21. Making Algorithms Work for Reporting

Jonathan Stray

Abstract

Sophisticated data analysis algorithms can greatly benefit investigative reporting, but most of the work is getting and cleaning data.

Keywords: algorithms, machine learning, computational journalism, data journalism, investigative journalism, data cleaning

The dirty secret of computational journalism is that the “algorithmic” part of a story is not the part that takes all of the time and effort.

Don’t misunderstand me: Sophisticated algorithms can be extraordinarily useful in reporting, especially investigative reporting. Machine learning (training computers to find patterns) has been used to find key documents in huge volumes of data. Natural language processing (training computers to understand language) can extract the names of people and companies from documents, giving reporters a shortcut to understanding who’s involved in a story. And journalists have used a variety of statistical analyses to detect wrongdoing or bias.

But actually running an algorithm is the easy part. Getting the data, cleaning it and following up algorithmic leads is the hard part.

To illustrate this, let’s take a success for machine learning in investigative journalism, *The Atlanta Journal-Constitution’s* remarkable story on sex abuse by doctors, “License to Betray” (Teegardin et al., 2016). Reporters analyzed over 100,000 doctor disciplinary records from every US state, and found 2,400 cases where doctors who had sexually abused patients were allowed to continue to practice. Rather than reading every report, they first drastically reduced this pile by applying machine learning to find reports that were likely to concern sexual abuse. They were able to cut down their

pile more than 10 times, to just 6,000 documents, which they then read and reviewed manually.

This could not have been a national story without machine learning, according to reporter Jeff Ernsthausen. “Maybe there’s a chance we would have made it a regional story,” he said later (Diakopoulos, 2019).

This is as good a win for algorithms in journalism as we’ve yet seen, and this technique could be used far more widely. But the machine learning itself is not the hard part. The method that Ernsthausen used, “logistic regression,” is a standard statistical approach to classifying documents based on which words they contain. It can be implemented in scarcely a dozen lines of Python, and there are many good tutorials online.

For most stories, most of the work is in setting things up and then exploiting the results. Data must be scraped, cleaned, formatted, loaded, checked, and corrected—endlessly prepared. And the results of algorithmic analysis are often only leads or hints, which only become a story after large amounts of very manual reporting, often by teams of reporters who need collaboration tools rather than analysis tools. This is the unglamorous part of data work, so we don’t teach it very well or talk about it much. Yet it’s this preparation and follow-up that takes most of the time and effort on a data-driven story.

For “License to Betray,” just getting the data was a huge challenge. There is no national database of doctor disciplinary reports, just a series of state-level databases. Many of these databases do not contain a field indicating why a doctor was disciplined. Where there is a field, it often doesn’t reliably code for sexual abuse. At first, the team tried to get the reports through freedom of information requests. This proved to be prohibitively expensive, with some states asking for thousands of dollars to provide the data. So, the team turned to scraping documents from state medical board websites (Ernsthausen, 2017). These documents had to be OCR’d (turned into text) and loaded into a custom web-based application for collaborative tagging and review.

Then the reporters had to manually tag several hundred documents to produce training data. After machine learning ranked the remaining 100,000, it took several more months to manually read the 6,000 documents that were predicted to be about sex abuse, plus thousands of other documents containing manually picked key words. And then, of course, there was the rest of the reporting, such as the investigation of hundreds of specific cases to flesh out the story. This relied on other sources, such as previous news stories and, of course, personal interviews with the people involved.

The use of an algorithm—machine learning—was a key, critical part of the investigation. But it was only a tiny amount of the time and effort spent. Surveys of data scientists consistently show that most of their work is data “wrangling” and cleaning—often up to 80%—and journalism is no different (Lohr, 2014).

Algorithms are often seen as a sort of magic ingredient. They may seem complex or opaque, yet they are unarguably powerful. This magic is a lot more fun to talk about than the mundane work of preparing data or following up a long list of leads. Technologists like to hype their technology, not the equally essential work that happens around it, and this bias for new and sophisticated tools sometimes carries over into journalism. We should teach and exploit technological advances, certainly, but our primary responsibility is to get journalism done, and that means grappling with the rest of the data pipeline, too.

In general, we underappreciate the tools used for data preparation. OpenRefine is a long-standing hero for all sorts of cleaning tasks. Dedupe.io is machine learning applied to the problem of merging near-duplicate names in a database. Classic text-wrangling methods like regular expressions should be a part of every data journalist’s education. In this vein, my current project, Workbench, is focused on the time-consuming but mostly invisible work of preparing data for reporting—everything that happens before the “algorithm.” It thus aims to make the whole process more collaborative, so reporters can work together on large data projects and learn from each other’s work, including with machines.

Algorithms are important to reporting, but to make them work, we have to talk about all of the other parts of data-driven journalism. We need to enable the whole workflow, not just the especially glamorous, high-tech parts.

Works Cited

- Diakopoulos, N. (2019). *Automating the news: How algorithms are rewriting the media*. Harvard University Press.
- Ernsthausen, J. (2017). Doctors and sex abuse. NICAR 2017, Jacksonville. https://docs.google.com/presentation/d/1keGeDk_wpBPQgUOOhbRarPPFbyCculTOb-GLeAhOMmEM/edit?usp=embed_facebook
- Lohr, S. (2014, August 17). For big-data scientists, “janitor work” is key hurdle to insights. *The New York Times*. <https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html>
- Teegardin, C., Robbins, D., Ernsthausen, J., & Hart, A. (2016, July 5). License to betray. *The Atlanta Journal-Constitution*, Doctors & Sex Abuse. http://doctors.ajc.com/doctors_sex_abuse/?ecmp=doctorssexabuse_microsite_nav

About the Author

Jonathan Stray, a Research Fellow at Partnership on AI, previously taught the dual masters degree in Computer Science and Journalism at Columbia University and built software for investigative journalism.