Technology and regulation must work in concert to combat hate speech online

March 12 2018, by Andre Oboler

Credit: Artem Podrez from Pexels

Online bullying, hate and incitement are on the rise, and new approaches are needed to tackle them. As the Australian Senate conducts hearings for its Inquiry into cyberbullying, it should consider a two-pronged approach to combating the problem.

First, the government should follow the lead of Germany in imposing financial penalties on major social media companies if they fail to reduce the volume of abusive content on their platforms.

Second, we must develop ways of correctly identifying and measuring the amount of abusive content being posted and removed to ensure that companies are complying.

Given the volume of data on social media, artificial intelligence (AI) must be a part of the mix in supporting regulation, but we need an appreciation of its limitations.

The impact on victims

In 2015, Australian lawyer Josh Bornstein was the victim of serious online abuse at the hands of a man in the United States, who impersonated Bornstein and published a racist article online in his name. Bornstein subsequently found himself on the receiving end of a barrage of hate from around the world.

The incident was highly distressing for Bornstein, but cyberhate can also have consequences for society at large. Acting under a cloak of anonymity, the same man used another fake identity to pose as an IS supporter calling for terror attacks in Australia and other Western countries. In December, he was convicted in the United States on terrorism charges.

Bornstein is now calling for both the regulation of social media companies by governments and legal remedies to enable action by victims.

Germany as a regulatory model

New legislation recently introduced in Germany requires companies to remove clear cases of hate speech within 24 hours.

In response, Facebook has employed 1,200 staff and contractors to more effectively process reports of abuse by German users. If the company fails to remove the majority of such content within the 24-hour limit, regulators can impose fines of up to €50 million (A$79 million).

These laws aren't perfect – within months of them coming into effect, Germany is already considering changes to prevent excessive caution by social media companies having a chilling effect on free speech. But the German approach gives us a window into what a strong state response to cyberbullying looks like.

This is only the cusp of a brave new world of technology regulation. Cyberbullying laws can't be enforced if we don't know how much abuse is being posted online, and how much abuse platforms are removing. We need tools to support this.

Employing artificial intelligence

At the Online Hate Prevention Institute (OHPI), we have spent the past six years both tackling specific cases – including Bornstein's – and working on the problem of measurement using world-class crowdsourcing and artificial intelligence approaches.

Others are also looking at identification and measurement as the next step. The Antisemitism Cyber Monitoring System (ACMS) – a new tool to monitor antisemitism on social media – has been under development by Israel's Diaspora Affairs Ministry since October 2016. It will be launched at the 2018 Global Forum for Combating Antisemitism in Jerusalem later this month.

The tool uses text analysis – a form of artificial intelligence – and works by searching social media sites for words, phrases and symbols that have been identified as indicators of possible antisemitic content. The tool then reviews the content and generates interactive graphs.

Similar approaches have been used by the World Jewish Congress and by Google's Conversation AI project, but the approach has limited effectiveness, particularly when applied to large social media sites.

Data from a one-month trial of ACMS was released ahead of the system's launch. While the software is being promoted as a major step forward in the fight against cyberhate, the data itself highlights serious methodological and technological limitations making it more of a distraction.

Limitations of the technology

One limitation ACMS has is detecting abuse that uses the coded language, symbols and euphemisms that are increasingly favoured by the far right.

Another is that ACMS only monitors content from Facebook and Twitter. YouTube, which accounted for 41% of the online antisemitism identified in a previous report, is not included. The automated system also only monitors content in English, Arabic, French and German.

What's more concerning is the Ministry's claim that the cities that produce the highest volume of racist content were Santiago (Chile), Dnipro (Ukraine), and Bucharest (Romania). These cities have primary languages the software is not programmed to process, yet they have somehow outscored cities whose primary languages the software does process.

Of particular concern to Australia is a graph titled Places of Interest: Level of Antisemitism by Location that shows Brisbane as the highest-ranked English-speaking city. This result has been explained by a later clarification suggesting the number is an amalgamation of global likes, shares and retweets that engaged with content originally posted from Brisbane. The data is therefore subject to a large degree of randomness based on which content happens to go viral.

Lawyers and data scientists must work together

There is a place for AI-based detection tools, but their limitations need to be understood. Text analysis can identify specific subsets of online hate, such as swastikas; language related to Hitler, Nazis, gas chambers and ovens; and antisemitic themes that are prominent among some far right groups. But they're not a silver bullet solution.

Moving beyond identification, we need both lawyers and data scientists to inform our approach to regulating online spaces. New artificial intelligence tools need to be verified against other approaches, such as crowdsourced data from the public. And experts must review the data for accuracy. We need to take advantage of new technology to support regulation regimes, while avoiding a form of failed robo-censorship akin to the robo-debt problems that plagued Centrelink.

The Inquiry into Cyberbullying is an important step, as long as it facilitates the solutions of tomorrow, not just the problems of today.

This article was originally published on The Conversation. Read the original article.

Provided by The Conversation

Citation: Technology and regulation must work in concert to combat hate speech online (2018, March 12) retrieved 4 May 2024 from https://phys.org/news/2018-03-technology-concert-combat-speech-online.html