Society | DOI:10.1145/3137764 Alex Wright

Censoring Sensors

Amid growing outcry over controversial online videos, tech firms grapple with how best to police online advertising.

Following the wave of U.K. terror attacks in the spring of 2017, prime minister Theresa May called on technology companies like Facebook and YouTube to create better tools for screening out controversial content—especially digital video—that directly promotes terrorism.

Meanwhile, in the U.S., major advertisers including AT&T, Verizon, and Wal-Mart have pulled ad campaigns from YouTube after discovering their content had been appearing in proximity to videos espousing terrorism, anti-Semitism, and other forms of hate speech.

In response to these controversies, Google expanded its advertising rules to take a more aggressive stance against hate speech, and released a suite of tools allowing advertisers to block their ads from appearing on certain sites. The company also deployed new teams of human monitors to review videos for objectionable content. In a similar vein, Facebook announced that it would add 3,000 new employees to screen videos for inappropriate content.

Yet these kinds of manual fixes won't solve the problem, especially as video becomes more and more ubiquitous thanks to smartphones, car cameras, and other embedded capture devices. Ultimately, monitoring user-generated content at scale will demand both computational and policy solutions.

At the heart of this problem lies a conflict that is deeply embedded in the history of the Internet itself: between the network's capacity for unleashing creative self-expression, and platform providers' business need to give advertisers' control over their messages. Caught between these opposing forces, companies like Google and Facebook make judgment calls about what kinds of content to deem acceptable—decisions that are often obscured from public view.

"These platforms and the firms that control them have largely been left to their own devices in terms of developing ethics," says Sarah T. Roberts, an assistant professor in the Department of Information Studies of the University of California, Los Angeles. "The public is starting to ask questions about the power these companies have."

While certain types of content clearly violate the law (child pornography, for example), far more material falls into a vast grey area ranging from the mildly insensitive or tasteless to outright hate speech. Much of this content falls well within the free speech protections of the First Amendment in the U.S., yet remains unpalatable to advertisers who fear being associated with anything that could alienate potential consumers.

In the absence of government-enforced laws and regulations, however, what ethical obligations do these companies have toward safeguarding the public digital commons that provides the foundation for their businesses?

"It's hard to argue with a straight face about freedom of speech online when that speech is commoditized and highly lucrative," says Roberts. "This is a particular version of free speech that is deeply influenced by the politics and ethos of Silicon Valley."

Laws and regulations around freedom of expression also vary widely from country to country. For example, most European Union countries place much tighter restrictions around hate speech than the U.S. does, and a political video commentary that's considered satirical in one country might be considered treasonous in another. Navigating this shifting terrain of international laws, regulations, and advertiser sensibilities—while continuing to provide as open a forum as possible to grow their audiences—presents companies like Google and Facebook with a complex, multidimensional challenge.

In an attempt to give the public some visibility into its internal dialogue around these questions, Facebook launched a series of blog posts called "Hard Questions" that provides a window into the company's current thinking.

"How aggressively should social media companies monitor and remove controversial posts and images from their platforms?," Facebook vice president of Public Policy and Communications Eric Schrage wrote in an introductory post. "Who gets to decide what's controversial, especially in a global community with a multitude of cultural norms?"

The company currently employs a combination of image matching, language understanding, and human monitoring to identify and remove Facebook groups that promote terrorist activity around the world.

"Although our use of AI against terrorism is fairly recent, it's already changing the ways we keep potential terrorist propaganda and accounts off Facebook," wrote Facebook's head of Global Policy Management Monika Bickert in a recent blog post.

Policy matters aside, the sheer vastness of the Internet—with over one billion videos on YouTube (and more than 400 hours of new content being uploaded every hour)—introduces further complexities in terms of both process and technology.

Many researchers think the long-term solution to monitoring online content will inevitably involve artificial intelligence. Recent advances in big data, machine learning, and embedded Graphics Processing Units (GPUs) are starting to pave the way for more scalable approaches to computer vision, allowing neural networks to identify emerging patterns in user-generated content that may demand further human scrutiny.

