The Wikipedia community, the free encyclopaedia that is built from a model of openly editable content, is notorious for its toxicity. The issue was so bad that the number of active contributors or editors—those that made one edit per month—had fallen by 40 percent during an eight-year period. Even though there's not one solution to combat this issue, Wikimedia Foundation, the nonprofit that supports Wikipedia, decided to useartificial intelligence to learn more about the problem and consider ways to combat it.

Collaboration with Wikimedia Foundation and Jigsaw to Stop Abusive Comments

In one effort to stop the trolls, Wikimedia Foundation partnered with Jigsaw (the tech incubator formerly known as Google Ideas) on a research project called Detox using machine learning to flag comments that might be personal attacks. This project is part of Jigsaw's initiative to build open-source AI tools to help combat harassment on social media platforms and web forums.

The first step in the project was to train the machine learning algorithms using 100,000 toxic comments from Wikipedia Talk pages that had been identified by a 4,000-person human team where every comment had ten different human reviewers. This annotated dataset was one of the largest ever created that looked at online abuse. Not only did these include direct personal attacks, but also third-party and indirect personal attacks ("You are horrible." "Bob is horrible." "Sally said Bob is horrible.") After training, the machines could determine a comment was a personal attack just as well as three human moderators.

Then, the project team had the algorithm review 63 million English Wikipedia comments posted during a 14-year period between 2001 to 2015 to find patterns in the abusive comments. What they discovered was outlined in the Ex Machina: Personal Attacks Seen at Scale paper:

- More than 80% of all comments characterised as abusive were made by more than 9,000 people who made less than five abusive comments in a year rather

than an isolated group of trolls.
- Nearly 10% of all attacks were made by just 34 users.
- Anonymous users made up 34% of all comments left on Wikipedia.
- More than half of the personal attacks are being carried out by registered users although anonymous users were six times more likely to launch personal attacks. (There are 20 times more registered users than anonymous users.)

Now that the algorithms have created more clarity about who is contributing to the community's toxicity, Wikipedia can figure out the best way to combat the negativity. Although human moderation is likely still needed, algorithms can help sort through the comments and flag those that require human involvement.

Objective Revision Evaluation Service (ORES System)

Another reason for the significant decline in editors to Wikipedia is thought to be the organisation's complex bureaucracy as well as its harsh editing tactics. It was common for first-time contributors/editors to have an entire body of work wiped out with no explanation. One way they hope to fight this situation is with the ORES system, a machine that acts as an editing system powered by an algorithm trained to score the quality of changes and edits. Wikipedia editors used an online tool to label examples of past edits, and that was how the algorithm was taught the severity of errors. The ORES system can direct humans to review the most damaging edit and determine the calibre of mistakes—rookie mistakes are treated more appropriately as innocent.

AI to Write Wikipedia Articles

Well, AI can do "OK" writing Wikipedia articles, but you have to start somewhere, right? A team within Google Brain taught software to summarise info on web pages and write a Wikipedia-style article. It turns out text summarization is more difficult than most of us thought. Google Brain's efforts to get a machine to summarise content is slightly better than previous attempts, but there is still work to be done before a machine can write with the cadence and flair humans can. It turns out we're not quite ready to have a machine automatically generate Wikipedia entries, but there are efforts underway to get us there.

While the use cases for artificial intelligence in the operations of Wikipedia are still being optimised, machines can undoubtedly help the organisation analyse the vast amount of data they generate daily. Better information and analysis can help Wikipedia create successful strategies to troubleshoot negativity from its community and recruitment issues for its contributors.

Written by

Bernard Marr

Bernard Marr is a bestselling author, keynote speaker, and advisor to companies and governments. He has worked with and advised many of the world's best-known organisations. LinkedIn has recently ranked Bernard as one of the top 10 Business Influencers in the world (in fact, No 5 - just behind Bill Gates and Richard Branson). He writes on the topics of intelligent business performance for various publications including Forbes, HuffPost, and LinkedIn Pulse. His blogs and SlideShare presentation have millions of readers.