# AUTOMATED TECHNIQUES IN SMALL NEWSROOMS

Yaryna Serkez

Mike McKean, Project Supervisor

## ANALYSIS

Throughout the history of media, abrupt changes – from radio's "hypodermic needle" effect in the early 1930s to today's social media revolution – have had one thing in common – new technologies. Indeed, the development of new information technologies goes side by side with their experimental implementation in the journalism industry. No matter if we are talking about innovations in data analysis, communication technologies, or even something as distant from shoe-leather journalism as satellite imagery, a broad scope of new information-related techniques eventually ends up being integrated into the journalist's toolkit, to some extent. With the emerging innovations in the field of Artificial Intelligence (AI) and new approaches in data analysis and collection, the implementation of these techniques looks extremely promising. As highlighted in this year's AP special report on augmented journalism "… it promises to reap many big rewards for journalism in the years to come. Greater speed, accuracy, scale and diversity of coverage are just some of the results media organizations are already seeing" (Marconi & Siegman, 2017). The leading role of algorithms in reshaping newsroom routines and work strategies was also recognized in the recent Tow Center report on automated journalism. As noted by the author, they "can create content on a large scale, personalizing it to the needs of an individual reader, quicker, cheaper, and potentially with fewer errors than any human journalist" (Graefe, 2016).

One of the most fascinating aspects of this phenomenon is how wide the spectrum of potential use is for these technologies. Being one of the first pioneers of automation in the early 1980s, Zuboff shaped the main directions of algorithmic applications for the decades to come: "… information technology is characterized by a fundamental duality that has not been fully appreciated. First, the technology can be applied to automate operations… Secondly, technology can be used to create information." (Zuboff, 1985, p.371). Dexterous programs designed to recognize emotions on video or wildfires from satellite imagery, smart bots to post the most notorious results of state procurements, sophisticated scripts to extract topics and bullet points from public speeches or complicated algorithms to spot propaganda and fake news; all these possibilities are on the cutting edge of what can be done by newsrooms once they master how to use automated techniques to their advantage. Moreover, automation of the most routine and repetitive work also means more time for in-depth analysis and investigations. In the era of post-truth and alternative facts, the resources of regular journalists are particularly scarce. Sound, reliable reporting demands exhaustive preparation and fact checking, — intrinsically cognitive tasks, which are still better performed by humans than machines. With this in mind, it's fair to say that automated techniques are a key not only for work optimization and higher productivity, but also for better and more accurate reporting. As Lynn and Hermida summarized in their study on computational journalism, "…journalists framed the robo-posts as a tool to facilitate systematic coverage and freed journalists to add more depth, context, and the human touch, as well as possibly decrease costs" (Lynn & Hermida, 2014, p. 392).

Another particularly promising aspect of the AI technologies and other automated techniques is their accessibility. Even though they demand highly qualified specialists and extensive training, their successful usage doesn't depend on the newsroom's size or budget. The availability of open source solutions and algorithms makes it possible to deploy such techniques even in small newsrooms with just a few skillful data journalists or programmers on board. These factors create a broad field of potential implementations in media of any scale, starting from the giant media corporations and finishing with small, niche teams of reporters. Empowering journalists with better and faster data gathering, analysis or even content generation, automation technologies and artificial intelligence provide a unique opportunity for more in-depth reporting, more precise investigations and much faster media coverage.

This paper will investigate how small newsrooms can take advantage of automated technologies to help generate content and conduct more thorough journalistic investigations and increase the overall productivity of the newsroom. To encapsulate the full potential of the emerging automated techniques, I took a rather broad approach by framing them as technologies "reducing human effort and squeezing time out of the many chores journalists must undertake to get the story and get the news out to the public" (Marconi & Siegman, 2017). For the purpose of this paper I conducted a case study scrutinizing projects and routines of the small, data-oriented Ukrainian newsroom *Texty.org.ua*. My approach includes structured interviews with *Texty's* leading reporters and programmers, analysis of their latest projects, along with my own observations acquired while being engaged with some of these projects myself.

*Texty.org.ua*, more commonly known as *Texty*, is an award winning Kyiv-based team of data journalists. They specialize in big data analysis and visualizations. In 2012, *Texty* won a silver prize at the Data Journalism Awards for their online database of state procurements. In 2016, their project about declarations of Ukrainian state officials was shortlisted for the 2016 Data Journalism Awards as a best data app of the year (small newsroom category). *Texty* enjoys tackling ambitious projects and right now is the leading data-oriented media outlet in Ukraine. Nevertheless, they are a tiny team with only seven people being employed on the newsroom side. However, it doesn't stop them from creating large-scale data projects. Their secret is simple: they are automating the lion's share of tasks, using open source algorithms and creating their own for the purposes of each particular project. Tedious data collection or cleaning, which can hardly be done by a team of ten people, is easily achieved with just a few scripts. Geocoding thousands of addressees, which would previously take a few weeks, is done in a couple of minutes with an in-house tool. Detection of the propagandistic trolls on Facebook, which would take months of qualitative human-made research, is now done overnight with a sophisticated algorithm.

