Social Media Censorship to Maintain Free Speech and Lessen Mental Health Problems While Preventing Offensive Content

Madison Reznicki

PHIL 213: Contemporary Moral Problems

Dr. Hoi Yee Chan (Phoebe)

May 10, 2023

Word count: 2,016

Abstract

The rise of social media has had many benefits and drawbacks. It can connect individuals, convey information, and share personal details about users' lives. However, freedom of speech online can lead to negative outcomes such as hate speech, violence, and bullying. All these situations can take a toll on users' mental health and result in stress, anxiety, depression, and more. Due to an influx of children and teenagers online, social media platforms have started censoring content with illegal activities, excessive nudity, and graphic images. Some platforms, such as Instagram, TikTok, and Twitter, also offer trigger warnings before graphic content to warn users. Some argue that social media platforms should hold themselves to higher standards of censorship, but which social media censorship solution should we endorse that balance considerations regarding users' mental health and their freedom of speech? By utilizing the temporarily quarantining method, suggested by Ullmann, of removing an offensive post before it is posted, social media companies can censor negative content to lessen the negative impact on users' mental health while maintaining freedom of speech.

Introduction

The internet and social media became widely adopted in the 1990s and 2000s and was considered one of the biggest social and technological advancements of our time. One of the first social media platforms was created in 1997 called SixDegrees.com. Users could set up their profile and send messages to their network. This led to further technological advancements that created the opportunity to post photos/videos, shop for products, stay up to date with current events, and communicate in general. Individuals from different countries finally had the opportunity to connect with each other, something that had not been as easy to do before. Businesses promote their products to increase shopping patterns, amount of customers, and brand

loyalty. Now, some individuals get paid for creating content that influences their followers on their platforms. Clearly, there are many positive aspects that social media has provided since its creation. However, social media has also resulted in negative side effects on users. With a wide variety of users on social media, there is room for conflict. Some social media companies have created terms of use/community guidelines eliminating pornographic and illegal posts on their platforms outright. Still, this leaves room for hate speech and offensive content that affects various users. This kind of content can be triggering and damaging to users' mental health, and can lead to issues such as stress, anxiety, and depression. The main issue that social media companies face when trying to manage this content is finding a way to maintain users' right to free speech, which is protected under the U.S. Constitution. The line between appropriate and inappropriate free speech is very gray when instances such as sarcasm are brought into account, but social media companies have a responsibility to control the type of content being posted on their platforms. Since offensive content can be damaging to users' mental health, better censorship solutions must be implemented to prevent negative effects and promote users' mental health.

Background

Offensive content such as hate speech, violence, and graphic content negatively harms users' mental health. There are many methods that have been suggested to combat this issue such as increasing AI moderators, community moderation, using content filters, or textual censorship. With the AI method though, AI may delete sarcastic posts that do not have ill-intent. The community moderation method results in posts being reported by other users before they're taken down, and the damage is already done. Content filtering can often result in false negatives and false positives. One of the most effective methods, suggested by Ullmann, is to temporarily

quarantine users' posts. With this method, users' comments are not restricted indefinitely unless they are offensive. Their comments are flagged by AI, temporarily quarantined, reviewed by human moderators, and then either reposted or deleted. This method allows for free speech to be maintained by the poster, while also keeping the interests of other users in mind to decrease offensive content and mental health issues that can occur from them.

Free Speech

The First Amendment in the U.S. Constitution protects citizens right to free speech, and this includes online platforms such as social media. Social media companies are often given the freedom to choose what type of content they want to censor, and terms differ depending on each platform. For instance, TikTok is very sensitive to content featuring minors, and their AI moderators frequently remove videos with minors in them. However, other platforms, such as Twitter, are known for allowing pornographic or graphic content onto certain accounts, simply offering a warning from time-to-time. Sorabji writes about how social media companies occasionally support bad content because it attracts viewers, which the social media company may be able to profit off. Sorabji says, "Section 230 of the U.S. Communications Decency Act, which says that a platform is not responsible for the content that passes through it. But Facebook and many other media are not a mere passage for whatever passes through. Facebook deliberately *feeds* its users, as the name *newsfeed* implies, with content—sometimes sensational content—and it is said to be willing, in return for payment, to give higher prominence to an item" (Sorabji). If certain types of content are appealing to the masses, even if it is bad/offensive content, social media companies can choose to ignore censoring it so that they can profit off it. However, offensive content on social media can be damaging to users' mental health, so it is

ethically fair for social media companies to put their users' mental health first and attempt to censor offensive content.

Mental Health and Social Media

What once was a taboo topic, mental health discussions are becoming more normalized with more effort put into researching causations, treatments, and solutions. Mental health issues can be triggered by a variety of factors such as home life, school, relationships, genetics, and much more. Once social media was created, there was much debate over what aspects of social media affect users' mental health. In the article titled "Does the Internet Make the World Worse? Depression, Aggression and Polarization in the Social Media Age," Ferguson analyzes various opinions and studies on the matter. Ferguson states that there is not a causal relationship between depression and social media based on how much time one spends on social media (Ferguson). However, there was a relationship between how one uses social media plus the types of content that they see and how it affects their mental health. Ferguson concludes that "(social media) usage focused on positive disclosures predicts positive outcomes whereas negative disclosures and interactions predict negative outcomes" (Ferguson). If users have view negative content that offends or hurt them, then a negative outcome will result where they become more upset than before they interacted with social media content. Negative outcomes seem to be increasing as 27% of respondents for a 2014 Pew Research Center study claimed that they experienced at least mild harassment on social media (Ferguson). These negative outcomes were found to be associated with greater mental health issue symptoms. Even though time spent on social media may not affect mental health, the type of content that users interact with has a significant effect. This shows that social media companies should take censorship seriously to prevent mental health harm.

