When computers learn to swear: Using machine learning for better online conversations

Imagine trying to have a conversation with your friends about the news you read this morning, but every time you said something, someone shouted in your face, called you a nasty name or accused you of some awful crime. You'd probably leave the conversation. Unfortunately, this happens all too frequently online as people try to discuss ideas on their favorite news sites but instead get bombarded with toxic comments.

Seventy-two percent of American internet users have witnessed harassment online and nearly half have personally experienced it. Almost a third self-censor what they post online for fear of retribution. According to the same report, online harassment has affected the lives of roughly 140 million people in the U.S., and many more elsewhere.

This problem doesn't just impact online readers. News organizations want to encourage engagement and discussion around their content, but find that sorting through millions of comments to find those that are trolling or abusive takes a lot of money, labor, and time. As a result, many sites have shut down comments altogether. But they tell us that isn't the solution they want. We think technology can help.

Today, Google and Jigsaw are launching Perspective, an early-stage technology that uses machine learning to help identify toxic comments. Through an API, publishers—including members of the Digital News Initiative—and platforms can access this technology and use it for their sites.

We've been testing a version of this technology with The New York Times, where an entire team sifts through and moderates each comment before it's posted—reviewing an average of 11,000 comments every day. That's a lot of comments. As a result the Times has comments on only about 10 percent of its articles. We've worked together to train models that allows Times moderators to sort through comments more quickly, and we'll work with them to enable comments on more articles every day.

Where we go from here

Perspective joins the TensorFlow library and the Cloud Machine Learning Platform as one of many new machine learning resources Google has made available to developers. This technology is still developing. But that's what's so great about machine learning—even though the models are complex, they'll improve over time. When Perspective is in the hands of publishers, it will be exposed to more comments and develop a better understanding of what makes certain comments toxic.

While we improve the technology, we're also working to expand it. Our first model is designed to spot toxic language, but over the next year we're keen to partner and deliver new models that work in languages other than English as well as models that can identify other perspectives, such as when comments are unsubstantial or off-topic.

In the long run, Perspective is about more than just improving comments. We hope we can help improve conversations online.