To get more insightful information about *Texty's* experience in automating its daily tasks and using algorithms to conduct advanced data journalism investigations, I've interviewed three main developers and analysts from their team: Anatolii Bondarenko, the head of the data team, Vlad Gerasymenko, leading web developer and Andrii Gazin, leading specialist in data analysis. Gazin recently left the team but was highly involved in *Texty's* investigations and is still engaged with a few projects as a consultant. The wide range of questions I've asked can be easily divided into two distinctive categories. The

first category is connected to automation techniques they have used or learned recently and the results of the implementation of these technologies. The other group of questions focuses on currently missing tools and opportunities, unsuccessful or unfinished projects, and potential tasks that can be automated. Beyond interviews, I also analyzed the tools used in *Texty's* recent projects. More precisely, I scrutinized 18 data projects published since January 1, 2016. I paid special attention to the particular technologies that helped to create these journalistic pieces and empowered a small team of reporters for such massive investigations. Last but not least, I also incorporate my own observations and experiences, since I have been engaged with some of these projects as a consultant or part-time designer and developer in the past. This mixed case study approach gives me the ability to examine algorithm usage in full depth and get a comprehensive picture of automated technology implementation in a small newsroom.

## Existing technologies and current practices

*Texty's* reporters agreed that automation techniques, to some extent or another, are being used in almost all their large-scale investigative projects. "Reproducibility is essential for us. Data may change, variables may change, but we should be able to replicate the same analysis to expand results or double check them and automation is a key to that," Gazin says. In fact, 14 out 18 recent projects by *Texty* used algorithms, either for data collection, data mining, modeling, or prediction.

As Bondarenko noticed, sometimes the application of such technologies is not only an attempt for faster reporting, but also a core part of the investigation. For instance, *Texty's* declaration project about the discrepancies between declared assets and official income of Ukrainian politicians would not be possible if Bondarenko would not have

created the model to determine car prices. The initial dataset included the declared salaries, cars, real estate, and money on bank accounts that belonged to officials and their family members. Even though this seems like a pretty rich dataset, in fact, the real value of the possessions of the government officials was missing. For example, how expensive is the MP's BMW from 2009 or his wife's Lexus from 2015, and how do those assets correspond to his $500 per month income and the housewife status of his spouse? To get this data and answer these questions he created a model trained on huge datasets downloaded from numerous classified advertisement websites. Separate models were created for each car model and brand to determine its current price based on such factors as year of issue, manufacture country, brand, liter capacity and mileage. As Bondarenko noted, after checking for multiple variables he surprisingly found out that the most influential variable was car age, which explained up to 70% of the variance. As a result, Bondarenko's model helped to estimate prices of more than thousands of cars declared by Ukrainian politicians and state officers. The final app based on the results of this modeling helped to compare car prices with the official salaries of government officials, and helped to assess the potential scale of corruption or at least present some vivid evidence for further investigations.

The same was true for a deforestation project, which I created for *Texty* in June 2016. The problem of illegal cuts in the Carpathians was becoming more and more apparent. Photos of treeless hills and mountains overwhelmed social media newsfeeds in Ukraine. These events raised the question of how to prove this deforestation was actually occurring. Even the biggest media organizations can hardly afford a huge expedition to check the whole mountain range acre by acre. So we decided to use satellite imagery and

train a supervised learning model, which would determine whether a particular area is bare soil or wood. Later I compared land cover classification results for different years and determined the scale of fresh cuts. This project was a classic example of how algorithms can generate data, which later results in further investigation and an independent journalistic piece.

Another great example of algorithm usage in Ukrainian media was *Texty's* Trolls project. The initial idea was to detect propagandistic trolls, who were spreading calls for riots and protests across the Ukrainian segment of Facebook. The trolls in question may have a lot in common, but it would normally be difficult to track and identify them. To accomplish such an ambitious goal, *Texty* developed a set of criteria, which then served as a basis for a troll detection algorithm. Their method started with manually creating the ultimate list of Facebook groups highly saturated with bots and propagandistic content, and which were connected to popular DNR separatist Stepan Mazura. Then, *Texty* scraped members of these groups and networks of their friends and so on. If a particular user from this giant network would meet the trolls' previously determined criteria, he or she will eventually end up in the so-called Trolls Network. As a result, journalists discovered almost 2,000 Facebook users involved in the propaganda machine. Approximately 80% of them were bots, the rest were real people who moderated the bot-saturated Facebook groups or just had a lot of trolls in their friend lists and actively reposted their content. *Texty* also created a separate network of troll-based Facebook groups' moderators and managed to trace which media resources these groups shared the most.

