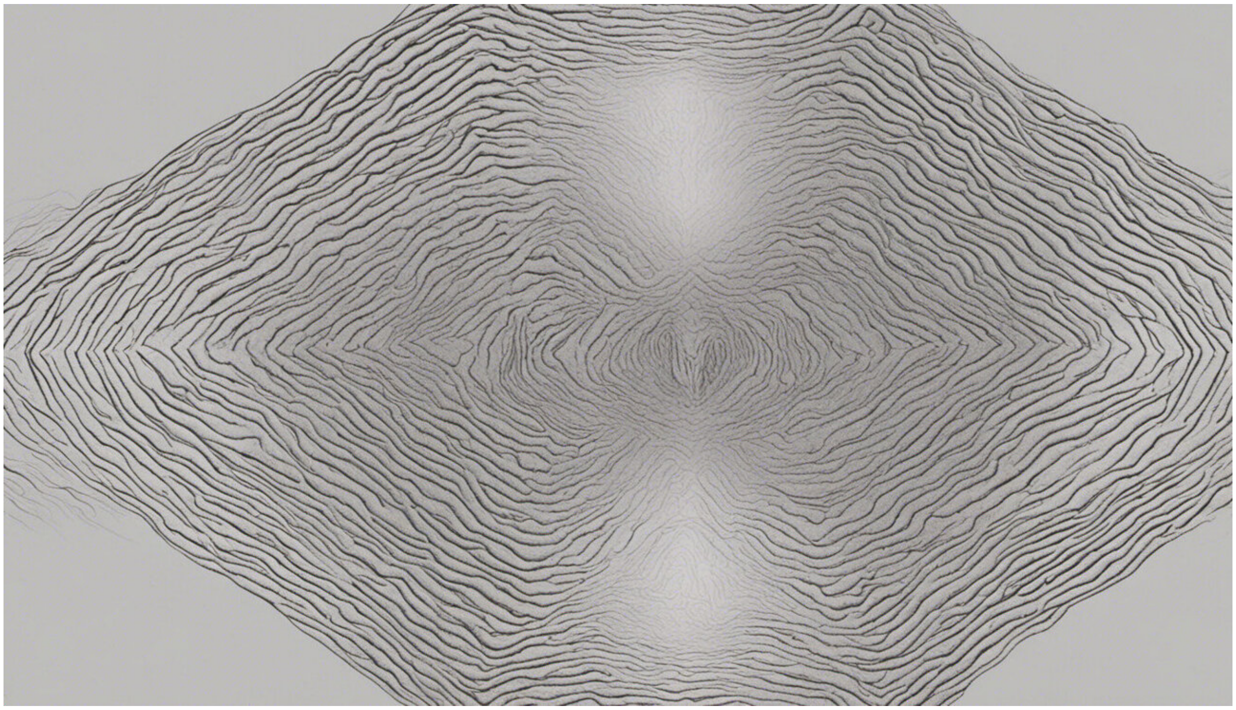
PHYS.ORG

Hate speech is still easy to find on social media

November 1 2018, by Jennifer Grygiel



Credit: AI-generated image (disclaimer)

Shortly after the synagogue shooting in Pittsburgh, I noticed that the word "Jews" was trending on Twitter. As a social media researcher and educator, I became concerned that the violence would spread online, as it has in the past.

The alleged synagogue shooter's activity on the Gab social media site has drawn attention to that site's role as a hate-filled alternative to more mainstream options like Facebook and Twitter. Those are among the social media platforms that have promised to fight hate speech and online abuse on their sites.

