AI service will boost Wikipedia's hunt for damaging edits

December 2 2015, by Nancy Owano

Credit: Illustration by Mun May Tee, freely licensed under CC BY-SA 4.0.

Bears can fly. Shoes taste like lemons. Swimming causes rabies.

Not the kind of sentences one wants to see on an information source and [editors](#) work hard to strike such statements out.

Announced on Monday in the [Wikimedia blog](), "Today, we're announcing the release of a new artificial intelligence service designed to improve the way editors maintain the quality of Wikipedia. This service empowers Wikipedia editors by helping them discover damaging edits and can be used to immediately 'score' the quality of any Wikipedia article. We've made this [artificial intelligence]() available as an open web service that anyone can use."

Tom Simonite, *MIT Technology Review*'s San Francisco bureau chief, referred to it as software "trained to know the difference between an honest mistake and intentional [vandalism]()."

Wikipedia's AI can automatically spot bad edits, said *Engadget*, and one might say farewell to spammers. Another beneficial result: the AI service could make the website a lot friendlier to [newbie]() contributors, said Mariella Moon on Wednesday.

This [service]() is a step in a needed direction. After all, according to the blog announcement, Wikipedia is edited about half a million times per day—a "firehose of new content," as the blog put it, which needs Wikipedians' review.

The Wikimedia Foundation has created ORES, the [Objective Revision Evaluation Service](). (The Wikimedia Foundation is the nonprofit organization that supports Wikipedia, the collection of free, collaborative knowledge.)

The blog said it functions like a pair of X-ray specs, referring to the toy kind of specs seen in novelty shops, for highlighting anything suspicious.

"This allows editors to triage them from the torrent of new edits and review them with increased scrutiny," said the blog.

How does ORES know if the edit is intentionally damaging and not just an unintentional slip? Moon said it uses Wiki teams' article-quality assessments as examples. "The AI's 'false' or not damaging probability score for it is 0.0837, while its 'true' or damaging probability score is 0.9163."

Why ORES matters: Some of the older tools could not tell the difference between a malicious edit and an honest error.

Another key plus delivered by ORES is that it will help avoid turning off newbies who sincerely want to participate. According to the blog, "we hope to pave the way for experimentation with new tools and processes that are both efficient and welcoming to new editors."

While some automated tools worked to keep up the quality of Wikipedia, they "also (inadvertently) exacerbated the difficulties that newcomers experience when learning about how to contribute to Wikipedia." commented the blog. "These tools encourage the rejection of all new editors' changes as though they were made in bad faith, and that type of response is hard on people trying to get involved in the movement."

What is more, "Our research shows that the retention rate of good-faith new editors took a nose-dive when these quality control tools were introduced to Wikipedia."

Simonite said that this was an effort to make editing Wikipedia less psychologically bruising. He reported that ORES is up to speed on the English, Portuguese, Turkish, and Farsi versions of Wikipedia so far.

 More information: blog.wikimedia.org/2015/11/30/ … ligence-x-ray-specs/