How social media firms moderate their content

February 3 2022

Credit: Unsplash/CC0 Public Domain

Content moderation is a delicate balancing act for social media platforms trying to grow their user base. Larger platforms such as Facebook and Twitter, which make most of their profits from advertising, can't afford to lose eyeballs or engagement on their sites. Yet they are under tremendous public and political pressure to stop disinformation and remove harmful content. Meanwhile, smaller

platforms that cater to particular ideologies would rather let free speech reign.

In their forthcoming paper, titled "Implications of Revenue Models and Technology for Content Moderation Strategies," Wharton marketing professors Pinar Yildirim and Z. John Zhang, and Wharton doctoral candidate Yi Liu show how a social media firm's content moderation strategy is influenced mostly by its revenue model. A platform under advertising is more likely to moderate its content than one under subscription, but it moderates less aggressively than the latter when it does. In the following essay, the authors discuss their research and its implications for policymakers who want to regulate social media platforms.

Every day, millions of users around the world share their diverse views on social media platforms. Not all these views are in harmony. Some are considered offensive, harmful, even extreme. With diverse opinions, consumers are conflicted: On the one hand, they want to freely express their views on ongoing political, social, and economic issues on social media platforms without intervention and without being told their views are inappropriate. On the other hand, when others express their views freely, they may consider some of that content inappropriate, insensitive, harmful, or extreme and want it removed. Moreover, consumers do not always agree about what posts are objectionable or what actions social media platforms should take. According to a survey by Morningconsult, for instance, 80% of those surveyed want to see hate speech—such as posts using slurs against a racial, religious, or gender group—removed, 73% wish to see videos depicting violent crimes removed, and 66% wish to see depictions of sexual acts removed.

Social media platforms face a challenge acting as the custodians of the internet, while at the same time being the center of self-expression and user-generated content. Indeed, content moderation efforts eat up

significant resources of firms. Facebook alone has committed to allocating 5% of the firm's revenue, $3.7 billion, on content moderation, an amount greater than Twitter's entire annual revenue. Yet neither consumers nor regulators seem to be satisfied with their efforts. In one form or another, firms need to decide how to moderate content to protect individual users and their interests. Should sensitive content be taken down from the internet? Or should free speech rule freely, indicating all are free to post what they want, and it is the consumer's decision to opt in or out of this free speech world? Taking down someone's content reduces that user's (and some other users') enjoyment of the site, while not taking it down can also offend others. Therefore, in terms of a social media platform's economic incentives, content moderation can affect user engagement, which ultimately can affect the platform's profitability.

Moderating Content, Maximizing Profits

In our forthcoming paper, "Implications of Revenue Models and Technology for Content Moderation Strategies," we study how social media platforms driven by profits may or may not moderate online content. We take into account the considerable user heterogeneity and different revenue models that platforms may have, and we derive the platform's optimal content moderation strategy that maximizes revenue.

When different social media platforms moderate content, the most significant determinant is their bottom line. This bottom line may rely heavily on advertising, or delivering eyeballs to advertisers, or the subscription fees that individual consumers are paying. But there is a stark contrast between the two revenue models. While advertising relies on delivering many, many eyeballs to advertisers, subscription revenues depend on being able to attract paying customers. As a result of the contrast, the content moderation policy in an effort to retain consumers also looks different under advertising vs. subscription. Social media

platforms running on advertising revenue are more likely to conduct content moderation but with lax community standards in order to retain a larger group of consumers, compared to platforms with subscription revenue. Indeed, subscription-based platforms like Gab and MeWe are less likely to do content moderation, claiming free speech for their users.

A second important factor in content moderation is the quality of the content moderation technology. A significant volume of content moderation is carried out with the help of computers and artificial intelligence. Why, then, do social media executives claim the technology is not sufficient? When asked about content moderation, most executives at Facebook emphasize that they care a lot about content moderation and allocate large amounts of firm revenue to the task.

We find that a self-interested social media platform does not always benefit from technological improvement. In particular, a platform whose main source of revenue is from advertising may not benefit from better technology, because less accurate technology creates a porous community with more eyeballs. This finding suggests that content moderation on online platforms is not merely an outcome of their technological capabilities, but their economic incentives.

The findings from the paper overall cast doubt on whether social media platforms will always remedy the technological deficiencies on their own. We take our analysis one step further and compare the content moderation strategy for a self-interested platform with that for a social planner, which is a government institution or similar acting body that sets rules for the betterment of societal welfare. A social planner will use content moderation to prune any user who contributes negatively to the total utility of society, whereas a self-interested platform may keep some of these users, if it serves its interests. Perhaps counter to lay beliefs, we find that a self-interested platform is more likely to conduct content moderation than a social planner, which indicates that individual

platforms have more incentives to moderate their content compared to the government.

However, more incentives do not mean right incentives. When conducting content moderation, a platform under advertising will be less strict than a social planner, while a platform under subscription will be stricter than a social planner. Moreover, a social planner will always push for perfect technology when the cost of developing technology is not an issue. Only a platform under subscription will have its interest aligned with a social planner in perfecting the technology for content moderation. These conclusions overall demonstrate that there is room for government regulations, and when they are warranted, they need to be differentiated with regard to the revenue model a platform adopts.

More information: Yi Liu et al, Implications of Revenue Models and Technology for Content Moderation Strategies, *SSRN Electronic Journal* (2021). DOI: 10.2139/ssrn.3969938

Provided by University of Pennsylvania