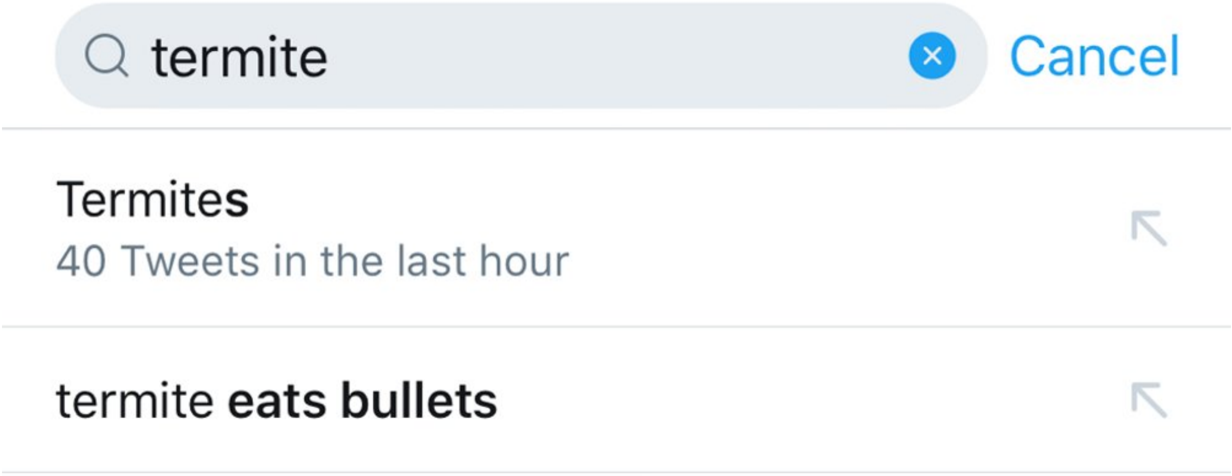
However, as I explored online activity in the wake of the shooting, it quickly became clear to me that the problems are not just on sites like Gab. Rather, hate speech is still easy to find on mainstream social media sites, including Twitter. I also identified some additional steps the company could take.

Incomplete responses to new hate terms

I was expecting new threats to appear online surrounding the Pittsburgh shooting, and there were signs that was happening. In a recent anti-Semitic attack, Nation of Islam leader Louis Farrakhan used the word "termite" to describe Jewish people. I searched for this term, knowing racists were likely to use the new slur as a code word to avoid detection when expressing anti-Semitism.

Twitter had not suspended Farrakhan's account in the wake of yet another of his anti-Semitic statements, and Twitter's search function automatically suggested I might be searching for the phrase "termite eats bullets." That turns Twitter's search box into a hate-speech billboard.

The company had, however, apparently adjusted some of its internal algorithms, because no tweets with anti-Semitic uses of the word "termite" showed up in my search results.



termite

Cancel

Termites

40 Tweets in the last hour

termite eats bullets

Credit: Jennifer Grygiel, CC BY-ND

Posts unnoticed for years

As I continued my searches for hate speech and calls for violence against Jewish people, I found even more disturbing evidence of shortfalls in Twitter's content moderation system. In the wake of the 2016 U.S. election and the discovery that Twitter was being used to influence the election, the company said it was investing in machine learning to "detect and mitigate the effect on users of fake, coordinated, and automated account activity." Based on what I found, these systems have not identified even very simple, clear and direct violent threats and hate speech that have been on its site for years.

When I reported a tweet posted in 2014 that advocated killing Jewish people "for fun," Twitter took it down the same day – but its standard automated Twitter notice gave no explanation of why it had been left untouched for more than four years.

