Unsplash © Jeremy Bezanger

Why Facebook’s content moderation system doesn’t work

Facebook and its apps have come under fire for allowing toxic content to cause harm across its platforms. Despite Facebook’s denial that its platforms do not benefit from harmful content and proactively remove it, toxic information is still spreading rapidly across the platforms.

Online abuse is rampant, and misinformation has been allowed to spread rapidly in the last year alone, generating followers and profits for its perpetrators.

There’s a gap between Facebook’s claims that it censors harmful content and the swathe of misinformation and harmful content getting traction online. At the heart of it is an obscure process of content moderation that allows toxic information to slip through.

“Extremist groups are very good at sanitizing their discourse,” said Maygane Janin, senior research analyst at Tech Against Terrorism. “They are aware that platforms remove content linked to terrorism and violent extremism, and they will adapt their online content to limit detection by platforms’ moderators and automatic detection tools.”

“They will try to present themselves as an alternative source of information and stay within the limits of non-violent speech. This is where violent extremist

content can sometimes cross with misinformation."

By the time violent, extremist content reaches mainstream platforms like Facebook and Twitter, it is sanitized to the extent that it doesn't trigger content moderation tools to step in.

But though the violent dimension might be gone, the harm is still there. Maygane said that this is the reason that COVID-19 vaccine disinformation was able to spread so rapidly in the last year, with bad actors often posing as journalists with new and viable information, targeting everyday users and high-profile figures with large followings.

"It really starts with the most violent discourse on niche platforms, and then you end up on the biggest platform with the largest audience. The discourse is the same, but it has changed so that any anti-vaxxer can be potentially attracted by that kind of discourse," she said.

Hate speech masked as symbols

Dangerous content circulates online because it's increasingly difficult for content moderation tools to pick up on harmful discourse. This is because those perpetuating it are constantly exhausting new methods of getting around moderation tools, such as symbols, memes and emojis, which are increasingly being used to mask *hateful* speech.

"They will know all the content moderation avoidance strategies," said Maygane. "They can hack a huge sum of accounts so as soon as one account is deactivated, they can use another to continue spreading content."

"Or they can also use what's called broken text. Instead of just saying 'the Taliban', for example, they'll add a symbol in the middle of it, and sometimes that's enough to block content moderation."

Since words and symbols can have multiple meanings in different contexts, tech platforms need to stay ahead of this evolving threat.

We only need to consider the multiple meanings of Pepe the Frog to understand this. The simple frog comic has been used both as a harmless meme, a pro-democracy symbol in places such as *Hong Kong* , and a weapon of the alt-right to promote racist, anti-Semitic and homophobic content online.

"It was just a frog emoji to begin with, so it's not easy for a platform to understand how it can be used and abused," said Maygane. "When it comes to sanitized discourse and hate speech, it becomes way more complicated to understand the world context of something posted online."

The presence of toxic information

Facebook and other social media giants insist they take proactive steps to remove harmful content online. According to a representative from Facebook, over 40M pieces of hate speech were removed from the platform between April and June last year, 95% of which was found before it was reported.

But other research suggests toxic information is allowed to stay up on the platforms, even after being reported. The non-profit Center for Countering Digital Hate found that when volunteers flagged misinformation that breached the tech giant's guidelines using reporting systems in the same period, action was taken against less than *one in ten* posts.

> *Our Will to Act study, undertaken in partnership with @RestlessDev, found that @Facebook removes just 1 in 10 Covid misinformation posts when reported by users.*
>
> *The most effective tool for tackling harmful misinformation is to remove it.https://t.co/1mYVkD59Nq*
>
> *— Center for Countering Digital Hate (@CCDHate) December 16, 2020*

Andrew Carter, cofounder and CEO of social network Podium says the problem is one of scale.

"Whatever your perspective on how willing Facebook and Twitter have been to try and fix the problem, ultimately, the root of their lack of willingness is the fact that if they were to actually invest the amount required to do their job properly, it would bankrupt them, "he said.

According to Facebook, the company has tripled its safety and security team in the last number of years and now has 15,000 content reviewers. But with nearly 3B users worldwide, Andrew says the giants would “never be able to turn profit” if they invested in hiring enough human moderators to pinpoint harmful content, including sanitised content that is harder for AI to discern.

He added: “That’s notwithstanding the fact that the people who are employed to do that job are looking at the very worst of the internet for eight hours every single day.”

The system amplifies hate

Content posted on Facebook and Instagram run on algorithms that recommend posts based on what we’ve already liked or posted to maximise time spent engaging with content online. But this means that those engaging with hateful and harmful content are likely to see and spread more of it.

Facebook insists that protecting people is more important than maximising usage of the platforms and profit. But the only way to alter content users see on their profiles – as recommended by Facebook – is by using self-protective tools. According to Facebook, this includes message controls, comment controls, blocking tools and limits, meaning users can limit unwanted content during spikes of activity, such as elections or sports games.

Meanwhile, harmful content can continue to gain traction online. Last year, the *Center for Countering Digital Hate* found that in reviewing Instagram’s ‘explore’ feature, the platform recommended vaccine misinformation to users not following anti-vaxx content, while those engaging with this type of content were also being recommended anti-Semitic content and election misinformation.

Andrew says that algorithms create a “bubble effect” in amplifying negative content, but that the very design of how users engage on these platforms also contributes to this.

“If you see some positive content, most users respond by liking or sharing, which creates no new content,” he explained. “But if you see some negative content, you don’t have a mechanism for expressing that negativity. If it’s strongly negative content, people will write a reply.”

“Out of all possible options, the only one that generates more content is a strong negative reaction. The positive content is diluted, and negative content is amplified.”

What’s the solution?

The ongoing problem of harmful content online has sparked calls for tech platforms to better regulate this. The UK’s Draft Online Safety Bill is part of the UK government’s attempt to do so. It would give Ofcom powers to impose fines of up to £ 18M, or 10% of annual global turnover, in cases where a social media platform fails to comply. But it places responsibility firmly on these platforms to take action.

> *We have responded to the UK Online Safety Bill consultation. We are concerned that the Bill does not consider smaller platforms, will be ineffective in tackling terrorism online, and risks harming digital rights in the process. Full submission here: https://t.co/XAxR1jjAWK pic.twitter.com/kXwAWKiTE7*
>
> *— Tech Against Terrorism (@techvsterrorism) September 22, 2021*

“The responsibilities it creates for the big tech platforms are still woefully insufficient,” said Andrew. “Any regulation will destroy all the competition before it takes down the bigger platforms who always have the capacity to adapt and defy the rules with impunity.”

“As we’ve seen, their attitude to EU regulations has been that it’s generally cheaper to just break them and pay the fine than it is to actually become compliant.”

Maygane is also concerned that the threat won’t go away even if big tech – who have the resources to comply – tightens strategies for blocking hateful content.

“The internet is not just Facebook or YouTube. It’s a whole lot of platforms that are also being exploited. If those tech giants come down on hate speech, there will just be a migration to other platforms. It doesn’t solve the problem: it just puts a band aid on it.”

While those who spread hate and misinformation online are a small minority, the consequences are dangerous – particularly for those targeted by abusive comments or led astray by misinformation.

Existing moderation technologies have proven that they don’t work. But if social media giants are here to stay, they need to invest in new strategies to tackle this evolving, dangerous threat.

Article by ABBY WALLACE