APPLICATION OF ALGORITHMS IN NEWSROOMS

Javkhlan Bold-Erdene

David Herzog, Project Supervisor

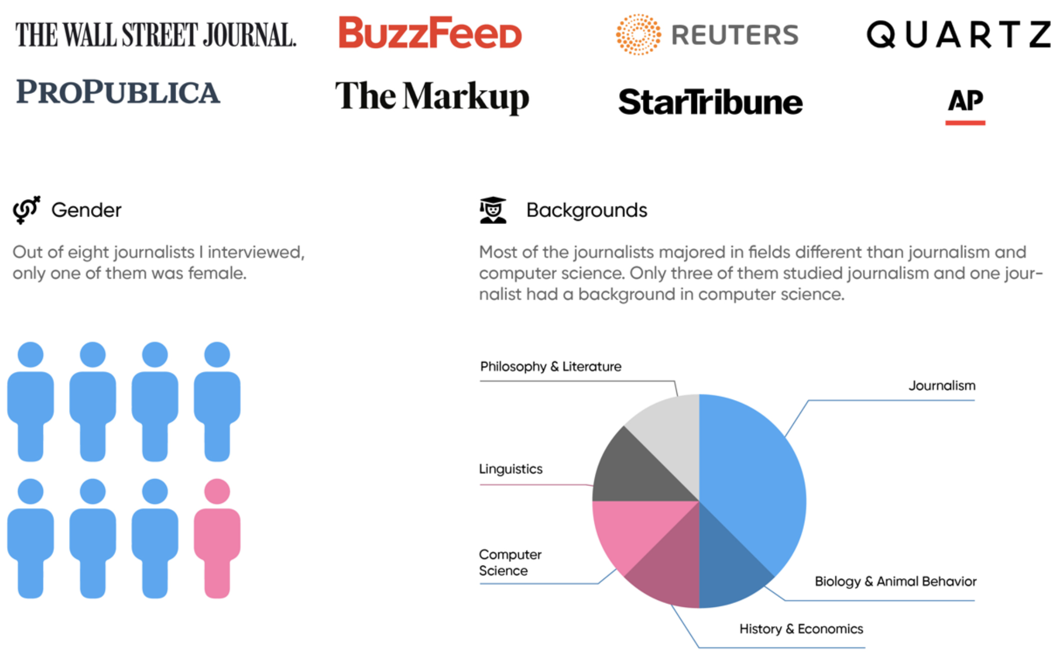
ANALYSIS

As automation, machine learning and artificial intelligence take a foothold in newsrooms, journalists have the opportunity to employ these technologies. This report provides insight and lessons from eight early adopters in medium to large news organizations. It answers the following questions:

What skills and knowledge do journalists need to be able to work with algorithms in newsrooms?

How did early journalistic adopters of AI gain those necessary skills?

Below is an infographic about the interviewees who participated in the study:



Out of eight journalists I interviewed, only one of them was female.

Most of the journalists majored in fields different than journalism and computer science. Only three of them studied journalism and one journalist had a background in computer science.

The necessary skills and knowledge differed, depending on the techniques used. Therefore, it is important to note the differences between those techniques.

Automation, Artificial Intelligence, Machine Learning defined

Automation allows for the collection of and analysis of data and automated journalism refers to the "process of using software or algorithms to automatically generate news stories without human intervention" as described by Andreas Graefe, whose research area focuses on computational journalism.

Artificial intelligence enables journalists analyze data, identify patterns and insights from multiple sources as defined by Francesco Marconi, former R&D Chief at The Wall Street Journal. More broadly, Marvin Minsky defined artificial intelligence as "the science of making machines do things that would require intelligence if done by men."

Machine learning is a subdomain of AI and its two common types are supervised and unsupervised. The technique is mostly used for document classifications and clustering.

Automation in use

Peter Aldhous, Science Reporter at BuzzFeed News, employs automation mostly, rather than something that involves AI or machine learning. He runs scripts in the cloud on a defined timescale that updates various data, maps, and graphics and monitors the data source for updates for their weather tweets.

Troy Thibodeaux, Data Science and News Applications Editor at Associated Press also started using automation for earnings reports by working with Automated Insights around 2012.

