Content Moderation of Digital Platforms

SID:500001356

Introduction

The increasing amount of content on the Internet not only represents more traffic, but also signals a huge content risk hiding in it. Pornography, drugs, flags, violence, gore, weapons and other undesirable and harmful information not only endanger the content ecology of internet platforms but may also lead to security problems and loss of business development. It is not an exaggeration to say that content security is the lifeline of the Internet platform's risk control. In the past, Internet companies have solved the problem by increasing the size of their content auditors, for example, in 2018, Today's Headlines had expanded its original 6,000-strong operational audit team to 10,000, and social giant Facebook has 15,000 content auditors worldwide. 2016, the European Commission led a partnership with Facebook, Twitter, YouTube, and Microsoft, among other Internet giants. In 2016, the European Commission took the lead in signing a code of conduct with internet giants such as Facebook, Twitter, YouTube, and Microsoft, pledging to "block and remove hate speech within 24 hours of receiving a report". "Hate speech is already an urgent internet content safety issue in Europe and the US, and in 2019, incidents of public harm in the US and New Zealand were identified as being perpetrated by defenders of racism who had previously shared their journeys and hateful ideas on the internet. (Gillespie, 2018) And these are just the tip of the iceberg of internet content safety issues. This article will discuss the problems that have arisen in the control of the content of speech posted on online digital platforms, and the various attempts to control speech on these platforms which have revealed various controversies and whether the relevant government departments should become more involved in the control of content on digital platforms on the internet.

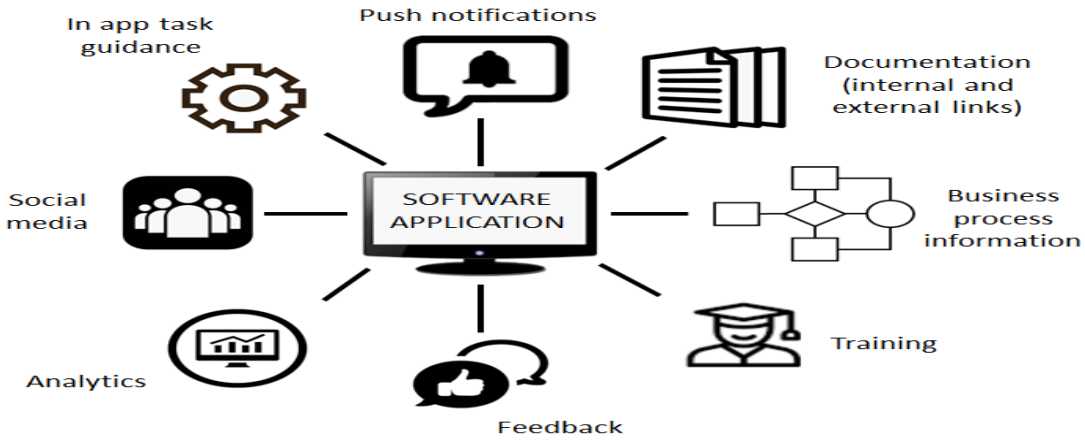

Figure 1. Digital platforms components by Wikimedia Commons are licensed under CC BY-SA 4.0.

Moderation Method

Content regulation, including the act of monitoring and withdrawing published information, is particularly challenging in an online environment where information comes in part or in whole from

