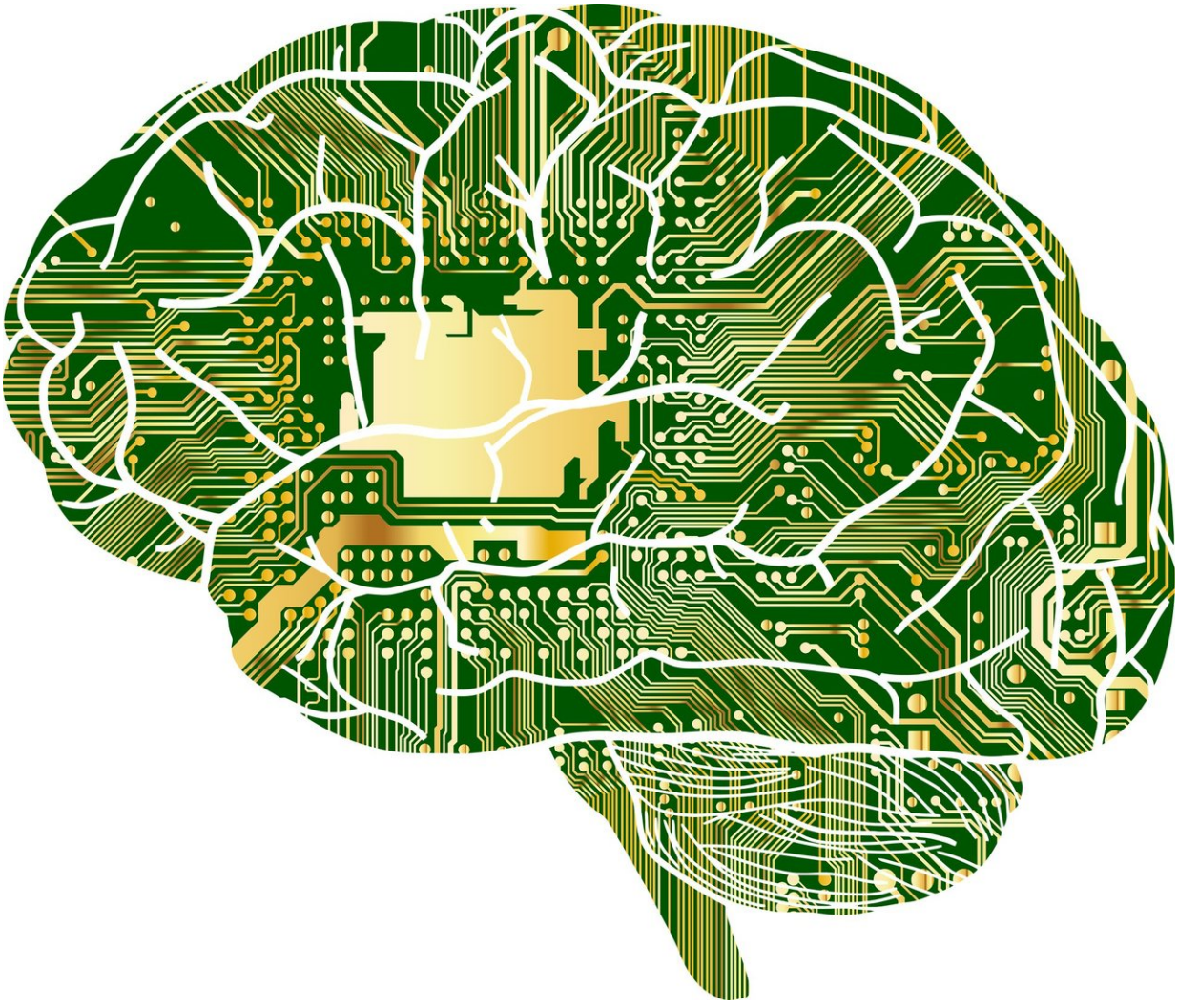
Humans and AI team up to improve clickbait detection

August 28 2019



Credit: CC0 Public Domain

Humans and machines worked together to help train an artificial intelligence—AI—model that outperformed other clickbait detectors, according to researchers at Penn State and Arizona State University. In addition, the new AI-based solution was also able to tell the difference between clickbait headlines that were generated by machines—or bots—and ones written by people, they said.