Maddy Verner, Investigative Data Journalist at The Markup, said that automation is about letting computers do work, collecting data from the internet using things like scrapers. Verner is the only female journalist I interviewed and the only journalist who had a background in computer science.

Artificial intelligence, Machine Learning in use

Supervised machine learning, a technique used to classify documents, is the most common technique used by these early adopters.

Aldhous used supervised machine learning for his project about identifying covert spy planes, "Hidden Spy Planes," which won 2018 Data Journalism Award. He had to train a dataset to get an estimate of whether something is a spy plane or not.

Aldhous is wary of using AI techniques in journalism. "The trouble is, with those approaches, in my view, is that you're not able to very well diagnose what's going on under the hood. You don't know what the AI is doing in that case, in the same way that I had a very clear idea by using the very simple and well used, well-known algorithm exactly what it was doing," he said. He prefers to use the term machine learning than AI due to transparency issues.

Aldhous is not the only one who prefers the term machine learning than AI. Thibodeaux, who started his journey from automating stories from earnings reports said the terms machine learning or algorithm make AI "less science fiction."

Although the needs for these journalists to use machine learning were varied, most of them started using the technique for working with huge documents and data.

When it comes to time journalists first started using AI and automation, it goes back to around 2006 at the earliest as Chase Davis, who at that time was working on a project with the National Center for Supercomputing Applications at the University of Illinois, where he used some of the techniques for analyzing documents. He is now Senior Digital Editor at Star Tribune.

Jeff Ernsthausen, Data Reporter for ProPublica, started with merging lists and got interested in machine learning applications in journalism from there. He wrote a program for logistic regression for the Doctors & Sex Abuse, investigative project by The Atlanta Journal-Constitution that discovered doctors were allowed continue practicing after found sexually violated patients. The first of the series ran on The Atlanta Journal-Constitution website in 2016.

Ernsthausen explains his approach to the project as “It's always a technique for helping me assemble a data set, possibly for helping me understand what's important inside the data set, but it's almost always going to be just part of the workflow.”
Jeremy Merrill, machine learning journalist at Quartz, was on a team with The International Consortium of Investigative Journalists that worked on Luanda Leaks project that revealed corruption of Angola’s former president’s daughter. He used machine learning to do a semantic search on leaked documents about the corruption and said that he faced a technical challenge to encode all of the documents because tools that were available were not fully mature. The stories ran on Quartz in January 2020.

Demystifying the technical skills

Most of these journalists started learning themselves the necessary skills to understand AI and automation. Among the learning methods, having projects or problems to solve at hand, learning from books, and talking to experts played the most significant roles in skills development.

Davis self-taught himself how to analyze documents using machine learning. He started by looking at books in the university library and finding codes online to learn from. Then he started to meet more people who were working with machine learning and asked questions. The best approach to him was working on big projects that needed those skills.

For a project that's about predicting whether bills would pass at the Georgia legislature, Ernsthausen picked a book about R and logistic regression and spoke with the author of the book while working on the project. He said, "Learning how to do something is always easier when I have a thing I'm trying to do."

Particularly for automation, Padraic Cassidy, Editor for News Technology at Reuters, suggests starting with some projects and start tackling them. "And then as your ambition grows to cover what you want, you realize you're going to need something bigger and bigger to handle it," he said about how people start thinking about what can be automated, which lead them to look for tools and experts who can help with their work.

Not everyone had to learn new skills immediately when they shifted some of their operations into automation. Thibodeaux 's newsroom hired an automation editor after they used Automated Insight for two years and started looking broadly at opportunities across the newsroom for automation. The automation editor brought knowledge in programming languages such as Python and JavaScript and they started building up their own system.

In fact, not all news organization can afford such skills, especially small and medium newsrooms.

Francesco Marconi, the former R&D Chief at The Wall Street Journal, suggests small and medium news organizations partnering with startups like Newlab, a news AI venture based in New York developing tools for the media industry with a goal of "democratizing its use across the industry", which will not require a big financial commitment.

Besides technical skills, core competencies expected from journalists are being able to get a dataset, quickly assess problems in the dataset and figuring out what kind of statements can be made in a story using that dataset. And such basic skills will make it easier for somebody to learn more advanced techniques like machine learning.

One journalist emphasized the importance of an adaptability and flexibility with regard to anything to do with technology as it's more about ability to learn and self-teach new things because things will change.

