Scientists teach neural network to identify a writer's gender

April 27 2018

A team of researchers from the National Research Nuclear University MEPhI, the National Research Center Kurchatov Institute and the Voronezh State University has developed a new learning algorithm that allows a neural network to identify a writer's gender by the written text on a computer with up to 80 percent accuracy.

This is a new development in the field of computational linguistics. The research was funded by a Russian Science Foundation grant. The findings were published in the *Procedia Computer Science* journal.

Many scientific studies show that writing style can reflect certain characteristics of a writer – gender, physiological personality traits, and level of education. Speech patterns are a valuable psycho-diagnostic tool, and are often used by human resources professionals and security services.

By analyzing a person's speech, researchers can diagnose certain illnesses such as dementia and depression, and the person's inclination toward suicidal behavior. The demand for identifying certain characteristics of a writer's personality is increasing against the backdrop of the development of internet communications—companies want to know which demographics like their products and services.

Using the numerical values for various parameters in a text, researchers in this area (linguists, psychologists, IT experts) have created mathematical models to identify certain traits in the writer's personality.

Using neural networks, the researchers analyzed the effectiveness of various machine-learning algorithms for text analysis.

During the research, the scientists compared the accuracy of gender identification by text based on two types of data-driven modeling: first, machine-learning algorithms (such as a support vector machine and gradient boosting), and, second, a deep learning neural network (such as convolutional neural networks and the long short-term memory recurrent neural networks).

"Using these advanced neural network models, we have achieved great results in identifying the gender of the writer based on text, under conditions in which the author is not attempting to hide his/her gender," said Alexander Sboyev, assistant professor at MEPhI. "Our next step is to teach the neural network to identify the gender of a writer who is deliberately trying to hide it."

Thus, in the following texts, originally published on dating websites, the neural network easily identified the writer's gender 10 out of 10 times, despite the fact that authors were free to sign their texts with a name typical of the opposite gender.

This text was written by a female: "I am a handsome, fit 30-year-old man. I have a high-paying job at a large oil and gas company. I live in my own flat in Moscow, and also own a small but nice house in an Italian village. I am into sports, mainly football. I love going out on weekends, I can't stand homebodies. My perfect girl would be modest and beautiful, and would have an attractive body, based on today's standards. She would share my interests and would not be jealous or try to make me jealous. In the future, I do not plan to be the sole provider in a family, as I believe that when it comes to families, both men and women must earn the money. I would like to have separate budgets as well. I will not tolerate cheating."

This text was written by a male: "Hello! I am very angry, very! Why do you keep treating us like this?! We are people, too, all of us are equal! Are you sexist? I will not tolerate this anymore! I'm going to smash your car into pieces; I will spray paint all over it. You just wait, you monster. It sucks to be you."

This research indicated that the approach based on using convolutional neural networks and methods of deep learning to identify a writer's gender, is the most optimal. The team of researchers is currently working on identifying a writer's age.

Provided by National Research Nuclear University