Instagram/Facebook and AI

Artificial Intelligence is a critical tool to help protect people from harmful content. It helps us scale the work of human experts, and proactively take action, before a problematic post or comment has a chance to harm people.

Facebook has implemented a range of policies and products to deal with misinformation on their platform. These include adding warnings and more context to content rated by third-party fact-checkers, reducing their distribution, and removing misinformation that may contribute to imminent harm. But to scale these efforts, they need to quickly spot new posts that may contain false claims and send them to independent fact-checkers — and then work to automatically catch new iterations, so fact-checkers can focus their time and expertise fact-checking new content.

From March 1st through Election Day, Facebook displayed warnings on more than 180 million pieces of content viewed on Facebook by people in the US that were debunked by third-party fact checkers. Their AI tools both flag likely problems for review and automatically find new instances of previously identified misinformation. They are making progress, but know their systems are far from perfect.

As with hate speech, this poses difficult technical challenges. Two pieces of misinformation might contain the same claim but express it very differently, whether by rephrasing it, using a different image, or switching the format from graphic to text. And since current events change rapidly, especially in the run-up to an election, a new piece of misinformation might focus on something that wasn't even in the headlines the day before.

To better apply warning labels at scale, Facebook needed to develop new AI technologies to match near-duplications of known misinformation at scale.

Facebook has deployed SimSearchNet++, an improved image matching model that is trained using self-supervised learning to match variations of an image with a very high degree of precision and improved recall. It's deployed as part of their end-to-end image indexing and matching system, which runs on images uploaded to Facebook and Instagram.

SimSearchNet++ is resilient to a wider variety of image manipulations, such as crops, blurs, and screenshots. This is particularly important with a visuals-first platform

such as Instagram. SimSearchNet++'s distance metric is more predictive of matching, allowing us to predict more matches and do so more efficiently. For images with text, it is able to group matches at high precision using optical character recognition (OCR) verification. SimSearchNet++ improves recall while still maintaining extremely high precision, so it's better able to find true instances of misinformation while triggering few false positives. It is also more effective at grouping collages of misinformation.

When fact-checkers have identified a piece of misinformation, we want to spot copies of it even when they've been cropped or altered.

Another challenge in detecting misinformation at scale is that false claims can appear in countless variations over time. We've developed a set of systems to predict when two pieces of content convey the same meaning even though they look very different. (For example, they might have captions with completely different text that makes the same claim, such as "masks are dangerous" and "face coverings are not safe." Or they might have different images that show the same subject.)

Moreover, identifying mutations is a highly contextual problem, given that two posts about similar concepts or entities can make very different claims.