Twitter's surge in harmful content a barrier to advertiser return

July 19 2023, by Aisha Counts and Eari Nakano, Bloomberg News

Credit: Unsplash/CC0 Public Domain

Elon Musk's Twitter acquisition, and the series of content policy changes that ensued, has led to a dramatic spike in hateful, violent and inaccurate posts on the platform, according to researchers. That's now the top

challenge for Twitter's new Chief Executive Officer Linda Yaccarino, who has to address advertisers' concerns about the trend in order to boost revenue and pay back the company's debts.

Musk and Yaccarino have touted updates to the site's policies, such as letting advertisers prevent their posts from showing up next to certain kinds of content. Still, advertising sales are down by half since Musk took control of the company in October, he said this week. That's in part because businesses don't believe there has been significant progress in resolving the problem.

"Musk is not keeping his promises to advertisers, and their ads are appearing next to really harmful content," said Callum Hood, director of research at the Center for Countering Digital Hate.

During Musk's tenure, hate speech towards minority communities increased, according to the CCDH. Reports of harassment went up and extremist content spiked, according to the Anti-Defamation League. And COVID-19 misinformation rose, according to Media Matters.

Twitter, after reviewing the research reports, said many of the harmful pieces of content have since been evaluated and addressed, in some cases through labeling, downranking or removing the posts. More than 99.99% of tweet impressions, or times a tweet was seen, are from content that does not violate Twitter's rules, according to the company.

After this story's publication, Yaccarino tweeted, calling the researchers' findings "incorrect, misleading and outdated."

Twitter has made a series of changes to its content safety efforts under Musk, such as loosening its rules, laying off trust and safety employees, reinstating accounts previously banned for violating the platform's policies and removing the verification labels on high-profile accounts

that don't want to pay for a checkmark. Those changes, in addition to turning off advertisers, have alienated many users. One out of every four Twitter users said they are unlikely to stay on the platform next year, according to a survey by Pew Research.

Twitter's Yaccarino, since starting the job in June, has talked about a "freedom of speech, not reach" strategy with brands, encouraging them to use the new controls for what ads show up next to. More than 1,600 brands now use them, according to a person familiar with the matter, who asked not to be identified sharing internal data. Yaccarino also has been soliciting plans from third parties for improving brand controls. Meanwhile, Musk has been saying that hate speech impressions are down.

Musk's argument "doesn't hold water," said Hood, who noted that both volume and engagement of hate speech has gone up, according to the CCDH. During the first three months of Musk's tenure the rate of daily tweets containing slurs against Black Americans more than tripled, the organization said, basing its research on social media analysis tool Brandwatch. From October through March, tweets referring to the LGBTQ+ community alongside slurs such as "groomer" rose 119%. Online hate often leads to real harm: reports of harassment on Twitter rose 6% this year, according to the ADL.

"Musk has repeatedly said that hate speech has decreased on the platform, but based on the data studies that we have done, we have not seen that," said Kayla Gorgarty, deputy research director at Media Matters. "We have seen the opposite."

Twitter's approach to managing hate speech focuses on limiting the number of times people see it, not the volume of the content itself, the company told Bloomberg. Twitter said impressions of hate speech content are 30% lower on average than before Musk's acquisition. The

company also noted that "groomer" is not considered a slur in their policy standards, but is a violation of their hateful conduct policy when grouped with words that are harmful towards a protected category group.

Twitter users have also reported seeing violent and sexually explicit content on the platform. Video from a mass shooting at a Texas mall earlier this year was shared openly on Twitter for hours before the company took action; so was a video of a cat in a blender.

More than 30% of U.S. adults that used Twitter between March and May reported seeing content they consider bad for the world, according to a survey conducted by the USC Marshall Neely Social Media Index. That percentage was higher than rivals Facebook, TikTok, Instagram and Snapchat. Many users reported seeing tweets that condoned or glorified violence towards marginalized groups or explicit videos easily accessible to underage children.

