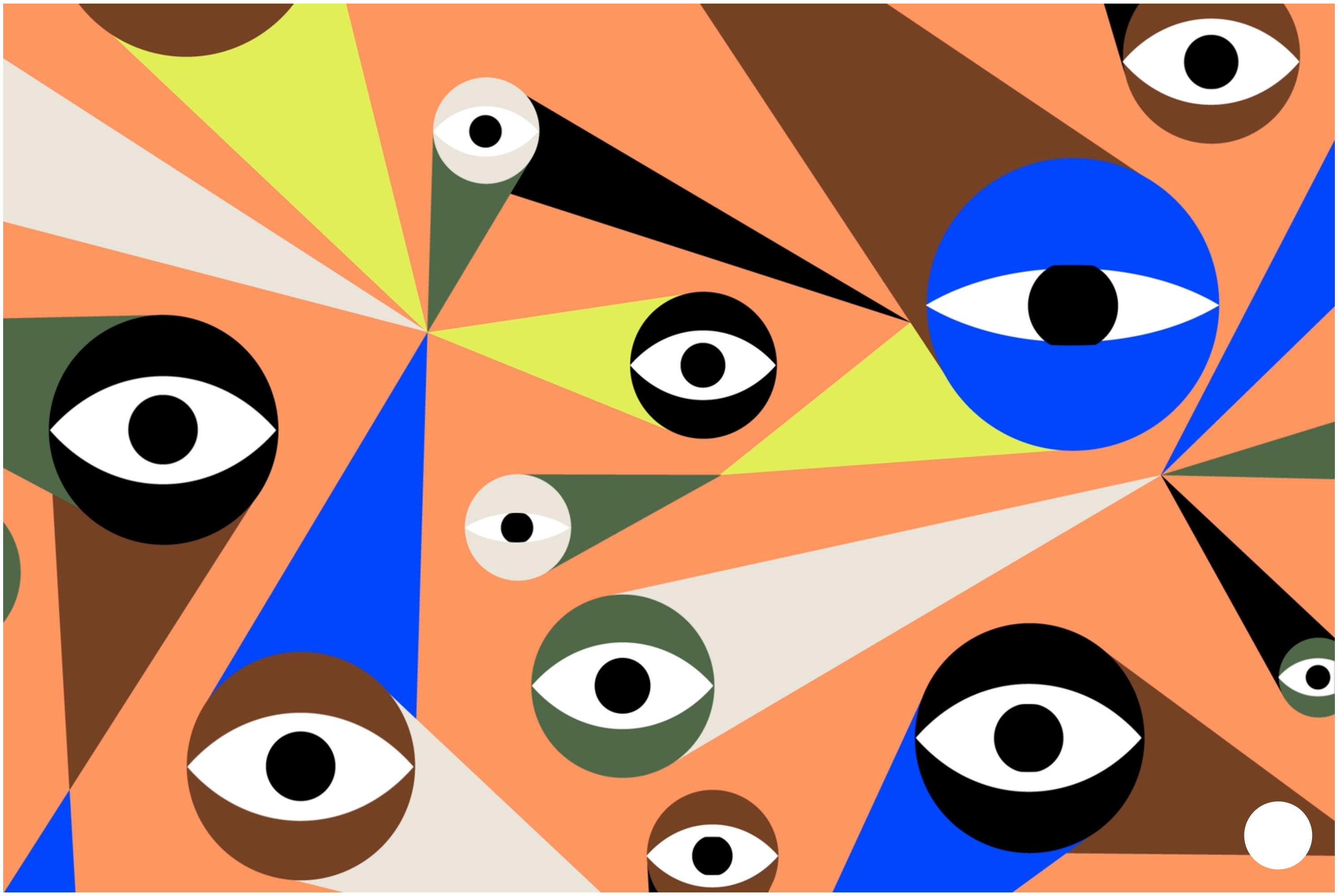
LOUISE MATSAKIS BUSINESS DEC 5, 2018 6:03 PM

Tumblr's Porn-Detecting AI Has One Job—and It's Bad at It

The blogging platform has a new policy forbidding "adult content"—but lots of innocuous posts are getting caught in the fray.



LA TIGRE

The AI Database →

APPLICATION: CONTENT MODERATION SECTOR: SOCIAL MEDIA SOURCE DATA: IMAGES TECHNOLOGY: MACHINE VISION

WHAT DO A patent application drawing for troll socks, a cartoon scorpion wearing a hard hat, and a comic about cat parkour have in common? They were all reportedly flagged by Tumblr this week after the microblogging platform announced that it would no longer allow "adult content." But so far, Tumblr's method for detecting posts that violate the new policy, which goes into effect December 17, isn't working too well, at least not according to many people on Twitter who have shared screenshots of innocent Tumblr posts that were mistakenly marked as NSFW.

The announcement was greeted with dismay in the Tumblr community, which has long been a bastion for DIY and non-mainstream porn. But the policy change appears to be having an even wider effect than anticipated. Posts are being flagged that seem to fall well outside Tumblr's definition of adult content, which "primarily includes photos, videos, or GIFs that show real-life human genitals or female-presenting nipples, and any content—including photos, videos, GIFs and illustrations—that depicts sex acts." (Users can appeal to a human moderator if they believe their posts were incorrectly labeled as adult content, and nothing will be censored until the new policy goes into effect later this month.)

