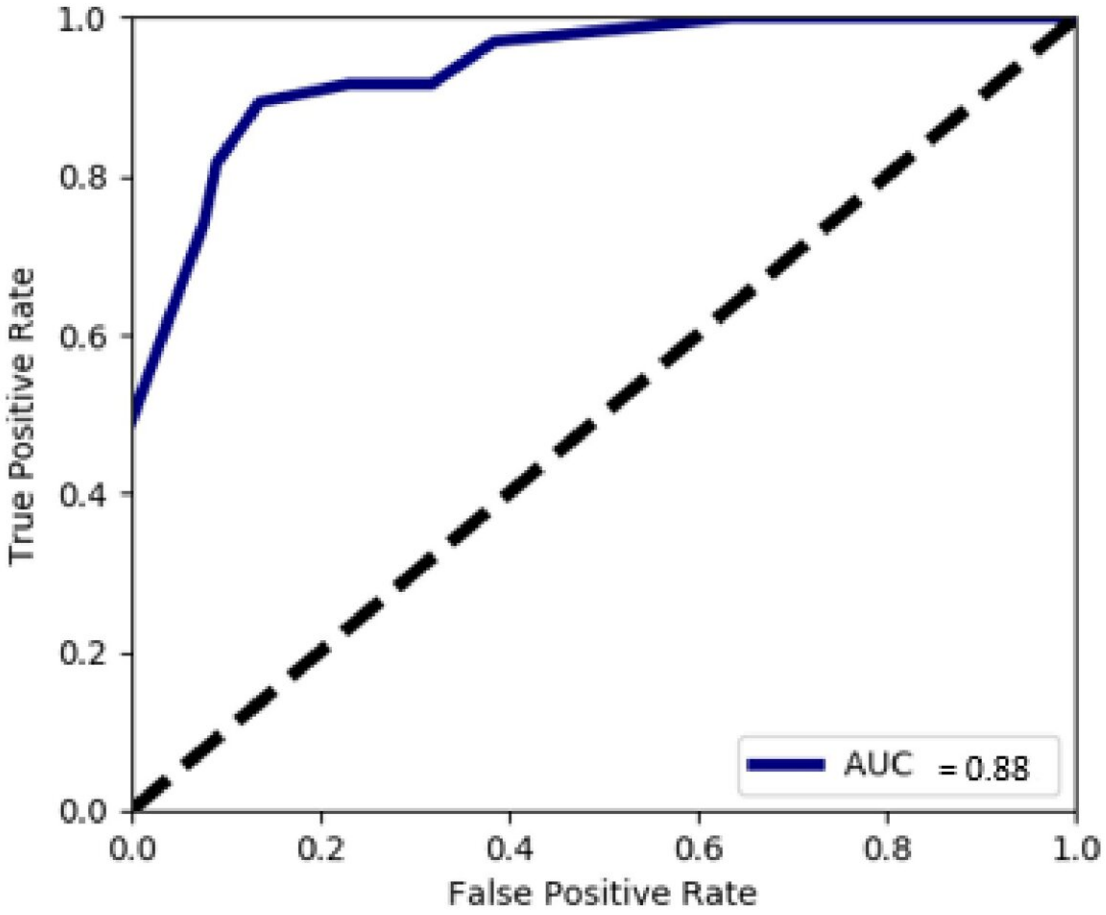
Young social media users help train machine learning program to flag sexual conversations

April 24 2023



CNN Sexual Risks Conversation Classifier ROC. Credit: *Proceedings of the ACM on Human-Computer Interaction* (2023). DOI: 10.1145/3579522

In a first-of-its-kind effort, social media researchers from Drexel University, Vanderbilt University, Georgia Institute of Technology and Boston University are turning to young social media users to help build a machine learning program that can spot unwanted sexual advances on Instagram. Trained on data from more than 5 million direct messages—annotated and contributed by 150 adolescents who had experienced conversations that made them feel sexually uncomfortable or unsafe—the technology can quickly and accurately flag risky DMs.

The project, which was recently published by the Association for Computing Machinery in its *Proceedings of the ACM on Human-Computer Interaction*, is intended to address concerns that an increase of teens using social media, particularly during the pandemic, is contributing to rising trends of child sexual exploitation.

"In the year 2020 alone, the National Center for Missing and Exploited Children received more than 21.7 million reports of child sexual exploitation—which was a 97% increase over the year prior. This is a very real and terrifying problem," said Afsaneh Razi, Ph.D., an assistant professor in Drexel's College of Computing & Informatics, who was a leader of the research.

Social media companies are rolling out new technology that can flag and remove sexually exploitative images and helps users to more quickly report these illegal posts. But advocates are calling for greater protection for young users that could identify and curtail these risky interactions sooner.

The group's efforts are part of a growing field of research looking at how machine learning and artificial intelligence be integrated into platforms to help keep young people safe on social media, while also ensuring their privacy. Its most recent project stands apart for its collection of a trove of private direct messages from young users, which

the team used to train a machine learning-based program that is 89% accurate at detecting sexually unsafe conversations among teens on Instagram.