a large, diverse and dispersed user base. (Roberts, 2019) Traditionally, given their role as 'passive intermediaries', social media platforms have been subject to less direct regulation than typical content creators such as broadcasters or newspapers. Social media platforms are exempt from liability for illegal content they host under laws such as Section 230 of the Communications Decency Act of 1996 in the US or the EU E-Commerce Directive in Europe. With the growing proliferation of new technologies such as fake news and deep-fake algorithms, calls for new specific regulations are growing, and in the case of Facebook, the issue has led to a massive boycott by advertisers. These problems are particularly evident in the West, where a fine balance is struck between excessive control, which can lead to protests and objections about the platform infringing on one's freedom of expression, and laxity, which can lead to an increase in racist or violent speech on the platform. At the same time, the large number of content reviewers employed by the company has come to light, with more than 3,000 content reviewers launching a lawsuit against Facebook, seeking financial compensation for the moral damage caused by their work. in May, Facebook had to agree to pay a settlement of $52 million to the content reviewers. These reviewers are faced with thousands of new posts every day and are required to carefully screen out objectionable comments that could have a negative impact, which is a serious challenge for both employees and companies. And the explosion of video, audio and other forms of media has created new challenges for manual auditing. But even with good training, manual vetting is not foolproof, as much of what is posted often strays into legal grey areas that are difficult to screen, and because users drive platform revenue, most digital platforms use a publish-and-vet approach that priorities user experience but also increases the likelihood that inappropriate content will be exposed to the public. With the development of technology, companies have gradually evolved into intelligent review, which is a combination of artificial intelligence and large databases to review users' content, but sometimes AI review is not as accurate as manual review to identify every inappropriate content. The last type of review is user-led, such as Reddit, where users elect a board leader to review the content posted each day, but this approach also has disadvantages because the leader is a voluntary worker and does not pay attention to the content posted each day like a corporate employee, so this approach often misses inappropriate content.

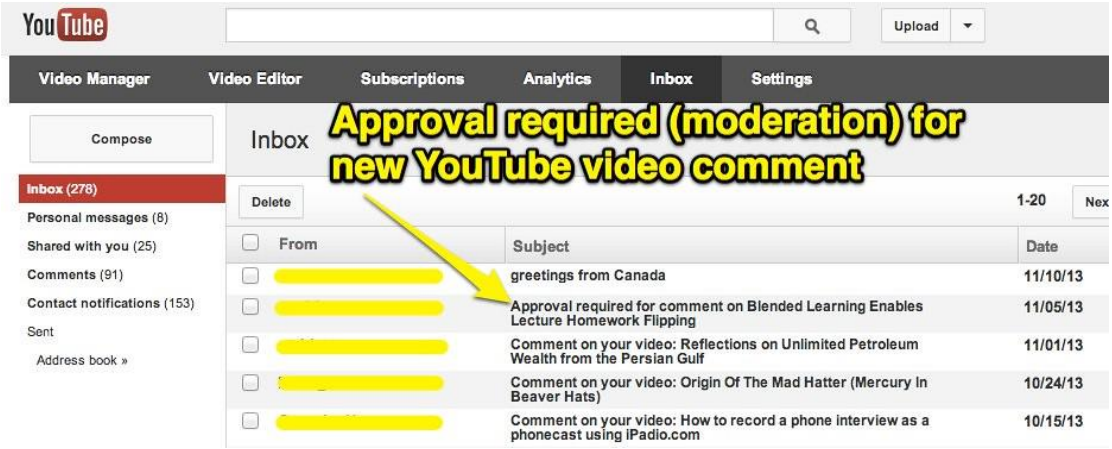

Figure 2. YouTube comment moderation by Skitch is licensed under CC BY-SA 2.0.

Controversies

On the flip side, the various platforms have brought a lot of controversy along with increased control over the content posted by users. When you give anyone the power to post anything, many people

will post anything that fits their preconceived agenda, whether it's true or not. This is especially true when they can hide behind a fake profile. Meanwhile, many people will publish hate speech under the guise of the greyest of areas - freedom of speech. When some content that is in a grey area or more ambiguous speech is banned, people feel that the platform is somehow interfering with their freedom of expression, a phenomenon that is particularly evident in Western countries, and in Asian countries, such as North Korea, where people almost completely lose their right to freedom of expression when internet content controls are within North Korea and when they reach a kind of limit, and in this environment although North Korea In this environment, although the North Korean people have become numb and accepting of this phenomenon, the practice still often attracts criticism from outside. Another extreme example is a platform called Parler, which is produced and marketed as a free speech alternative to the mainstream 'fake news' media, (Chotiner ,2019) meaning that this platform imposes almost no content controls, a feature that has attracted a large and loyal group of users who want to express their opinion or actual fake news and have it accepted as fact without the perceived But the platform soon evolved into a hotbed of racist and violent speech and has now been taken down from the Google Play Store, which in another way is a silent protest against the harsh content regulation of today's internet. While governments can make progress in preventing hate speech or fake news, the sheer volume of both of these things being generated is more than almost anyone can imagine. Platforms need to be held accountable for their products and swift and severe penalties should be imposed on anyone who knowingly shares false information or hate speech. The government should not play a major role in trying to intervene in the control of social media content, as is obvious from the example of North Korea. It would be a good idea for the government to enact a relevant content control act as soon as possible as a reference for the platforms and to leave the censorship to the individual internet companies. (Nast,2020)

