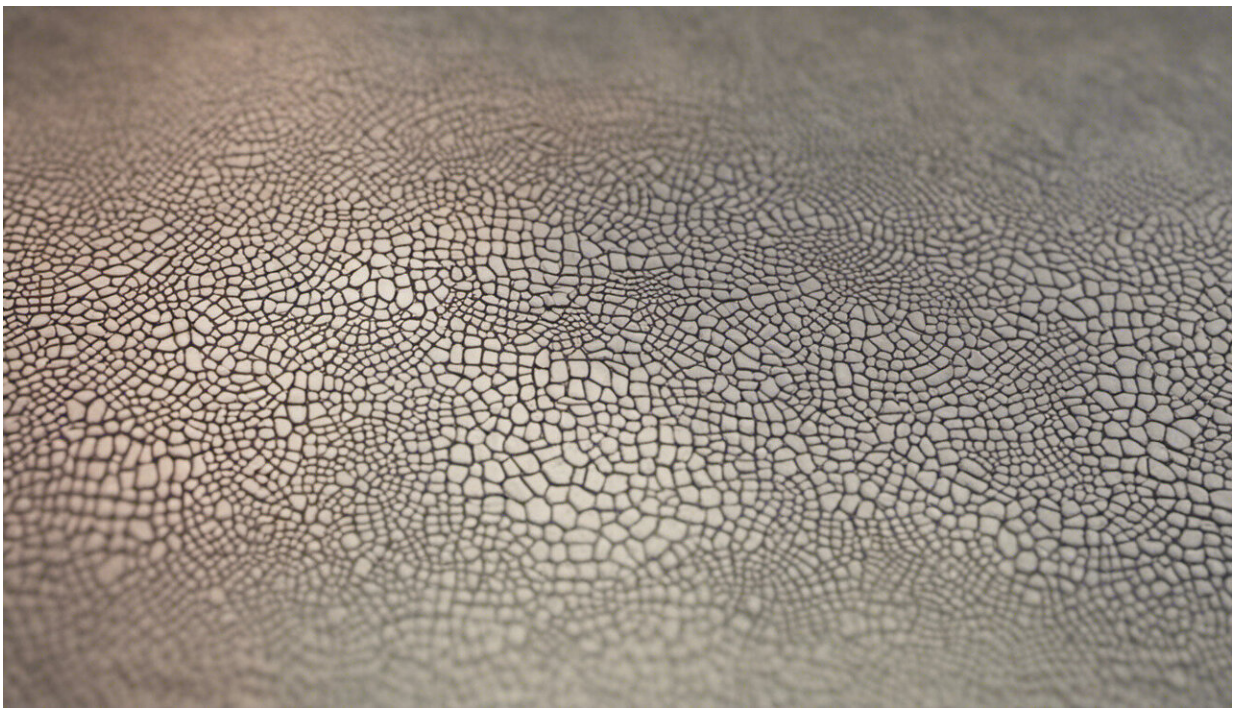
Four ways social media platforms could stop the spread of hateful content in aftermath of terror attacks

March 19 2019, by Bertie Vidgen



Credit: AI-generated image (disclaimer)

The deadly attack on two mosques in Christchurch, New Zealand, in which 50 people were killed and many others critically injured, was streamed live on Facebook by the man accused of carrying it out. It was then quickly shared across social media platforms.

Versions of the livestream attack video stayed online for a worrying amount of time. A report by the Guardian found that one video stayed on Facebook for six hours and another on YouTube for three. For many, the quick and seemingly unstoppable spread of this video typifies everything that is wrong with social media: toxic, hate-filled content which goes viral and is seen by millions.

But we should avoid scapegoating the big platforms. All of them (Twitter, Facebook, YouTube, Google, Snapchat) are signed up to the European Commission's #NoPlace4Hate programme. They are committed to removing illegal hateful content within 24 hours, a time period which is likely to come down to just one hour.

Aside from anything else, they are aware of the reputational risks of being associated with terrorism and other harmful content (such as pornography, suicide, paedophilia) and are increasingly devoting considerable resources to removing it. Within 24 hours of the Christchurch attack, Facebook had banned 1.5m versions of the attack video – of which 1.2m it stopped from being uploaded at all.

Monitoring hateful content is always difficult and even the most advanced systems accidentally miss some. But during terrorist attacks the big platforms face particularly significant challenges. As research has shown, terrorist attacks precipitate huge spikes in online hate, overrunning platforms' reporting systems. Lots of the people who upload and share this content also know how to deceive the platforms and get round their existing checks.

So what can platforms do to take down extremist and hateful content immediately after terrorist attacks? I propose four special measures which are needed to specifically target the short term influx of hate.

1. Adjust the sensitivity of the hate detection tools

All tools for hate detection have a margin of error. The designers have to decide how many false negatives and false positives they are happy with. False negatives are bits of content which are allowed online even though they are hateful and false positives are bits of content which are blocked even though they are non-hateful. There is always a trade off between the two when implementing any hate detection system.

The only way to truly ensure that no hateful content goes online is to ban all content from being uploaded – but this would be a mistake. Far better to adjust the sensitivity of the algorithms so that people are allowed to share content but platforms catch a lot more of the hateful stuff.

Mourning the victims of the Christchurch mosque attacks. Credit: EPA-EFE

2. Enable easier takedowns

Hateful content which does get onto the big platforms, such as Twitter and Facebook, can be flagged by users. It is then sent for manual review by a content moderator, who checks it using predefined guidelines. Content moderation is a fundamentally difficult business, and the platforms aim to minimise inaccurate reviews. Often this is by using the "stick": according to some investigative journalists, moderators working on behalf of Facebook risk losing their jobs unless they maintain high moderation accuracy scores.

During attacks, platforms could introduce special procedures so that staff can quickly work through content without fear of low performance evaluation. They could also introduce temporary quarantines so that content is flagged for immediate removal but then re-examined at a later date.

3. Limit the ability of users to share

Sharing is a fundamental part of social media, and platforms actively encourage sharing both on their sites (which is crucial to their business models) and between them, as it means that none of them miss out when anything goes viral. But easy sharing also brings with it risks: research shows that extreme and hateful content is imported from niche far-right sites and dumped into the mainstream where it can quickly spread to large audiences. And during attacks it means that anything which gets past one platform's hate detection software can be quickly shared across all of the platforms.

Platforms should limit the number of times that content can be shared within their site and potentially ban shares between sites. This tactic has already been adopted by WhatsApp, which now limits the number of times content can be shared to just five.

4. Create shared databases of content

All of the big platforms have very similar guidelines on what constitutes "hate" and will be trying to take down largely the same content following attacks. Creating a shared database of hateful content would ensure that content removed from one site is automatically banned from another. This would not only avoid needless duplication but enable the platforms to quickly devote resources to the really challenging content that is hard to detect.

Removing hateful content should be seen as an industry-wide effort and not a problem each platform faces individually. Shared databases like this do also exist in a limited way but efforts need to be hugely stepped up and their scope broadened.

In the long term, platforms need to keep investing in content moderation and developing advanced systems which integrate human checks with machine learning. But there is also a pressing need for special measures to handle the short-term influx of hate following terrorist attacks.

This article is republished from The Conversation under a Creative Commons license. Read the original article.

Provided by The Conversation