A solution for moderating junk senders on WhatsApp

25 April 2022

Credit: Pixabay/CC0 Public Domain

A Rutgers researcher has developed techniques to help WhatsApp identify junk senders in public groups and automatically filter junk and spam messages for WhatsApp users.

The study, "Jettisoning Junk Messaging in the Era of End-to-End Encryption: A Case Study of WhatsApp," will be presented at The Web Conference 2022. The researchers examined 2.6 million messages from 5,051 public-politics-related WhatsApp public political WhatsApp groups in India, analyzing content, URLs and patterns of spam messages over time.

WhatsApp is the most popular mobile messaging app globally, with more than 2 billion users.

The prevalence of junk—defined as messages not of interest or suitable by administrators for a group—was much higher than researchers anticipated. According to the study, nearly one in 10 messages posted to these groups were junk messages.

"Eliminating unwanted messages is key for improving information consumption for people who are bombarded by spam and for reducing users' economic concerns," said Kiran Garimella, an assistant professor of Library and Information Science at the Rutgers School of Communication and Information. "Some junk senders aim to steal users' credit card information."

The study found the most widespread junk is advertisements for jobs, which comprised nearly 30 percent of the data set. Other junk messages included "click and earn," which encourages clicks to a URL and promises a reward. 7.7 percent of junk messages offered items for sale, while 7.5 percent offer a gift in return for referrals of users to an online service subscription, and consist mostly of a URL to click.

The researchers developed methods for moderating WhatsApp public groups. Unlike messaging systems such as email and Twitter, WhatsApp can't read or moderate user content because of end-to-end encryption. While this ensures user privacy, WhatsApp's inability to moderate content means spam and unwanted messages posted by junk senders can impact user experience on the platform.

According to the study, spam-senders post across many groups and typically appear and disappear several times to avoid being detected and removed by administrators.

Junk senders spread the same spam messages over a few "active" days. Garimella said this strategy might improve the visibility of junk by providing a longer 'shelf life' in the recent messages.

A key indicator of junk are URLs and phone numbers. Nearly 90 percent of junk messages contained a phone number, a URL or both (in contrast to 36 percent for nonjunk). The researchers created a coding model to

automatically detect junk using URLs and phone numbers. This can assist WhatsApp administrators in quickly flagging and removing these messages, they said.

From a user stand-point, the researchers created a model in which users encode a signal that detects whether a message contains a phone number, a URL, both or neither.

"Our methods are very practical and applicable," said Garimella. "WhatsApp can apply them to stop the spread of spam in their groups, and our techniques can be used on the platform centrally while still respecting the end-to-end encryption guarantees WhatsApp offers users to protect their privacy."

As part of a broad effort to reduce the spam on WhatsApp public groups, Garimella and his co-authors are sharing their annotated dataset and code with WhatsApp and making it publicly available for other researchers to use.

The research was published on *arXiv*.

More information: Pushkal Agarwal et al, Jettisoning Junk Messaging in the Era of End-to-End Encryption: A Case Study of WhatsApp, arXiv:2106.05184 [cs.CR], arxiv.org/abs/2106.05184

Provided by Rutgers University