Earlier this year, researchers at the Stanford Internet Observatory found that Twitter failed to take down dozens of images of child sex abuse. The team identified 128 Twitter accounts selling child sex abuse material and 43 instances of known CSAM. "It is very surprising for any known CSAM to publicly appear on major social media platforms," said lead author and chief technologist David Thiel. Twitter responded to the issue after being contacted by researchers. This year Twitter removed 525% more accounts related to child sexual exploitation content than a year ago, according to the company.

Twitter has been slow to catch and remove some harmful content since Musk fired or faced resignations for nearly 75% of Twitter's staff, including the bulk of the trust and safety team, which is responsible for managing responses to content reports. On average, only 28% of antisemitic tweets reported by the ADL between December and January were removed or sanctioned. The group found the posts by drawing a

1% sample of all posts from Twitter's API, or application programming interface. Twitter has since restricted the reported tweets that were found to violate policies, the company said.

"Since Elon Musk took over Twitter, we have seen the platform go from having one of the best trust and safety divisions in the industry, to one of the worst," said Nadim Nashif, director at the Arab Center for the Advancement of Social Media.

During Musk's tenure, content from extremist political groups and misinformation related to national politics have increased. QAnon-related hashtags rose 91% in May compared with a year earlier, with the majority of those tweets occurring in the last six months, according to research from the ADL. In the first six months under Musk, nearly a quarter of the top COVID-19 related tweets included information about vaccines that is unproved and untested, according to research done by Media Matters, a left-leaning nonprofit media watchdog group funded by donors. In November, Twitter removed bans against COVID-19 misinformation.The challenge with the freedom of speech, not reach policy is, "there's no way to verify what's actually de-amplified," said Yael Eisenstat, head of the Center for Technology and Society at the ADL. Meanwhile, Musk himself has also engaged with extremist voices, replying to antisemitic conspiracy theories and anti-trans narratives, which boosts those posts because he is followed by 148 million people.

Independently verifying the safety of Twitter's platform will get harder as time goes on. In February, Twitter began charging for access to its application programming interface, or API. Twitter's API is used by third-party apps and researchers to analyze tweets. While researchers could previously access millions of tweets for free to conduct research, now they are being charged thousands of dollars for access to a fraction of that amount. To research online hate speech following the 2016 presidential election, the NYU Center for Social Media and Politics

analyzed more than 750 million tweets. Today, the university wouldn't be able to afford that research.

"Now it costs $42,000 a month to get access to just 10 million tweets," Joshua Tucker, co-director of the NYU center. Researchers at Stanford, Berkeley, the CCDH and the ADL also said they can no longer afford access to Twitter data. CCDH is U.S. and UK-based nonprofit funded by philanthropy and donations, aiming to protect human rights online. ADL is a New York-based nonprofit, also funded by donations, which says it fights all forms of extremist hate.

To reassure the public and advertisers, Twitter says it is working with independent companies Sprinklr, DoubleVerify and Integral Ad Science to assess the content on its platform. According to Twitter, the company's brand safety controls are now more than 99% effective in placing ads next to safe content.

But the damage may already be done. Advertisers have said they left Twitter because of concerns over harmful content—including Musk's own posts. The businesses generally have been working with smaller marketing budgets, and have more options for where to spend their money on digital platforms. Competition for their ads will soon get even tighter; Meta Platforms Inc., for instance, plans to eventually introduce advertising to its Twitter clone, Threads.

Twitter, which is still losing money, is trying to come up with business models that are alternatives to advertising. The company's Twitter Blue premium subscription, which costs $8 per month, has seen little uptake. This month Twitter began paying a share of ad revenue to some Twitter Blue subscribers, based on the amount of engagement with their tweets. That rewarded accounts that interact heavily with Musk himself.

2023 Bloomberg L.P.

Distributed by Tribune Content Agency, LLC.

Citation: Twitter's surge in harmful content a barrier to advertiser return (2023, July 19) retrieved 11 May 2024 from https://techxplore.com/news/2023-07-twitter-surge-content-barrier-advertiser.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.