compliance with stated policies. This component is in a very *nascent stage* and *requires more hands on deck*. Several interesting design aspects need to be considered when creating the datasets: i) is the data conversational and contextual?; ii) does it violate stated platform policies?; iii) is it culturally representative in terms of its undesirable impact? We envision the data used for testing to be a combination of existing datasets and newly generated examples for the purposes of this evaluation.

Phase 3: Resilience Testing [Future] Given the results from Phase 2, we will identify successful instances of moderation and tweak the input data in order to induce moderation failure. Character substitution and black-box adversarial examples are two potential techniques, and others can be explored. The main aim would be to check whether semantics can be maintained while bypassing moderation.

At the end of Phase 3, the paper will be submitted to ACM FaccT 2025, which has a deadline on January 22, 2025.

Future Work

There are several potential directions for future work. The first would involve extending the study to other domains such as images, videos and audio, as well as to other platforms such as Discord. Each of these domains will present its own challenges in terms of dataset creation for testing, and the subsequent evaluation, and resilience testing. Once extended, a suite of such probe datasets, and the tools for their creation, can be made available for the community to audit other models.

A second, more ambitious direction, would involve automated compliance testing. Given a set of policies regarding platform governance, how can researchers create probe data in an automated fashion to check for compliance, and can compliance be proved in any formal manner?

Finally, the creation of open-source automated content moderation systems that learn from human feedback, are resilient to manipulated inputs, and can be effective under multi-lingual and cross-cultural contexts remains an open question. In particular, it is unclear whether a single large model or several smaller models should be deployed for such purposes.

References

[1] A. Chen. The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed, 2023.

[2] V. Elliott. Meta's Gruesome Content Broke Him. Now He Wants It to Pay, 2023.

[3] E. Goldman. Content moderation remedies. *Michigan Technology Law Review, Forthcoming*, 2021.

[4] R. Gorwa, R. Binns, and C. Katzenbach. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1):2053951719897945, 2020.

[5] T. Gröndahl, L. Pajola, M. Juuti, M. Conti, and N. Asokan. All you need is" love" evading hate speech detection. In *Proceedings of the 11th ACM workshop on artificial intelligence and security*, pages 2–12, 2018.

[6] C. Hiruncharoenvate, Z. Lin, and E. Gilbert. Algorithmically bypassing censorship on sina weibo with nondeterministic homophone substitutions. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9, pages 150–158, 2015.

[7] D. Keller, P. Leerssen, et al. Facts and where to find them: Empirical research on internet platforms and content moderation. *Social media and democracy: The state of the field and prospects for reform*, 220:224, 2020.

[8] S. Kiesler, R. E. Kraut, P. Resnick, A. Kittur, M. Burke, Y. Chen, N. Kittur, J. Konstan, Y. Ren, and J. Riedl. *Regulating Behavior in Online Communities*, pages 125–178. The MIT Press, 2011.

[9] M. Ozanne, A. Bhandari, N. N. Bazarova, and D. DiFranzo. Shall ai moderators be made visible? perception of accountability and trust in moderation systems on social media platforms. *Big Data & Society*, 9(2):20539517221115666, 2022.

[10] Sabin, Sam. On policing content, social media companies face a trust gap with users. `https://pro.morningconsult.com/articles/on-policing-content-social-media-companies-face-a-trust-gap-with-users`, 2019. Accessed: 2023-09-14.

[11] B. Schaffner, A. N. Bhagoji, S. Cheng, J. Mei, J. L. Shen, G. Wang, M. Chetty, N. Feamster, G. Lakier, and C. Tan. " community guidelines make this the best party on the internet": An in-depth study of online platforms' content moderation policies. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2024.

[12] L. Taylor. How Twitter lost its place as the global town square, July 2023.