Automated detection of doxing on Twitter with over 96% accuracy

December 12 2022, by Jess Hallman

Credit: Unsplash/CC0 Public Domain

A new automated approach to detect doxing—a form of cyberbullying in which certain private or personally identifiable information is publicly shared without an individual's consent or knowledge—may help social

media platforms better protect their users, according to researchers from Penn State's College of Information Sciences and Technology.

The research on doxing could lead to more immediate flagging and removal of sensitive personal information that has been shared without the owner's authorization. To date, the research team has only studied Twitter, where their novel proposed approach uses machine learning to differentiate which tweet containing personally identifiable information is maliciously shared rather than self-disclosed.

They have identified an approach that was able to automatically detect doxing on Twitter with over 96% accuracy, which could help the platform—and eventually other social media platforms—more quickly and easily identify true cases of doxing.

"The focus is to identify cases where people collect sensitive personal information about others and publicly disclose it as a way of scaring, defaming, threatening or silencing them," said Younes Karimi, doctoral candidate and lead author on the paper. "This is dangerous because once this information is posted, it can quickly be shared with many people and even go beyond Twitter. The person to whom the information belongs needs to be protected."

In their work, the researchers collected and curated a dataset of nearly 180,000 tweets that were likely to contain doxed information. Using machine learning techniques, they categorized the data as containing personal information tied to either an individual's identity—their social security number—or an individual's location—their IP address—and manually labeled more than 3,100 of the tweets that were found to contain either piece of information.

They then further classified the data to differentiate malicious disclosures from self-disclosures. Next, the researchers examined the

tweets for common potential motivations behind disclosures, determined whether the intent was likely defensive or malicious, and indicated whether it could be characterized as doxing.

"Not all doxing instances are necessarily malicious," explained Karimi. "For example, a parent of a missing child might benignly share their private information with the desperate hope of finding them."

Next, the researchers used nine different approaches based on existing natural language processing methods and models to automatically detect instances of doxing and malicious disclosures of two types of most sensitive private information, social security number and IP address, in their collected dataset.

They compared the results and identified the approach with the highest accuracy rate, and presented their findings in November at the 25th ACM Conference on Computer-Supported Cooperative Work and Social Computing.

According to Karimi, this work is especially critical in a time when leading social media platforms—including Twitter—are conducting mass layoffs, minimizing the number of workers responsible for reviewing content that may violate the platforms' terms of service.

One platform's policy, for example, states that unless a case of doxing has clearly abusive intent, the owner of the publicly shared information or their authorized representative must contact the platform before enforcement action is taken. Under this policy, private information could remain publicly available for long periods of time if the owner of the information is not aware that it has been shared.

"While there exist some prior studies on detection of private information in general and some automated approaches for detecting cyberbullying

are applied by social media platforms, they do not differentiate self-disclosures from malicious disclosures of second- and third-parties in tweets," he said.

"Fewer people are now in charge of taking action for these manual user reports, so adding automation can help them to narrow down the most important and sensitive reports and prioritize them."

The paper is published as part of the *Proceedings of the ACM on Human-Computer Interaction*.

More information: Younes Karimi et al, Automated Detection of Doxing on Twitter, *Proceedings of the ACM on Human-Computer Interaction* (2022). DOI: 10.1145/3555167

Provided by Pennsylvania State University