NO TRUE SCOTSMAN

How a Scots Wikipedia scandal highlighted AI's data problem

REUTERS/YVES HERMAN

A Reddit user discovered that most articles aren't written in Scots, raising questions for AI models trained on Wikipedia data.

FROM OUR OBSESSION
Beyond Silicon Valley
The next big battles in tech are happening outside the Bay Area.

By **Nicolás Rivero**
Reporter

56 minutes ago

Most of the English language technology you use on a daily basis—voice assistants, spell checkers, translation tools, search functions—share a common origin story. They're built using AI language models, and many of those models are trained on millions of Wikipedia articles.

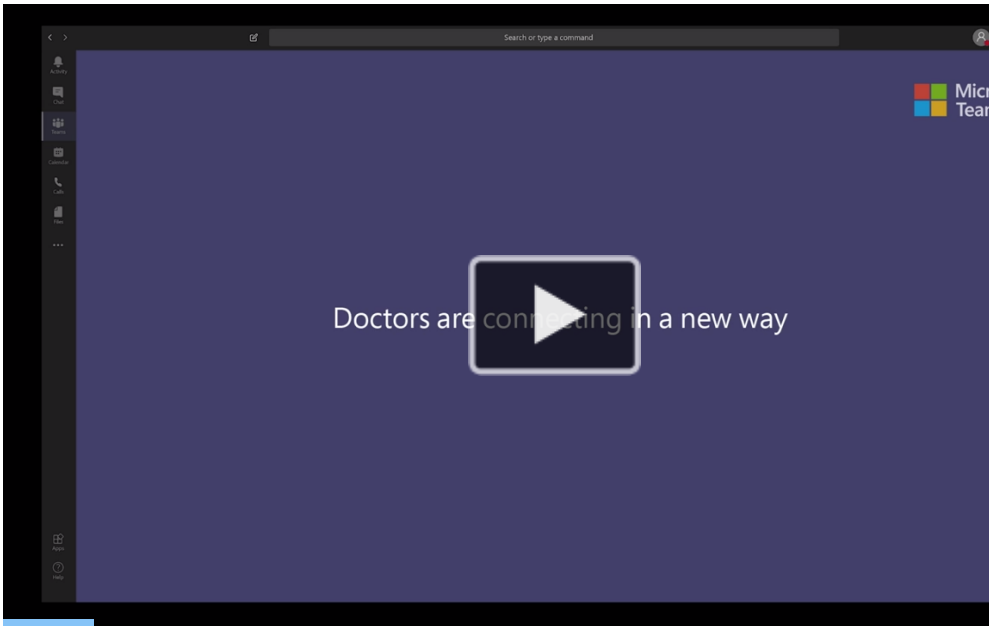
But a bizarre discovery made this week by a Scottish Reddit sleuth has highlighted a worrisome problem for that data pipeline. Most of the Scots language edition of Wikipedia was written by an American teenager who doesn't actually speak the language. Instead, the teen

wrote tens of thousands of articles in English with a put-on Scottish accent, ignoring actual Scots grammar and vocabulary.

For a low-resource language like Scots, which has few digital archives of written text to pull from, it could mean that some models base their entire understanding of the language on the phony version written in the Scots Wikipedia. That limits the amount of access native speakers have to tech tools in their language.

"I don't think people necessarily realize how important Wikipedia is for training all of our language technologies," said David Yarowsky, a computer science professor at Johns Hopkins University. "When these problems crop up, it really is impacting our ability to do a high quality job on the technologies that these communities want."

The Scots Wikipedia is a rather unique example of a misguided editor dedicating a decade to writing out articles in what they believed to be (but absolutely was not) genuine Scots. More often, Yarowsky says, the issue is that Wikipedia editors decide to fill out a language edition with machine-translated text that is not corrected by fluent speakers. The second largest Wikipedia edition behind English—the Cebuano edition, catering to just 16 million speakers primarily in the Philippines—was almost entirely written by a single bot. Unlike Scots speakers, Cebuano speakers aren't likely to have language AI tools available in another language they speak just as well.



A bot-based translation approach can create a vicious cycle when future algorithms use those Wikipedia pages as training data. "You're basically learning from a bad version of what you already have," Yarowsky said. "If we train machine translation systems on machine translation output, we are just reinforcing existing failings like an echo chamber."

Shoddy Wikipedia pages can also make it harder to verify that language algorithms actually work. Jeanna Matthews, a computer science professor at Clarkson University, says that models are often tested against Wikipedia data, meaning that AI tools built for languages with high-quality entries, like English, keep improving, while others don't.

"The people who develop these tools can work out the bugs and kinks for languages that are well-represented," she said. "It's a snowball effect, and the advantages for these languages get further amplified."

There is, however, a clear way to break the cycle: "One of the best things that the community can do is write a lot in their language and post it online, whether as news or stories or Wikipedia articles or discussion forums," Yarowsky said. That way language researchers will have good text data from fluent speakers to incorporate into their models.

Scots speakers have begun to organize to do just that. Seventy-four people have joined a new "Scots Wikipedia editors" Facebook group, and they've scheduled their first "editathon" to begin rewriting articles for Aug. 30.