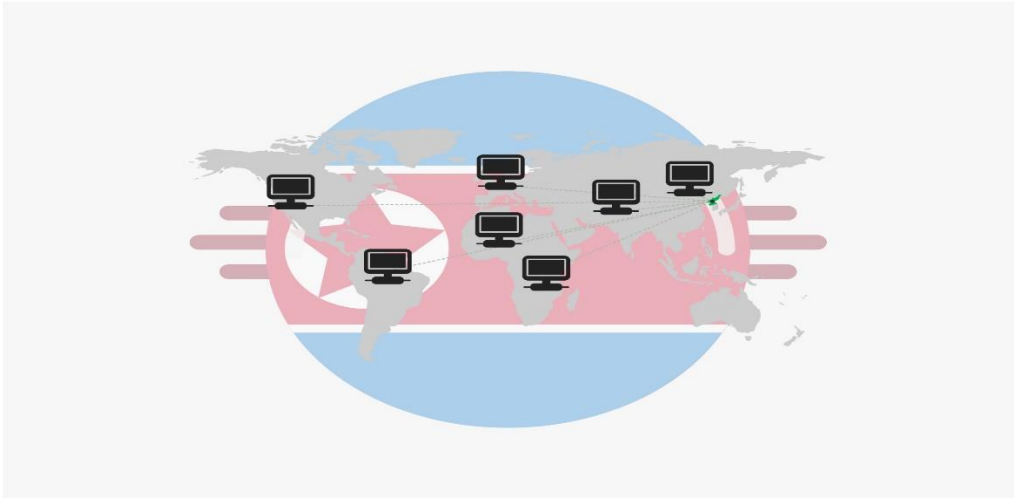

Figure 3. How North Korea Revolutionized the Internet as a Tool for Rogue Regimes by Wikimedia Commons are licensed under CC BY-SA 4.0.

Conclusion

In conclusion, as national controls continue to strengthen, to ensure the continued and stable development of products and companies, companies need to strengthen their values and manual auditing, but also require both product managers and technical skills to cope with the various uncertainties of content. For companies in general, continuous, high-cost technical investment will

inevitably increase the cost of their operations, so entrusting content security to a professional third-party organization is probably the safest thing for them to do at the moment. For large enterprises with strength and capital, both strategic and tactical attention must be paid, because it is not only a matter of values and technology, but also requires long-term training of personnel, long-term accumulation of spam profiles and a series of training. (Massanari,2017)

Reference List

1. Gillespie, Tarleton. (2018) All Platforms Moderate. In Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. Yale University Press. pp. 1-23.
2. Roberts, Sarah T. (2019) Behind the Screen: Content Moderation in the Shadows of Social Media. New Haven, CT: Yale University Press, pp. 33-72.
3. Massanari, Adrienne (2017) #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. New Media & Society, 19(3): 329–346.
4. Chotiner, I. (2019). The Underworld of Online Content Moderation. The New Yorker. Retrieved 13 October 2021, from https://www.newyorker.com/news/q-and-a/the-underworld-of-online-content-moderation.
5. Nast, C. (2020). More Content Moderation Is Not Always Better. Wired. Retrieved 13 October 2021, from https://www.wired.com/story/more-content-moderation-not-always-better/.