Peter, who has an educational background in biology and animal behavior said, "If you think that everything is going to have to be formally taught, that's going to be quite limiting."

Cassidy added on that by saying that he wouldn't recommend tying it down to a specific language because they change so fast and frequently. "It's really just knowing what the concept of machine learning is and what it does and what it can do," he said. "I know a little bit of Python, but not I couldn't pass a test with any serious competency, but I know what it's doing, and I know how to formulate queries and to get things out of data. So, I can show our development team the kinds of things we're looking for."

Most of the early adopters agreed that not all journalists have to learn such skills or are interested in AI or automation and the needs to learn those skills depend on the newsroom and the role of the journalist. Important questions to define those needs include what opportunities would be there and whether there is an environment in the newsroom to discuss such techniques.

Large newsrooms such as AP has a dedicated data team with people from background in machine learning as they either did that in school or through their own work.

Thibodeaux said, “And so for us really is a question that how do we keep them current? How do we make sure that they're aware of new developments, or seeing other applications of this technology? And a lot of that is to create the conditions for constant learning on your team, and to challenge them.”

Thibodeaux’s team set a learning goal every year and every quarter and set questions such as whether they are fully aware of the latest cloud-based resources for machine learning in the latest machine learning platforms to build up expectations for the year. “Because your skill sets could improve, and our team is going to benefit because we're going to be up to date and making use of the best available technology. And AP will benefit because we're going to get the stories faster that we might not have ever gotten to before,” he said.

About commonly used skills these days in terms of AI and automation, although these journalists name programming languages and mathematics, that’s not all it takes. For example, for Ernsthausen, one of the most important things is knowing how wrong things can go so that it gives the journalist a sense of developing a good sense of caution for using them.

He explains that rather than understanding exactly the math or the way individual statistics are calculated, understanding the general intuition of how things work is important. Besides that, being able to clean documents, PDFs and make them usable and being able to search through things are useful these days.

Verner also emphasized the importance of being able to look at data and see if there is anything “weird.” For instance, being able to see how a curve looks different than the usual normal distribution curve.

Davis said, “If you don't understand exactly what that algorithm is doing, and why it's making the decisions that it is and what trade-offs you're making, you could mischaracterize the work that algorithm has done very easily.”

Whether it’s necessary to train reporters in those skills, Ernsthausen said that most of the time, problems can be solved with techniques that are not artificial intelligence. For example, after doing his project about doctors and sex abuse, Ernsthausen would get asked by people if he could do the same thing for millions of documents related to a scandal going on in City Hall. He said that it was about finding all documents related to one person, which wouldn’t need statistical models but could be done by a search feature. “I think sometimes people get caught up in AI and AI is so important. But most of the challenges we face are still ones that can just be done with sort of simple rules-based programming kind of approaches,” he said.

In addition to the skills and knowledge these journalists have today, they said they are curious about more tools for analyzing huge piles of documents besides learning other languages such as JavaScript and R or generally more about machine learning. Such skills would create opportunities for them like creating databases that never existed before as Ernsthausen explained as “I think just for practitioners of journalism, especially data journalism, that's a real sweet spot for finding and new stories and bringing new light to topics because it's difficult to do without some data journalism skills. And chances are that if the data doesn't exist yet, then no one studied it, and so there's plenty of room to bring new insights to that area."

Marconi said that implementation of automation, analytics and AI processes in newsrooms require significant human labor. He said that as AI enters the newsroom, the tasks of creating and managing these tools will also change the makeup of the newsroom skillset.

"In the future, we will see more newsrooms asking for writers that understand how to work with AI, editors that understand how to oversee smart tools, programmers that can design journalistic computer programs, and designers who can evaluate the user experience of reading AI-generated content," Marconi said. He mentioned automation editor, computational journalist, newsroom tool manager, AI ethics editor as possible new roles in newsrooms.

Early challenges

When journalists first started using AI or automation, their most common challenges were around knowledge about mathematics, statistics, programming language, and lack of resources to learn.

"Not being computer scientists and mathematician was the hardest thing," said Davis.

Verner explained her challenge when first started as "The real problem is not really using the technology for the data we have, but it's actually finding and working with the data that we need to work with them in the program." She worked on a project about Allstate's secret auto insurance algorithm that was squeezing money out of big spenders in Maryland in 2019.

