How AI Can Moderate Comments, Eliminate Trolls

- *6/19/2017 by Ariella Brown 41*

(Image: pixone/iStockphoto)

As any online publisher or even blogger knows, reader comments are a great sign of engagement. But the comments also open up the possibility for spam and verbal abuse. Consequently, many publishers restrict comments or keep them on hold until a human moderator can assess whether or not they should be posted. Now AI can help speed up that process tremendously.

The problem with human moderators is that they have human limitations that cannot keep up with a huge influx of comments. That was the problem the *New York Times* faced in balancing reader comments demand with an editorial standard of civility for all published comments. Its solution was a partnership with an incubator owned by Google's parent company, Alphabet.

Back in September, the Times announced the partnership in an article that set out the challenge faced by its 14 moderators tasked with reviewing the comments on the 10% of articles that do allow them. That alone amounted to 11,000 comments a day. As the article invited readers to try their hand at moderating, it was entitled Approve or Reject: Can You Moderate Five New York Times Comments?

I took the test. The official summary of my results were: "You moderated 4 out of 5 comments as the Community desk would have, and it took you 81 seconds. Moderating the 11,000 comments posted to nytimes.com each day would take you 49.5 hours."

That summation was followed by this: "Don't feel too bad; reviewing all of these comments takes us a long time, too." According to my calculations, however, the Times actually allows more time for their moderators. Given 14 people working 8 hours a day, the number of working hours each day would be 112, or more than twice the number of hours they said would be required for my rate of moderation.

That investment of so many hours is not something they regret, as they regard it as a requisite part of building up "one of the best communities on the web." However, they recognize that needs must dictate a new approach. That's where the machine assistance enters into the picture, enabling the same

number of humans to effectively moderate a much larger number of comments and reduce the delay for reviewing time.

Flash forward to June 13, 2017, and the Gray Lady herself announces: The Times Sharply Increases Articles Open for Comments, Using Google's Technology. Using what they call "Moderator," the digital paper now allows comment on "all our top stories" for a span of eight hours and extending 24 hours for "comments in both the News and Opinion sections."

This comment expansion is made possible by the addition of Jigsaw's machine learning technology that can "prioritize comments for moderation," and even let comments post without human intervention. "Its judgments are based on more than 16 million moderated Times comments, going back to 2007." That formed the basis for Jigsaw's "machine learning algorithm that predicts what a Times moderator might do with future comments."

As with most machine learning projects, Moderator will evolve. Initially, the majority of comment are to be assigned a "summary score." That is based on "three factors: their potential for obscenity, toxicity, and likelihood to be rejected." But as the system continues to learn, and the Times editors believes they can rely on it, the plan is to advance to allowing automated moderation for the comments that show strong indications of qualifying as approved.

FastCompany explained that in this scoring system, zero is the best and 100 is the worst comment ranking. Either extreme wouldn't need further moderation, but the human moderators can work more efficiently if they read over the comments that fall into a certain range. Bassey Etim, the Community Editor for nytimes.com, who wrote the Times piece introducing Moderator, expects that would result in an eight to 10 times increase of efficiency for the humans involved (and that explains how the online paper can increase comments from 10% to 80% of its content).

Certainly, that algorithm benefits the Times in fostering engagement without allowing it to run amok. But what does Jigsaw gain from the partnership aside from a huge amount of data to play with in developing its machine learning algorithm? According to FastCompany it's about common goals. The article quotes Patricia Georgiou, Jigsaw's head of partnership, saying that this media outlet like a few others it selected to work with "aligned with our goal," namely, to find a way to "improve online conversation."

Georgiou also clarified that it is challenging to train the machine learning to recognize "what is a toxic comment," as well as the attributes within "a comment that would cause somebody leave the conversation." That's why the algorithm requires such a large body of data that extends beyond the Times' comments to include input from other publishers that have contributed to the project.

One interesting note about the state of online article comments emerges in contrasting the FastCompany article and the Times article that address the same topic and were published the same day. The former has no comments; the latter has close to 500.

While I didn't read through all the comments, I did read through a number of them. Some pointed out that the machine learning will absorb whatever biases the human moderators may have if it is trained on what they have approved or disproved. That is a valid point, though I think it is naïve for anyone to consider a media outlet to be completely objective.

I find that those comments that extend the conversation on the topic raised in the article really enhance the reader experience. It's a shame to have that spoiled by people who cannot distinguish between reasoned argument and ad hominem attacks.

In my view, even if the AI system is influenced human biases, it is still a good thing if it allows the community members to comment. What's your view? You can comment here and be heard.