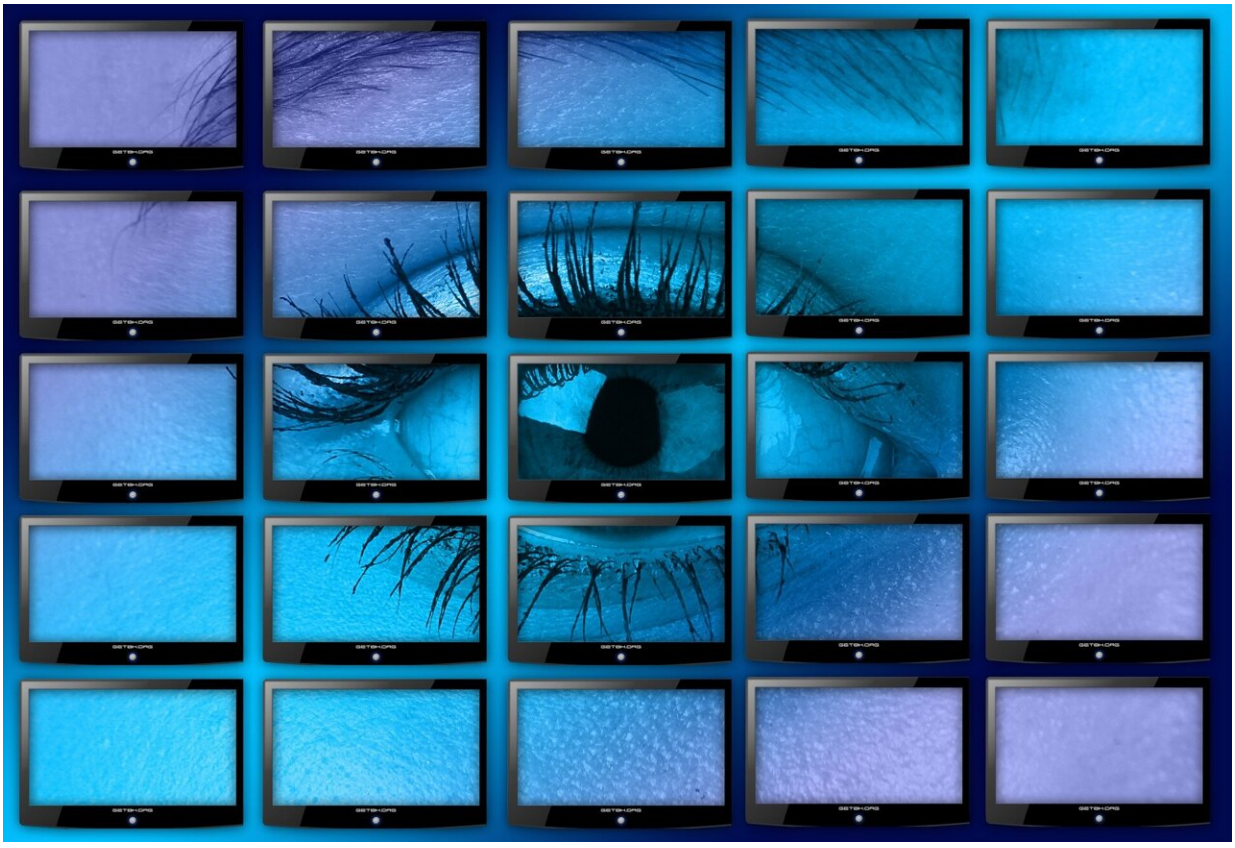
Using AI to monitor the internet for terror content is inescapable—but also fraught with pitfalls

February 7 2024, by Stuart Macdonald, Ashley A. Mattheis and David Wells



Credit: Pixabay/CC0 Public Domain

Every minute, millions of social media posts, photos and videos flood the internet. On average, Facebook users share 694,000 stories, X (formerly Twitter) users post 360,000 posts, Snapchat users send 2.7 million snaps and YouTube users upload more than 500 hours of video.

This vast ocean of online material needs to be constantly monitored for harmful or illegal content, like promoting terrorism and violence.

The sheer volume of content means that it's not possible for people to inspect and check all of it manually, which is why automated tools, including artificial intelligence (AI), are essential. But such tools also have their limitations.