“With machine learning, we can now understand a lot more about what’s going in a video or an image,” says Reza Zadeh, an adjunct professor at Stanford University and founder and CEO of Matroid, a Palo Alto, CA-based computer vision software start-up that is developing tools for video analysis.

Built atop TensorFlow (Google’s open-source library for machine intelligence), Matroid uses a video player coupled with a so-called detector program to identify similar images in a given video stream. For example, a Matroid detector could look for images of Donald Trump across five hours of video—like a few weeks’ worth of network news broadcasts, or large volumes of YouTube videos—and pinpoint the spots where those images appear. It can also easily detect images containing gore or violence, nudity and other forms of NSFW (not safe for work) content, and look for “more like this” elsewhere in other video streams.

The company offers a self-service tool for non-technical users to train the system to spot particular images, as well as a more advanced version geared toward machine learning engineers that enables them to edit the neural network architecture, explore histograms, and ultimately create their own detectors for others to use.

While deep learning approaches are yielding advances in analyzing videos and other image-based content, the wide variety in the type and quality of video across different capture devices poses additional obstacles.

“Applying machine learning techniques to analyzing video content works reasonably well when the right conditions exist,” says George Awad, project director of TRECVID, a U.S. National Institute of Standards and Technology (NIST)-sponsored project that evaluates video search engines and explores new approaches to content-based video retrieval. “The major challenges occur when dealing with user videos in the wild that are not professionally edited.”

Even if it were possible to identify visual elements across all kinds of video files with 100% accuracy, that alone wouldn’t solve the problem of screening for objectionable material.

Even if it were possible to identify visual elements across all kinds of video files with 100% accuracy, that alone wouldn’t solve the problem of screening for potentially objectionable material. Much of the “content” of a video involves spoken words, after all, or other contextual cues that won’t be readily apparent from simply identifying an image. It’s notoriously difficult for computers to distinguish news from satire, for example—or an editorial opinion piece about terrorism from a call to arms.

In order to automate the process of screening video content at scale, researchers will likely need to apply natural language search techniques to begin parsing videos for deeper levels of nuance. “The gap between what the videos demonstrate and what an automated system would generate for a natural language description is still very big and challenging,” says Awad.

Looking ahead, Zadeh also sees plenty of opportunity on the hardware front, with semiconductor makers devising computer vision-capable chips that can work on devices like next-generation smartphones and cameras, self-driving cars, and a wide range of other video-capable devices throughout our homes and offices.

Whereas today, machine learning happens primarily over the network—with supercomputers in datacenters analyzing big datasets stored in the cloud—eventually some of those processes will migrate toward edge-layer devices. Over time, the task of identifying objectionable content may become more diffuse, as computer vision algorithms increasingly come pre-coded on chips embedded on these devices. “The more algorithms move to the source of data capture, the more challenging it will be to cope with real-world factors,” says Awad.

Ultimately, the future of monitoring digital content may have less to do with policy-making and brute-force processing at the platform provider level, and more to do with algorithmic filters making their way into the devices all around us—for better or worse.

As Zadeh puts it: “Computers are opening their eyes.”

Further Reading

Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., and Vijayanarasimhan, S.
YouTube-8M: A Large-Scale Video Classification Benchmark. arxiv.org/abs/1609.08675

Bickert, M.
Hard Questions: How we Combat Terrorism, Facebook blog post, https://newsroom.fb.com/news/2017/06/how-we-counter-terrorism/

Hegde, V., Zadeh, R.
FusionNet: 3D Object Classification Using Multiple Data Representations, 3D Deep Learning Workshop at NIPS 2016. Barcelona, Spain, 2016. http://3ddl.cs.princeton.edu/2016/papers/Hegde_Zadeh.pdf

Real, S., Shlens, J., Mazzocchi, S., Pan, X., and Vanhoucke, V.
YouTube-Bounding Boxes: A Large High-Precision Human-Annotated Data Set for Object Detection in Video. (preprint) Accepted at the Conference on Computer Vision and Pattern Recognition (CVPR) 2017. arXiv:1702.00824.

Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., and Saenko, K.
Translating Videos to Natural Language Using Deep Recurrent Neural Networks. arXiv:1412.4729 [cs.CV]

Alex Wright is a writer and researcher based in Brooklyn, NY.