In a study, the researchers asked people to write their own clickbait—an interesting, but misleading, news headline designed to attract readers to click on links to other online stories. The researchers also programmed machines to generate artificial clickbaits. Then, the headlines made by both people and machines were used as data to train a clickbait-detection algorithm.

The resulting algorithm's ability to predict clickbait headlines was about 14.5 percent better than other systems, according to the researchers, who released their findings today (Aug. 28) at the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis (ASONAM) at Vancouver, Canada.

Beyond its use in clickbait detection, the team's approach may help improve machine learning performance in general, said Dongwon Lee, the principal investigator of the project and an associate professor in the College of Information Sciences and Technology. Lee is also an affiliate of Penn State's Institute for CyberScience (ICS), which provides Penn State researchers access to supercomputing resources.

"This result is quite interesting as we successfully demonstrated that machine-generated clickbait training data can be fed back into the training pipeline to train a wide variety of machine learning models to have improved performance," said Lee. "This is the step toward addressing the fundamental bottleneck of supervised machine learning that requires a large amount of high-quality training data."

According to Thai Le, a doctoral student in the College of Information Sciences and Technology, Penn State, one of the challenges confronting the development of clickbait detection is the lack of labeled data. Just like people need teachers and study guides to help them learn, AI models need data that is labeled to help them learn to make the correct connections and associations.

"One of the things we realized when we started this project is that we don't have many positive data points," said Le. "In order to identify clickbait, we need to have humans label that training data. There is a need to increase the amount of positive data points so that, later on, we can train better models."

While finding clickbait on the internet can be easy, the many variations of clickbait add another layer of difficulty, according to S. Shyam Sundar, James P. Jimirro Professor of Media Effects and co-director of the Media Effects Research Laboratory in the Donald P. Bellisario College of Communications, and an ICS affiliate.

"There are clickbaits that are lists, or listicles; there are clickbaits that are phrased as questions; there are ones that start with who-what-where-when; and all kinds of other variations of clickbait that we have identified in our research over the years," said Sundar. "So, finding sufficient samples of all these types of clickbait is a challenge. Even though we all moan about the number of clickbaits around, when you get around to obtaining them and labeling them, there aren't many of those datasets."

According to the researchers, the study revealed differences in how people and machines approached the creation of headlines. Compared to the machine-generated clickbait, headlines generated by people tended to have more determiners—words such as "which" and "that"—in their headlines.

Training also seemed to prompt differences in clickbait creation. For example, trained writers, such as journalists, tended to use longer words and more pronouns than other participants. Journalists also were likely to use numbers to start their headlines.

The researchers plan to use these findings to guide their investigations into a more robust fake-news detection system, among other applications, according to Sundar.

"For us, clickbait is just one of many elements that make up fake news, but this research is a useful preparatory step to make sure we have a good clickbait detection system set up," said Sundar.

To find human clickbait writers for the study, the researchers recruited journalism students and workers from Amazon Turk, an online crowdsource site. They recruited 125 students and 85 workers from the site. The participants first read a definition of clickbait and then were asked to read a short—about 500 words—article. The participants were then asked to write a clickbait headline for each article.

The machine-generated clickbait headlines were developed by using a machine learning model called a Variational Autoencoders—or VAE—generative model, which relies on probabilities to find patterns in data.

The researchers tested their algorithm against top-performing systems from Clickbait Challenge 2017, an online clickbait detection competition.

Provided by Pennsylvania State University

Citation: Humans and AI team up to improve clickbait detection (2019, August 28) retrieved 4

May 2024 from https://techxplore.com/news/2019-08-humans-ai-team-clickbait.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.