The concerted effort in recent years to develop tools for the identification and removal of online terrorist content has, in part, been fueled by the emergence of new laws and regulations. This includes the EU's terrorist content online regulation, which requires hosting service providers to remove terrorist content from their platform within one hour of receiving a removal order from a competent national authority.

Behavior and content-based tools

In broad terms, there are two types of tools used to root out terrorist content. The first looks at certain account and message behavior. This includes how old the account is, the use of trending or unrelated hashtags and abnormal posting volume.

In many ways, this is similar to spam detection, in that it does not pay attention to content, and is valuable for detecting the rapid dissemination of large volumes of content, which are often bot-driven.

The second type of tool is content-based. It focuses on linguistic characteristics, word use, images and web addresses. Automated content-

based tools take one of two approaches.

1. Matching

The first approach is based on comparing new images or videos to an existing database of images and videos that have previously been identified as terrorist in nature. One challenge here is that terror groups are known to try and evade such methods by producing subtle variants of the same piece of content.

After the Christchurch terror attack in New Zealand in 2019, for example, hundreds of visually distinct versions of the livestream video of the atrocity were in circulation.

So, to combat this, matching-based tools generally use perceptual hashing rather than cryptographic hashing. Hashes are a bit like digital fingerprints, and cryptographic hashing acts like a secure, unique identity tag. Even changing a single pixel in an image drastically alters its fingerprint, preventing false matches.

Perceptual hashing, on the other hand, focuses on similarity. It overlooks minor changes like pixel color adjustments, but identifies images with the same core content. This makes perceptual hashing more resilient to tiny alterations to a piece of content. But it also means that the hashes are not entirely random, and so could potentially be used to try and recreate the original image.

2. Classification

The second approach relies on classifying content. It uses machine learning and other forms of AI, such as natural language processing. To achieve this, the AI needs a lot of examples like texts labeled as terrorist

content or not by human content moderators. By analyzing these examples, the AI learns which features distinguish different types of content, allowing it to categorize new content on its own.

Once trained, the algorithms are then able to predict whether a new item of content belongs to one of the specified categories. These items may then be removed or flagged for human review.

This approach also faces challenges, however. Collecting and preparing a large dataset of terrorist content to train the algorithms is time-consuming and resource-intensive.

The training data may also become dated quickly, as terrorists make use of new terms and discuss new world events and current affairs. Algorithms also have difficulty understanding context, including subtlety and irony. They also lack cultural sensitivity, including variations in dialect and language use across different groups.

These limitations can have important offline effects. There have been documented failures to remove hate speech in countries such as Ethiopia and Romania, while free speech activists in countries such as Egypt, Syria and Tunisia have reported having their content removed.

We still need human moderators

So, in spite of advances in AI, human input remains essential. It is important for maintaining databases and datasets, assessing content flagged for review and operating appeals processes for when decisions are challenged.

But this is demanding and draining work, and there have been damning reports regarding the working conditions of moderators, with many tech companies such as Meta outsourcing this work to third-party vendors.

To address this, we recommend the development of a set of minimum standards for those employing content moderators, including mental health provision. There is also potential to develop AI tools to safeguard the well-being of moderators. This would work, for example, by blurring out areas of images so that moderators can reach a decision without viewing disturbing content directly.

But at the same time, few, if any, platforms have the resources needed to develop automated content moderation tools and employ a sufficient number of human reviewers with the required expertise.

Many platforms have turned to off-the-shelf products. It is estimated that the content moderation solutions market will be worth $32bn by 2031.

But caution is needed here. Third-party providers are not currently subject to the same level of oversight as tech platforms themselves. They may rely disproportionately on automated tools, with insufficient human input and a lack of transparency regarding the datasets used to train their algorithms.

So, collaborative initiatives between governments and the private sector are essential. For example, the EU-funded Tech Against Terrorism Europe project has developed valuable resources for tech companies. There are also examples of automated content moderation tools being made openly available like Meta's Hasher-Matcher-Actioner, which companies can use to build their own database of hashed terrorist content.

International organizations, governments and tech platforms must prioritize the development of such collaborative resources. Without this, effectively addressing online terror content will remain elusive.

Provided by The Conversation