Machine learning tool could provide unexpected scientific insights into COVID-19

28 April 2020

Berkeley Lab researchers (clockwise from top left) Kristin Persson, John Dagdelen, Gerbrand Ceder, and Amalie Trewartha led development of COVIDScholar, a text-mining tool for COVID-19-related scientific literature. Credit: Berkeley Lab

A team of materials scientists at Lawrence Berkeley National Laboratory (Berkeley Lab) - scientists who normally spend their time researching things like high-performance materials for thermoelectrics or battery cathodes—have built a text-mining tool in record time to help the global scientific community synthesize the mountain of scientific literature on COVID-19 being generated every day.

The tool, live at covidscholar.org, uses natural language processing techniques to not only quickly scan and search tens of thousands of research papers, but also help draw insights and connections that may otherwise not be apparent. The hope is that the tool could eventually enable "automated science."

"On Google and other search engines people search for what they think is relevant," said Berkeley Lab scientist Gerbrand Ceder, one of the project leads. "Our objective is to do information extraction so that people can find nonobvious information and relationships. That's the whole idea of machine learning and natural language processing that will be applied on these datasets."

COVIDScholar was developed in response to a March 16 call to action from the White House Office of Science and Technology Policy that asked artificial intelligence experts to develop new data and text mining techniques to help find answers to key questions about COVID-19.

The Berkeley Lab team got a prototype of COVIDScholar up and running within about a week. Now a little more than a month later, it has collected over 61,000 research papers—about 8,000 of them specifically about COVID-19 and the rest about related topics, such as other viruses and pandemics in general—and is getting more than 100 unique users every day, all by word of mouth.

And there are more papers added all the time—200 new journal articles are being published every day on the coronavirus. "Within 15 minutes of the paper appearing online, it will be on our website," said Amalie Trewartha, a postdoctoral fellow who is one of the lead developers.

This week the team released an upgraded version ready for public use—the new version gives researchers the ability to search for "related papers" and sort articles using machine-learning-based relevance tuning.

The volume of research in any scientific field, but especially this one, is daunting. "There's no doubt we can't keep up with the literature, as scientists," said Berkeley Lab scientist Kristin Persson, who is co-leading the project. "We need help to find the

relevant papers quickly and to build correlations between papers that may not, on the surface, look like they're talking about the same thing."

The team has built automated scripts to grab new papers, including preprint papers, clean them up, and make them searchable. At the most basic level, COVIDScholar acts as a simple search engine, albeit a highly specialized one.

"Google Scholar has millions of papers you can search through," said John Dagdelen, a UC Berkeley graduate student and Berkeley Lab researcher who is one of the lead developers. "However, when you search for 'spleen' or 'spleen damage' - and there's research coming out now that the spleen may be attacked by the virus—you'll get 100,000 papers on spleens, but they're not really relevant to what you need for COVID-19. We have the largest single-topic literature collection on COVID-19."

In addition to returning basic search results, COVIDScholar will also recommend similar abstracts and automatically sort papers in subcategories, such as testing or transmission dynamics, allowing users to do specialized searches.

Now, after having spent the first few weeks setting up the infrastructure to collect, clean, and collate the data, the team is tackling the next phase. "We're ready to make big progress in terms of the natural language processing for 'automated science,'" Dagdelen said.

For example, they can train their algorithms to look for unnoticed connections between concepts. "You can use the generated representations for concepts from the machine learning models to find similarities between things that don't actually occur together in the literature, so you can find things that should be connected but haven't been yet," Dagdelen said.

Another aspect is working with researchers in Berkeley Lab's Environmental Genomics and Systems Biology Division and UC Berkeley's Innovative Genomics Institute to improve COVIDScholar's algorithms. "We're linking up the unsupervised machine learning that we're doing with what they've been working on, organizing all the information around the genetic links between diseases and human phenotypes, and the possible ways we can discover new connections within our own data," Dagdelen said.

The entire tool runs on the supercomputers of the National Energy Research Scientific Computing Center (NERSC), a DOE Office of Science user facility located at Berkeley Lab. That synergy across disciplines—from biosciences to computing to materials science—is what made this project possible. The online search engine and portal are powered by the Spin cloud platform at NERSC; lessons learned from the successful operations of the Materials Project, serving millions of data records per day to users, informed development of COVIDScholar.

"It couldn't have happened somewhere else," said Trewartha. "We're making progress much faster than would've been possible elsewhere. It's the story of Berkeley Lab really. Working with our colleagues at NERSC, in Biosciences [Area of Berkeley Lab], at UC Berkeley, we're able to iterate on our ideas quickly."

Also key is that the group has built essentially the same tool for materials science, called MatScholar, a project supported by the Toyota Research Institute and Shell. "The main reason this could all be done so fast is this team had three years of experience doing natural language processing for materials science," Ceder said.

They published a study in *Nature* last year in which they showed that an algorithm with no training in materials science could uncover new scientific knowledge. The algorithm scanned the abstracts of 3.3 million published materials science papers and then analyzed relationships between words; it was able to predict discoveries of new thermoelectric materials years in advance and suggest as-yet unknown materials as candidates for thermoelectric materials.

Beyond aiding in the effort to combat COVID-19, the team believes they will also be able to learn a lot about text mining. "This is a test case of whether

an algorithm can be better and faster at information assimilation than just all of us reading a bunch of papers," Ceder said.

Provided by Lawrence Berkeley National Laboratory