"I'll admit I was naive—when I saw the announcement about the new 'adult content' ban I never thought it would apply to my blogs," says Sarah Burstein, a professor at the University of Oklahoma College of Law who noticed many of her posts were flagged. "I just post about design patents, not 'erotica.'"

Tumblr did acknowledge in a blog post announcing its new rules that "there will be mistakes" as it begins enforcing them. "Filtering this type of content versus say, a political protest with nudity or the statue of David, is not simple at scale," Tumblr's new CEO Jeff D'Onofrio wrote. This also isn't the first time a social media platform has erroneously flagged PG-rated images as sexual. Last year, for example, Facebook mistakenly barred a woman from running an ad that featured a nearly 30,000-year-old statue because it contained nudity.

But unlike with Facebook's error, many of Tumblr's mistakes concern posts that don't feature anything looking *remotely* like a naked human being. In one instance, the site reportedly flagged a blog post about wrist supports for people with a type of connective tissue disorder. Computers are now generally very good at identifying what's in a photograph. So what gives?

While it's true that machine learning capabilities have improved dramatically in recent years, computers still don't "see" images the way humans do. They detect whether groups of pixels appear similar to things they've seen in the past. Tumblr's automated content moderation system might be detecting patterns the company isn't aware of or doesn't understand. "Machine learning excels at identifying patterns in raw data, but a common failure is that the algorithms pick up accidental biases, which can result in fragile predictions," says Carl Vondrick, a computer vision and machine learning professor at Columbia Engineering. For example, a poorly trained AI for detecting pictures of food might erroneously rely on whether a plate is present rather than the food itself.

Image-recognition classifiers—like the one Tumblr ostensibly deployed—are trained to spot explicit content using datasets typically containing millions of examples of porn and not-porn. The classifier is only as good as the data it learned from, says Reza Zadeh, an adjunct computer science professor at Stanford University and the CEO of computer vision company Matroid. Based on looking at examples of flagged content users at posted on Twitter, he says it's possible Tumblr neglected to include enough instances of things like NSFW cartoons in its dataset. That might account for why the classifier mistook Burstein's patent illustrations for adult content, for example. "I believe they've forgot about adding enough cartoon data in this case, and probably other types of examples that matter and are SFW," he says.

> "Computers are only recently opening their eyes, and it's foolish to think they can see perfectly."
>
> — REZA ZADEH, MATROID

WIRED tried running several Tumblr posts that were reportedly flagged as adult content through Matroid's NSFW natural imagery classifier, including a picture of chocolate ghosts, a photo of Joe Biden, and one of Burstein's patents, this time for LED light-up jeans. The classifier correctly identified each one as SFW, though it thought there was a 21 percent chance the chocolate ghosts might be NSFW. The test demonstrates there's nothing inherently adult about these images—what matters is how different classifiers look at them.

"In general it is very easy to think 'image recognition is easy,' then blunder into mistakes like this," says Zadeh. "Computers are only recently opening their eyes, and it's foolish to think they can see perfectly."

Tumblr has had issues with flagging NSFW posts accurately before. Back in 2013, Yahoo bought Tumblr—a social network that never quite figured out how to make much money—for $1.1 billion in cash. Then four years later, like Russian nesting dolls, Verizon bought Yahoo for around $4.5 billion. (Both Yahoo and Tumblr are now part of a subsidiary of Verizon called Oath.) Right after the second acquisition—possibly in an attempt to make the site more appealing to advertisers—Tumblr introduced "Safe Mode," an opt-in feature that purported to automatically filter out "sensitive" content on its dashboard and in search results. Users quickly realized that Safe Mode was accidentally filtering normal content, including LGBTQ+ posts. In June of last year, Tumblr apologized, and said it had mostly fixed the issue.

Now the blogging platform is getting rid of the feature, because soon all of Tumblr will be in Safe Mode, permanently. It's not clear whether the company will be borrowing the same artificial intelligence technology it used for Safe Mode across the site. When asked, Tumblr didn't specify what tech it would be using to enforce its new rules for adult content. A source familiar with the company said it's using modified proprietary technology. The company did say in a support post that like most user-generated social media platforms, it plans to use a mix of "machine-learning classification and human moderation by our Trust & Safety team—the group of individuals who help moderate Tumblr." The company also says it will soon be expanding the number of human moderators it employs.

Tumblr's competitors have also benefited from over a decade head start. While Tumblr has always permitted porn—its former CEO defended allowing explicit content on the site even after it was acquired by Yahoo—other sites like Facebook have long banned explicit media. Those platforms have spent years accumulating NSFW training data to hone their the image-recognition tools. Every time a human moderator removes porn from Facebook, that example can be used to teach its AI to spot the same sort of thing on its own, as Tarleton Gillespie, a researcher at Microsoft and the author of *Custodians of the Internet* pointed out on Twitter.

Platforms like Facebook and Instagram have also already run into many of the more philosophical issues Tumblr has yet to grapple with, like when a nipple should count as being in violation of its policies or not. Tumblr will soon need to decide where it wants to draw the line between art—which it says it will allow—and pornographic material, for instance. In order to evolve into a platform free from adult content, Tumblr will have to refine its automated tools and likely train its classifiers on more expansive datasets. But the company will also need to answer lots of hard questions—ones that can only be decided by humans.

Louise Matsakis is a freelance writer covering tech. She was formerly a staff writer at WIRED covering Amazon, TikTok, and digital platforms.

CONTRIBUTOR

TOPICS TUMBLR PORN ARTIFICIAL INTELLIGENCE