For automation, technical challenges seem to have been much less than using machine learning and those challenges were more related to understanding processes in newsrooms.

Peter said he didn't have many challenges when he first started as his main role is switching scripts on a cloud server and a cron runs the scripts every certain minute and process the data, which requires less to no human interaction.

Another journalist who works specifically with automation, Thibodeaux, said that technical skills were not necessary whereas developing process and building up templates was more of a challenge. He said, "The idea is based on the inputs and some logic, you can construct a fairly reasonable story from it. And so, developing that process and understanding how we can get the insights of our reporters into that automated content was a real challenge."

Most of the journalists agreed on the lack of time as a barrier for journalists to learn new skills.

Merrill, who has a background in linguistics said, "That just takes a long time to learn these skills and experiment with them and be able to use them confidently."

Resources for developing the skills

Compared to when these journalists first started getting their hands around such technologies, many more resources have become available today that support self-learning approach.

Among them, conferences such as NICAR, Strata Data, Computation + Journalism and online resources including Coursera, Codeacademy, fast.ai, and MOOC classes included. Online books such as Text Analysis with R were also helpful. Additionally, contacting the book author or experts is one of the useful ways to learn.

Besides conferences and online resources, self-devotion, a hands-on project to work on seem to have helped these journalists to keep up with the advancement of technology.

Ernsthausen said, “I think it's helpful to have something in front of you that you're really passionate about that helps you give you a reason to really, you know, really sink into something.”

Cloud service platforms for machine learning also have been useful resources for news organizations.

For example, AP focuses on such services available from Amazon Web Services and Google. It helps them to speed a lot faster than they could because when they first started doing things, they had to code everything from scratch and there was a steep learning curve just to get started.

“But now with this, or the Google platforms, or the Amazon platforms and some of the Microsoft platforms, when we looked at all of them, and we're experimenting with them, it gets such a head start, you know, and then you're able to do some things that were very difficult. You can do much, much faster now,” said Thibodeaux.

Expectations from early career journalists or new graduates

When asked about skills expectations from newsrooms today, most of these journalists agreed that detailed technical skills are not priority, but a basic understanding of what AI is, and its capabilities is important because there will be always something where specialists are going to be required.

Davis explained it as “It's almost like the internet. The internet is such an important part of what we do, that everyone, who considers themselves kind of curious and educated member of society should have a basic understanding of how what it is and how it works. It doesn't mean they have to learn how to write HTML and CSS and do all this code, but they should just generally know what the internet is and how it works. And I think the same is true with a base understanding of artificial intelligence at this point.”

Similarly, Merrill said, “Just not everyone needs to be a photographer and not everyone needs to be a really incredibly literary writer. These are different strengths and different people in the newsroom will have different strengths. And that's okay.”

For AP, they’ve come to the conclusion that they are not going to train every journalist at AP to be a data journalist or data scientist. “But being able to be comfortable with numbers, being able to think about stories quantitatively and ask what kind of data centric approach might provide more better sources and more insight into the story,” said Troy. He said the core skill would be around having the capacity to speak the comfort level with working with numbers and ability to go out and get the data and begin looking at it. For anything beyond that, AP has a small team to help with coding and advanced statistical techniques.

Conclusion

Automation, as a practice, requires more of understanding of process in newsrooms overall. In contrast, machine learning can be applied in different ways depending on the nature of the reporting projects that journalists work on.

Here are three main takeaways from the early adopters:

First, not every journalist needs to learn skills related to AI and automation, but having a basic concept about what they are, and their capabilities would be benefiting. In fact, it's more about data work such as cleaning that is used more often than actual AI or automation;

Second, it has become a lot easier to learn automation, AI or machine learning-related skills as most of my interviewees self-learned such skills using the abundance of resources that are available today. Moreover, it's more about learning such skills at conceptual levels without having to learn highly technical mathematics and statistical skills behind them as there is no need to be proficient in all different techniques used. It's also easier to learn from best practices as these journalists share their works and explain how they did their projects on their GitHub accounts, which makes more resources accessible for anyone interested in applying the same techniques.

Third, knowledge about common failure cases or being able to see where things can go wrong with automation and AI is as equally important as technical and professional skills. In other words, gut check pays off.