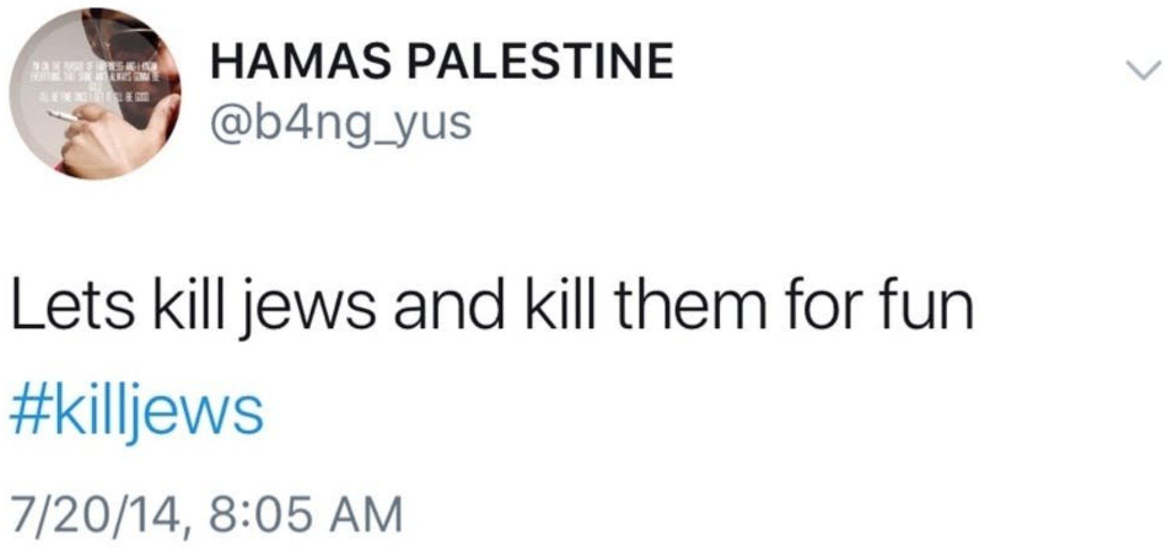
Hate games the system

When I reviewed hateful tweets that had not been caught after all those years, I noticed that many contained no text – the tweet was just an image. Without text, tweets are harder for users, and Twitter's own hate-identifying algorithms, to find. But users who specifically look for hate speech on Twitter may then scroll through the activity of the accounts they find, viewing even more hateful messages.

Twitter seems to be aware of this problem: Users who report one tweet are prompted to review a few other tweets from the same account and submit them at the same time. This does end up subjecting some more content to review, but still leaves room for some to go undetected.

Help for struggling tech giants

As I found tweets that I thought violated Twitter's policies, I reported them. Most of them were removed quickly, even within an hour. But some obviously offensive posts took as long as several days to come down. There are still a few text-based tweets that have not been removed, despite clearly violating Twitter's policies. That shows the company's content review process is not consistent.



HAMAS PALESTINE
@b4ng_yus

Lets kill jews and kill them for fun #killjews

7/20/14, 8:05 AM

One example of a hateful tweet allowed to remain on Twitter for more than four years. Credit: Jennifer Grygiel, CC BY-ND

It may seem that Twitter is getting better at removing harmful content and that it's taking down a lot of content and memes and suspending accounts, but a lot of that activity is not related to hate speech. Rather, much of Twitter's attention has been on what the company calls "coordinated manipulation," such as bots and networks of fake profiles run by government propaganda units.

In my view, the company could take a significant step to solicit the help of members of the public, as well as researchers and experts like my colleagues and me, to identify hateful content. It's common for technology companies – including Twitter – to offer payments to people who report security vulnerabilities in their software. However, all the company does for users who report problematic content is send an automatically generated message saying "thanks." The disparity in how Twitter treats code problems and content reports delivers a message that the company prioritizes its technology over its community.

Instead, Twitter could pay people for reporting content that is found to violate its community guidelines, offering financial rewards for stamping out the social vulnerabilities in its system, just as if those users were helping it identify software or hardware problems. A Facebook executive expressed concern that this potential solution could backfire and generate more online hate, but I believe the reward program could be structured and designed in a way to avoid that problem.

Much more to be done

There are further problems with Twitter that go beyond what's posted directly on its own site. People who post hate speech often take advantage of a key feature of Twitter: the ability of tweets to include links to other internet content. That function is central to how people use Twitter, sharing content of mutual interest from around the web. But it's also a method of distributing hate speech.

For instance, a tweet can look totally innocent, saying "This is funny" and providing a link. But the link – to content not posted on Twitter's servers – brings up a hate-filled message.

In addition, Twitter's content moderation system only allows users to report hateful and threatening tweets – but not accounts whose profiles themselves contain similar messages. Some of these accounts – including ones with profile pictures of Adolf Hitler, and names and Twitter handles that advocate burning Jews – don't even post tweets or follow other Twitter users. Sometimes they may simply exist to be found when people search for words in their profiles, again turning Twitter's search box into a delivery system. These accounts may also – though it's impossible to know – be used to communicate with others on Twitter via direct message, using the platform as a covert communication channel.

With no tweets or other public activity, it's impossible for users to report these accounts via the standard content reporting system. But they are just as offensive and harmful – and need to be evaluated and moderated just like other content on the site. As people seeking to spread hate become increasingly sophisticated, Twitter's community guidelines – but more importantly its enforcement efforts – need to catch up, and keep up.

If social media sites want to avoid becoming – or remaining – vectors for information warfare and plagues of hateful ideas and memes, they need to step up a lot more actively and, at the very least, have their thousands of full-time content-moderation employees search like a professor did over the course of a weekend.

Provided by The Conversation

Citation: Hate speech is still easy to find on social media (2018, November 1) retrieved 10 April 2024 from https://phys.org/news/2018-11-speech-easy-social-media.html