Critique: Content Moderation

Over the last decade, the surge of new internet businesses and social media platforms has opened more opportunities for people to share content easier than ever before. Without some form of moderation, various explicit content and toxic behavior can unwillingly be posted, ruining a platform's brand reputation and overall customer experience. To mitigate this, automated content moderators have become more prominent than ever before. Their algorithms can scan through thousands of pieces of content at a rapid rate compared to human content moderators. However, these systems still contain many flaws, so it is also important to address possibilities on how they can be improved upon.

One instance of an artificial intelligence content moderator in use is the New York Times' Moderator. In 2017, the New York Times (NYT) decided to implement an automated moderator system called Moderator, powered by Perspective. The algorithm would assign obscenity scores to content and flag them if deemed toxic. For many years, the NYT used solely human moderators, which was a time-consuming process, resulting in a small number of stories to be open for commentary. With the implementation of the automated moderator, the NYT can sift through comments much faster and allow the Community Desk, consisting of trained human moderators, to review the algorithm's results to decide whether to keep or delete the comment. Although this system may sound practical, researchers found that there was a correlation between the comments flagged as inappropriate and names that were strongly associated with particular racial groups. This is an example of algorithmic bias, in which implicit human biases are incorporated into the system, causing the machine to generate unfair outcomes by further discriminating against marginalized groups.

Many content moderating algorithms utilize some form of priority queue, where the toxic content is ranked by priority for review. Priority queues are useful in finding the most toxic content, but they face difficulty in identifying changing patterns like new slang terms. Language evolves over time, and without careful supervision and updates, a trained model can quickly become ineffective.

In addition to algorithmic bias, there are issues surrounding the mental well-being of human moderators. One example of this can be found from Facebook, one of the largest content-moderated social media platforms. Facebook's content moderation operates in a similar method to the NYT, where human moderators screen through the content that was flagged by the automated system to make the final judgment on whether to display or block that content from viewers' pages. The main difference between the two platforms is that Facebook affords users to post a variety of content, such as articles, photos, and videos, whereas the NYT is solely based on text content. This affordance allows users to share more visually disturbing posts, ranging from hate speech to recorded murders. Since human moderators are constantly exposed and forced to review sensitive content, they face high emotional tolls, or may develop anxiety or depression. According to The Verge, workers receive considerably low pay and can be immediately fired for making small mistakes, and they often cope with stress by telling dark humor or taking drugs. Therefore, working as a content moderator is a difficult job with questionable ethics.

Based on our research, we advocate for the use of an automated content moderating system along with human moderators. As the platform scales up, algorithmic content moderation enables for more user engagement and interactions on the platform because more content can be reviewed compared to only relying on human review. With less inappropriate, radical content being displayed, it enables broader audiences to have a great experience on the platform. Those who interact need to be protected from spam, trolling, and explicit content to bond together as a community. Today, there are still many issues with moderating that need to be improved upon, such as algorithmic bias, mental health of human moderators, and balance between automation and human decision-making. It is critical to address particular groups who may be affected and how we can improve the environment of those who work in the field. One of our suggestions for improving the priority queue system is by continuously learning how language use evolves to appropriately assign priority values and toxicity categories. Additionally, we need to push for improvement of the human moderators' conditions by increasing pay and providing access to mental health clinic centers. Without taking these steps towards empathy, the entire content moderation system will not be sustainable. Making significant changes to moderating will not only benefit our company in the long run, but also create a lasting societal impact on future online platforms to come.

References

Chotiner, I. (2019, July 5). *The Underworld of Online Content Moderation.* The New Yorker. https://www.newyorker.com/news/q-and-a/the-underworld-of-online-content-moderation

Newton, C. (2019, February 25). *The Secret Lives of Facebook Moderators in America.* The Verge. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

Salganik, M. J., Lee, R. C. (2020, April 30). *To Apply Machine Learning Responsibly, We Use It in Moderation.* NYT Open. https://open.nytimes.com/to-apply-machine-learning-responsibly-we-use-it-in-moderation-d001f49e0644

Schoolov, K. (2021, February 27). *Why Content Moderation Costs Billions and is so Tricky for Facebook, Twitter, Youtube, and others.* CNBC. https://www.cnbc.com/2021/02/27/content-moderation-on-social-media.html

Trebacz, A. (2020, July 19). *Improving Content Moderation is How Platforms Get Better.* Medium. https://medium.com/the-innovation/improving-content-moderation-is-how-platforms-get-better-bfb8114180f

York, J. C., McSherry, C. (2019, April 29). *Content Moderation is Broken. Let Us Count the Ways.* Electronic Frontier Foundation. https://www.eff.org/deeplinks/2019/04/content-moderation-broken-let-us-count-ways