PHYS.ORG

Using Big Data to monitor societal events shows promise, but the coding tech needs work

September 30 2016, by Thea Singer

Distinguished Professor David Lazer and his colleagues analyzed global-scale databases of news events and found them wanting. Their recommendations for improvements would enable researchers to build models anticipating everything from the escalation of conflicts to the progression of epidemics. Credit: Matthew Modoono/Northeastern University

In the age of Big Data, automated systems can track societal events on a global scale. These systems code and collect vast stores of real-time "event data"—happenings gleaned from news articles covering everything from political protests to ecological shifts around the world.

In new research published Thursday in the journal *Science*, Northeastern network scientist David Lazer and his colleagues analyzed the effectiveness of four global-scale databases and found they are falling short when tested for reliability and validity.

Misclassification and duplication

The fully automated systems studied were the International Crisis Early Warning System, or ICEWS, maintained by Lockheed Martin, and Global Data on Events Language and Tone, or GDELT, developed and run out of Georgetown University. The others were the hand-coded Gold Standard Report, or GSR, generated by the nonprofit MITRE Corp., and the Social, Political, and Economic Event Database, or SPEED, at the University of Illinois, which uses both human and automated coding.

First the researchers tested the systems' reliability: Did they all detect the same protest events in Latin America? The answer was "not very well." ICEWS and GDELT, they found, rarely reported the same protests, and ICEWS and SPEED agreed on just 10.3 percent of them.

Next they assessed the systems' validity: Did the protest events reported actually occur? Here they found that only 21 percent of GDELT's reported events referred to real protests. ICEWS' track record was better, but the system reported the same event more than once, jacking up the protest count.

The systems were also vulnerable to missing news. "If something doesn't get reported in a newspaper or a similar outlet, it will not appear in any of these databases, no matter how important it really is," says Lazer, Distinguished Professor of Political Science and Computer and Information Sciences who also co-directs the NULab for Texts, Maps, and Networks.

"These global-monitoring systems can be incredibly valuable, transformative even," added Lazer. "Without good data, you can't develop a good understanding of the world. But to gain the insights required to tackle global problems such as national security and climate change, researchers need more reliable event data."

And what about the reported protests that actually weren't protests at all? "Automated systems can misclassify words," says Lazer. For example, the word "protest" in a news article can refer to an actual political demonstration, but it can also refer to, say, a political candidate "protesting" comments from a rival candidate.

"It's so easy for us as humans to read something and know what it means," says Lazer. "That's not so for a set of computational rules."

Analysis begets policy

From community building among scholars and the formation of multidisciplinary groups—which were among the policy recommendations by the researchers—teams within the group could compete against one another to spur innovation.

"Transparency is key," says Lazer. In the best-case scenario, the development methods, the software, and the source materials would be available to everyone involved. "But many of the source materials have copyright protection, and so they can't be shared widely," he says. "So one question is: How do we develop a large publicly shareable corpus?"

Participants should be able to test their varying coding methods on open, representative sets of event data to see how the methods compare, Lazer says. Contests could be used as a catalyst. Finally, the researchers recommend that a consortium should be established to balance the business needs of the news providers with the source needs of the developers and event-data users.

The authors suggest that reliable data-tracking systems can be used to build models that anticipate the escalation of conflicts, forecast the progression of epidemics, or trace the effect of global warming on the ecosystem.

More information: W. Wang et al. Growing pains for global monitoring of societal events, *Science* (2016). DOI: 10.1126/science.aaf6758

Provided by Northeastern University

Citation: Using Big Data to monitor societal events shows promise, but the coding tech needs work (2016, September 30) retrieved 20 April 2024 from https://phys.org/news/2016-09-big-societal-events-coding-tech.html