Researchers develop method that boosts performance of moderation models on live platforms

December 7 2023

Credit: Pixabay/CC0 Public Domain

Twitch. Some see it as a fun online community of gamers and good-natured e-sports fandom. For others, it's a perilous stream of potentially toxic content and hate speech.

In the ever-evolving landscape of digital communication, the real-time nature of messages on live-stream platforms like Twitch and YouTube Live brings with it unique challenges for content moderation. At present, effective tools for moderating content in live streams are lacking because existing models have been trained on non-real-time social media platforms like Facebook or Twitter.

Research Assistant Dong-Ho Lee and Principal Scientist Jay Pujara, both from USC Viterbi's Information Sciences Institute (ISI), set out to change that. They have developed an innovative method that boosts the performance of moderation models on live platforms by 35%.

Getting in sync

Pujara said, "If I post something on Twitter or Reddit, someone might respond hours or days later. But if we're looking at Twitch, it's a very different environment. People are sending messages every second."

It all comes down to timing. Twitter, Facebook, and Reddit are asynchronous—where users post their thoughts, but the responses are not immediate. On the other hand, Twitch, YouTube Live, and other live-streaming platforms are synchronous—which is the equivalent of being in a live conversation.

In conversations on asynchronous platforms, thoughts are typically grouped into a structure of threads that allow for conversational context. And users have no time constraints, so they can comment with better thought-out responses. Whereas on synchronous platforms, thoughts are presented in real time, consecutively, with no structure to indicate

context. The fast-paced nature encourages quick responses and multiple short comments.

A first-of-its-kind approach

Seeing this gap in the research, Lee and Pujara conducted the first NLP study of detecting norm violations in live-stream chat.

"Norm violations" refer to instances where users on online platforms breach the established rules or guidelines for acceptable behavior. Pujara explained, "Typically there will be a set of rules that are published when you join [a live stream], and there are moderators who are trying to figure out if people are breaking these rules. Are you harassing someone? Are you trying to change the topic? Are you sending spam messages?"

The team of authors, including ISI Ph.D. students Justin Cho and Woojeong Jin, and Jonathan May, a research associate professor at the USC Viterbi Thomas Lord Department of Computer Science, used a dataset of 4,583 norm-violating comments on Twitch that were moderated by human channel moderators.

"They gathered chat rules of each Twitch streamer, held iterative meetings to categorize types of norm violations, and managed annotators in labeling various live streaming sessions to analyze norm violations in Twitch," said Lee, who continued, "This involved a significant joint effort between various industry partners and academic institutions for the first study of norm violations in live-stream chat."

Bring in the humans… and the details

Pujara said, "An interesting thing about the way we did this is that, to get

the label for the data, we crowdsourced. We had humans label it and then those humans would basically get three levels of detail. So, we were giving them progressively more information to be able to evaluate what's going on."

What kind of details were provided? The team designed a process that would determine the impact of varying levels of context surrounding the moderated comment. For example, did the chat history have an impact—either the commenter's last message before the moderated content or the broader chat around the time of the moderated comment? What was happening on the video as the comment was posted? And was there any external knowledge related to the content that is specific to the comment (i.e., particular emojis or slang within the channel).

Context is crucial

Turns out, when it comes to moderating live streams, context counts.

Pujara explains, "You can improve the quality of the moderation by using different amounts of information. And so, if you're designing an automated moderation system for Twitch, you really need to think about what the right context is to interpret what people are saying."

The team used this information, identified the informational context that best helped the human moderators, and trained models to identify norm-violations by leveraging this contextual information. Their results showed that contextual information can boost model moderation performance by 35%.

Pujara and Lee's paper, Analyzing Norm Violations in Live-Stream Chat, is available on the *arXiv* preprint server and will be presented at the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 23), which takes place in Singapore from December 6—10,

2023.

Lee said, "I'm thrilled to be participating in EMNLP and present our research. Moreover, I'm eager to present two additional papers—Temporal Knowledge Graph Forecasting Without Knowledge Using In-Context Learning and Making Large Language Models Better Data Creators—that I've worked on with Jay.

More information: Jihyung Moon et al, Analyzing Norm Violations in Live-Stream Chat, *arXiv* (2023). DOI: 10.48550/arxiv.2305.10731

Provided by University of Southern California