Temporarily Quarantining Method

Recently, hate speech detecting systems have been implemented into online platforms in forms of AI to decrease the amount of offensive content in videos, comments, sounds, captions, and more. Ullmann suggested that if the AI from these systems becomes advanced enough, they could classify posts as harmful and temporarily quarantine them upon attempt to be posted. Then, human moderators could review the content, as they are more in-touch with sarcasm, and decide whether it is allowed or not. This allows for free speech to be maintained because posts may only be temporarily quarantined rather than removed immediately. If the post is offensive, it would be removed for violating the company's guidelines. If it was not offensive, it would be reuploaded and the user's free speech would be upheld. Ullmann explains how this approach takes into consideration both users' free speech and mental health, "…quarantining has been explored as a viable approach that strikes an appropriate balance between libertarian and authoritarian tendencies" (Ullmann). Libertarian tendencies (free speech) and authoritarian tendencies (censorships) are both upheld under the temporarily quarantining method. Since this method fairly upholds free speech and mental health considerations, it is the most viable option for social media companies to follow.

Unsuccessful Methods

Currently, most social media companies use a reactive censoring method called community moderation where human moderators review content after users have already complained and reported the content themselves. Then, the human moderators decide whether it violates guidelines and should/should not be deleted. Companies such as Facebook, Twitter, and Google use this method on their platforms currently. With this method, users will have already seen the content and been caused psychological harm, so it is unsuccessful in considering the effect on

mental health. Ullmann refers to this as the Too Little Too Late method because the damage is already done by the time the post is removed. A new method suggested to censor negative social media content is to increase the amount of AI detectors used to moderate content. This idea would prevent offensive content, but it could also result in content being removed that should not be (Ullmann). Most AI is not advanced enough to understand sarcasm, which many social media users use. Sarcastic posts with no intent of harm could be removed without viable reason, violating the user's free speech and categorizing it as an unsuccessful method. Another method to censor content is by using content filters which automatically remove inappropriate words. This works for content that has text, as companies or users can filter out words of their choice. This method does not work well with videos as it cannot understand the audible content in the same way it reads the text. This leads to false negatives and positives, meaning it is an unsuccessful method as well since many platforms such as Instagram and TikTok rely on video content. A final option is textual censorship/word-filters where offensive text-based content is not removed from the platform, but instead it is visually impaired through strikethroughs, asterisks, or ellipses. This is one of the most damaging methods because the users' ill-intent is still shared, even though the inappropriate words may be visually impaired. Other users seeing the content are able to understand the idea of the message, which could be upsetting mentally. This method is unsuccessful for not taking mental health considerations into account. There are many ideas for various types of censorships methods, however, temporarily quarantining considers both free speech and mental health to the greatest extent.

Conclusion

While social media has many positive effects, many negative effects are also the result from various types of people and personalities online. Offensive content such as hate speech, violence,

and graphic material is prominent online, which can be damaging to users' mental health. Social media companies have the freedom to choose what type of content that they censor, but it is also their responsibility to look out for the well-being of their users and prevent negative effects. When deciding what censorship standards to maintain, social media companies must remember to uphold users' free speech at the same time. To reduce the strain on mental health and maintain appropriate free speech, social media companies should allow AI moderators to temporarily quarantine offensive content. Once the content is quarantine, it can be reviewed by human moderators. Then, these humans will be able to understand intention/tone and will decide if the content will be deleted or reposted. Out of all the methods presented above, this is the strongest suggestion because it does not delete potentially offensive content right away, it simply temporarily quarantines them. This ensures that this method does not fall under Ullmann's Too Little Too Late example. In conclusion, the temporarily quarantining method suggested by Ullmann is the best way to censor content while maintaining free speech on social media.

Sources

Ferguson, C. J. (2021). Does the Internet Make the World Worse? Depression, Aggression and Polarization in the Social Media Age. *Bulletin of Science, Technology & Society*, 41(4), 116–135. https://doi.org/10.1177/02704676211064567

Sorabji, R. (2020). FREE SPEECH ON SOCIAL MEDIA: HOW TO PROTECT OUR FREEDOMS FROM SOCIAL MEDIA THAT ARE FUNDED BY TRADE IN OUR PERSONAL DATA. *Social Philosophy and Policy, 37*(2), 209-236. doi:10.1017/S0265052521000121

Ullmann, S., Tomalin, M. Quarantining online hate speech: technical and ethical perspectives. *Ethics Inf Technol* 22, 69–80 (2020). https://doi.org/10.1007/s10676-019-09516-z