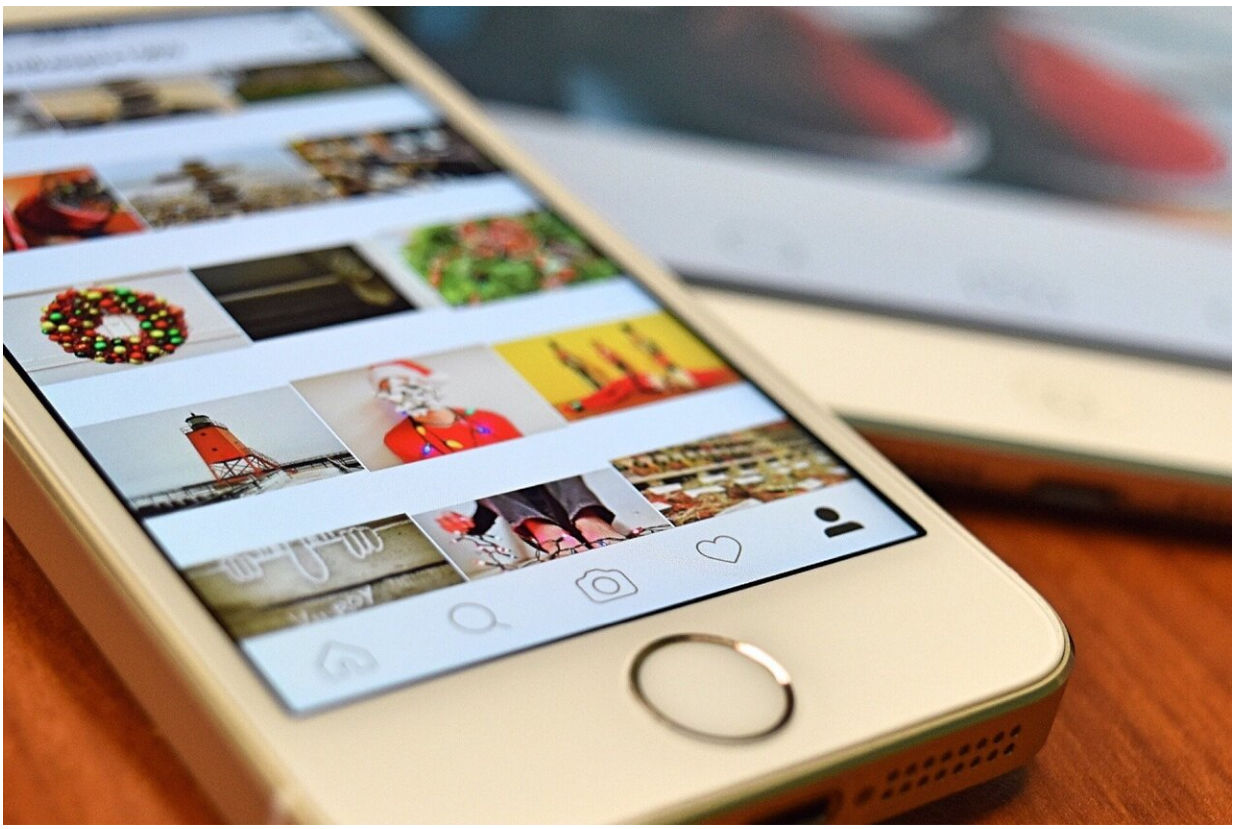
Machine learning can help to flag risky messages on Instagram while preserving users' privacy

April 17 2023



Credit: CC0 Public Domain

As regulators and providers grapple with the dual challenges of protecting younger social media users from harassment and bullying,

while also taking steps to safeguard their privacy, a team of researchers from four leading universities has proposed a way to use machine learning technology to flag risky conversations on Instagram without having to eavesdrop on them. The discovery could open opportunities for platforms and parents to protect vulnerable, younger users, while preserving their privacy.

The team, led by researchers from Drexel University, Boston University, Georgia Institute of Technology and Vanderbilt University recently published its timely work—an investigation to understand what type of data input, such as metadata, text, and image features could be most useful for machine learning models to identify risky conversations—in the *Proceedings of the Association for Computing Machinery's Conference on Human-Computer Interaction*. Their findings suggest that risky conversations can be detected by metadata characteristics, such as conversation length and how engaged the participants are.

Their efforts address a growing problem on the most popular social media platform among 13-to-21-year-olds in America. Recent studies have shown that harassment on Instagram is leading to a dramatic uptick of depression among its youngest users, particularly a rise in mental health and eating disorders among teenage girls.

"The popularity of a platform like Instagram among young people, precisely because of how it makes its users feel safe enough to connect with others in a very open way, is very concerning in light of what we now know about the prevalence of harassment, abuse, and bullying by malicious users," said Afsaneh Razi, Ph.D., an assistant professor in Drexel's College of Computing & Informatics, who was a co-author of the research.