"Most of the research in this area uses public datasets which are not representative of real-word interactions that happen in private," Razi said. "Research has shown that machine learning models based on the perspectives of those who experienced the risks, such as cyberbullying, provide higher performance in terms of recall. So, it is important to include the experiences of victims when trying to detect the risks."

Each of the 150 participants—who range in age from 13- to 21-years-old—had used Instagram for at least three months between the ages of 13 and 17, exchanged direct messages with at least 15 people during that time, and had at least two direct messages that made them or someone else feel uncomfortable or unsafe.

They contributed their Instagram data—more than 15,000 private conversations—through a secure online portal designed by the team. And were then asked to review their messages and label each conversation, as "safe" or "unsafe," according to how it made them feel.

"Collecting this dataset was very challenging due to sensitivity of the topic and because the data is being contributed by minors in some cases," Razi said. "Because of this, we drastically increased the precautions we took to preserve confidentiality and privacy of the participants and to ensure that the data collection met high legal and ethical standards, including reporting child abuse and the possibility of uploads of potentially illegal artifacts, such as child abuse material."

The participants flagged 326 conversations as unsafe and, in each case, they were asked to identify what type of risk it presented—nudity/porn, sexual messages, harassment, hate speech, violence/threat, sale or

promotion of illegal activities, or self-injury—and the level of risk they felt—either high, medium or low.

This level of user-generated assessment provided valuable guidance when it came to preparing the machine learning programs. Razi noted that most social media interaction datasets are collected from publicly available conversations, which are much different than those held in private. And they are typically labeled by people who were not involved with the conversation, so it can be difficult for them to accurately assess the level of risk the participants felt.

"With self-reported labels from participants, we not only detect sexual predators but also assessed the survivors' perspectives of the sexual risk experience," the authors wrote. "This is a significantly different goal than attempting to identify sexual predators. Built upon this real-user dataset and labels, this paper also incorporates human-centered features in developing an automated sexual risk detection system."

Specific combinations of conversation and message features were used as the input of the machine learning models. These included contextual features, like age, gender and relationship of the participants; linguistic features, such as wordcount, the focus of questions, or topics of the conversation; whether it was positive, negative or neutral; how often certain terms were used; and whether or not a set of 98 pre-identified sexual-related words were used.

This allowed the machine learning programs to designate a set of attributes of risky conversations, and thanks to the participant's assessments of their own conversations, the program could also rank the relative level of risk.

The team put its model to the test against a large set of public sample conversations created specifically for sexual predation risk-detection

research. The best performance came from its "Random Forest" classifier program, which can rapidly assign features to sample conversations and compare them to known sets that have reached a risk threshold. The classifier accurately identified 92% of unsafe sexual conversations from the set. It was also 84% accurate at flagging individual risky messages.

By incorporating its user-labeled risk assessment training, the models were also able to tease out the most relevant characteristics for identifying an unsafe conversation. Contextual features, such as age, gender and relationship type, as well as linguistic inquiry and wordcount contributed the most to identifying conversations that made young users feel unsafe, they wrote.

This means that a program like this could be used to automatically warn users, in real-time, when a conversation has become problematic, as well as to collect data after the fact. Both of these applications could be tremendously helpful in risk prevention and the prosecution of crimes, but the authors caution that their integration into social media platforms must preserve the trust and privacy of the users.

"Social service providers find value in the potential use of AI as an early detection system for risks, because they currently rely heavily on youth self-reports after a formal investigation had occurred," Razi said.

"But these methods must be implemented in a privacy-preserving matter to not harm the trust and relationship of the teens with adults. Many parental monitoring apps are privacy invasive since they share most of the teen's information with parents, and these machine learning detection systems can help with minimal sharing of information and guidelines to resources when it is needed."

They suggest that if the program is deployed as a real-time intervention,

then young users should be provided with a suggestion—rather than an alert or automatic report—and they should be able to provide feedback to the model and make the final decision.

While the groundbreaking nature of its training data makes this work a valuable contribution to the field of computational risk detection and adolescent online safety research, the team notes that it could be improved by expanding the size of the sample and looking at users of different social media platforms. The training annotations for the machine learning models could also be revised to allow outside experts to rate the risk of each conversation.

The group plans to continue its work and to further refine its risk detection models. It has also created an open-source community to safely share the data with other researchers in the field—recognizing how important it could be for the protection of this vulnerable population of social media users.

"The core contribution of this work is that our findings are grounded in the voices of youth who experienced online sexual risks and were brave enough to share these experiences with us," they wrote. "To the best of our knowledge, this is the first work that analyzes machine learning approaches on private social media conversations of youth to detect unsafe sexual conversations."

More information: Afsaneh Razi et al, Sliding into My DMs: Detecting Uncomfortable or Unsafe Sexual Risk Experiences within Instagram Direct Messages Grounded in the Perspective of Youth, *Proceedings of the ACM on Human-Computer Interaction* (2023). DOI: 10.1145/3579522