At the same time, platforms are under increasing pressure to protect their users' privacy, in the aftermath of the Cambridge Analytica scandal

and the European Union's precedent-setting privacy protection laws. As a result, Meta, the company behind Facebook and Instagram, is rolling out end-to-end encryption of all messages on its platforms. This means that the content of the messages is technologically secured and can only be accessed by the people in the conversation.

But this added level of security also makes it more difficult for the platforms to employ automated technology to detect and prevent online risks—which is why the group's system could play an important role in protecting users.

"One way to address this surge in bad actors, at a scale that can protect vulnerable users, is automated risk-detection programs," Razi said. "But the challenge is designing them in an ethical way that enables them to be accurate, but also non-privacy invasive. It is important to put younger generation's safety and privacy as a priority when implementing security features such as end-to-end encryption in communication platforms."

The system developed by Razi and her colleagues uses machine learning algorithms in a layered approach that creates a metadata profile of a risky conversation—it's likely to be short and one-sided, for example—combined with context clues, such as whether images or links are sent. In their testing, the program was 87% accurate at identifying risky conversations using just these sparse and anonymous details.

To train and test the system, the researchers collected and analyzed more than 17,000 private chats from 172 Instagram users ages 13-21 who volunteered their conversations—more than 4 million messages in all—to assist with the research. The participants were asked to review their conversations and label each one as "safe" or "unsafe." About 3,300 of the conversations were flagged as "unsafe" and additionally categorized in one of five risk categories: harassment, sexual message/solicitation, nudity/porn, hate speech and sale or promotion of

illegal activities.

Using a random sampling of conversations from each category, the team used several machine learning models to extract a set of metadata features—things like average length of conversation, number of users involved, number of messages sent, response time, number of images sent, and whether or not participants were connected or mutually connected to others on Instagram—most closely associated with risky conversations.

This data enabled the team to create a program that can operate using only metadata, some of which would be available if Instagram conversations were end-to-end encrypted.

"Overall, our findings open up interesting opportunities for future research and implications for the industry as a whole," the team reported. "First, performing risk detection based on metadata features alone allows for lightweight detection methods that do not require the expensive computation involved in analyzing text and images. Second, developing systems that do not analyze content eases some of the privacy and ethical issues that arise in this space, ensuring user protection."

To improve upon it—making a program that could be even more effective and able to identify the specific risk type, if users or parents opt into sharing additional details of the conversations for security purposes—the team performed a similar machine learning analysis of linguistic cues and image features using the same dataset.

In this instance advanced machine learning programs combed through the text of the conversations and, knowing which contact the users had identified as "unsafe," pinpointed the words and combinations of words that are prevalent enough in risky conversations that they could be used to trigger a flag.

For analysis of the images and videos—which are central to communication on Instagram—the team used a set of programs, one that that can identify and extract text on top of images and videos, and another that can look at and generate a caption for each image. Then, using a similar textual analysis the machine learning programs again created a profile of words indicative of images and videos shared in a risky conversation.

Trained with these risky conversation characteristics, the machine learning system was put to the test by analyzing a random sampling of conversations from the larger dataset that had not been used in the profile-generation or training process. Through a combination of analyses of both metadata traits, as well as linguistic cues and image features the program was able to identify risky conversations with accuracy as high as 85%.

"Metadata can provide high-level cues about conversations that are unsafe for youth; however, the detection and response to the specific type of risk require the use of linguistic cues and image data," they report. "This finding raises important philosophical and ethical questions in light of Meta's recent push towards end-to-end encryption as such contextual cues would be useful for well-designed risk mitigation systems that leverage AI."

The researchers acknowledge that there are limitations to their research because it only looked at messages on Instagram—though the system could be adapted to analyze messages on other platforms that are subject to end-to-end encryption. They also note that the program could become even more accurate if its training were to continue with a larger sampling of messages.

But they note that this proves that this work shows that effective automated risk detection is possible, and while protecting privacy is a

valid concern, there are ways to making progress and these steps should be pursued in order to protect the most vulnerable users of these popular platforms.

"Our analysis provides an important first step to enable automated—machine learning-based—detection of online risk behavior going forward," they write. "Our system is based on reactive characteristics of the conversation however our research also paves the way for more proactive approaches to risk detection which are likely to be more translatable in the real world given their rich ecological validity."

More information: Shiza Ali et al, Getting Meta: A Multimodal Approach for Detecting Unsafe Conversations within Instagram Direct Messages of Youth, *Proceedings of the ACM on Human-Computer Interaction* (2023). DOI: 10.1145/3579608

Provided by Drexel University