Nevertheless, in the majority of the cases automation was still connected with work optimization and faster data collection. “The idea of our investigation about reporters being paid to promote politicians by writing about them, originally came from a reporter who reached out to us because he didn’t have time to collect all of the headlines by hand. After hearing this, I wrote a script capable of collecting news headlines automatically, and at a later stage of the project I developed several scripts capable of tracking mentions of different politicians across different media websites,” Gazin says. He also noted that when a particular problem is seen as repeating throughout many projects, *Texty* prefers to develop in-house solutions to obtain the best and most correct results as fast as possible. This was the case with geocoding Ukrainian addresses, which are usually in Cyrillic and are poorly recognized by the majority of the available open source solutions. Since this was a common problem for several projects, *Texty* created their own tool based on the Yandex API, that can geocode up 25,000 addresses per day with more precision than similar Google Maps or OpenStreetMap based equivalents.

Data cleaning is another major application for automated techniques. Often, available data is inconsistent or flooded with irrelevant pieces. As Gerasymenko told me, his project on state procurement would never have happened if not for the R scripts he wrote to clean records. The original dataset predominantly consisted of Office Word files, which, in turn, had tables with data inside. Manually extracting data from hundreds of such files is inconceivable, so he used the available automated techniques to extract tables from Word files and parse them into data frames suitable for further analysis.

## Experimental projects and room for growth

But what about more advanced machine learning and artificial intelligence? During the interviews I've learned that, in fact, many such attempts are still being developed or have been suspended because of weak performance or unsatisfactory results.

For instance, one such endeavor was an attempt to process and cluster Twitter anti-Ukrainian bots' posts, detect their main topics, and create a separate Twitter bot that will post messages summarizing themes or the news that these propagandistic bots are pushing on a particular day. Even though the topic modeling part of this project was rather successful, the results were still very messy. As Gazin noted "These bots are trying to pretend they are not bots, so they are diluting the flow of striking propaganda with some random irrelevant messages. The latter usually make no sense, which made our results rather blurry". Notwithstanding this previous failure, *Texty* is still developing this project. The topic of Russian propaganda is extremely important for Ukrainian readers and Bondarenko believes this deserves a second try. That's why right now they have a separate team exclusively focused on this issue.

There was also an idea to build a Twitter bot that would automatically tweet the most notorious declaration results, but it was suspended because of a declaration API malfunction. Finally, the third content generating project was aimed to create a Twitter bot that would post alerts about press conferences in which paid sociologists spread false surveys. The list of these pseudo-sociologists resulted from *Texty's* previous project when they analyzed the large amount of suspicious agencies that would publish survey results before elections, and would then vanish the next day after these elections are over. Eventually, this project was postponed, but might be revisited in the future.

However, it was not always poor data that made a particular project hard to accomplish. Sometimes it was just an absence of particular technologies. Bondarenko confessed that over the last few years he often noticed that he was repeating the same work in different projects, and every time it was similarly tedious and exhausting. This repetition was record linkage, a particular problem in text analysis when you have multiple names of the same object but they are written slightly differently each time. Bondarenko had to tackle this problem for his project on school education in Ukraine. The existing database included thousands of state schools across Ukraine and standardized test results of its students for many years. But each year the name of many of these schools was recorded differently. It could get as diverse as “School #67”, “General School #67”, “Stus General School”, “Stus School #67” etc. For this particular project he solved this problem, but usually results are far from being satisfactory. All my interviewees complained that the whole cluster of the natural language processing problems is poorly developed for Slavic languages, including Ukrainian. Techniques of text and language processing are still at the early stages of their development for these languages. There is no comprehensive method for tagging parts of the Ukrainian language, neither are there reliable algorithms for detecting main themes. “But we do recognize the power and opportunity of implementing AI and machine learning. Some of those projects are still experimental for us, but I hope in the near future they will yield the first successful results,” Bondarenko concluded.

*Texty’s* experience signifies how substantially the newsroom’s workflow can change once they master the use of algorithms and automated technologies. Moreover, implementation of such techniques is directly connected to the scope of topics journalists

can cover and depth of analysis they can conduct. Not only can automated techniques considerably boost the productivity, they can also generate data and content, which then will serve as a basis for future journalistic investigations. However, these technologies also have their limitations and drawbacks. As my colleagues noticed, language-processing algorithms are still pretty weak, especially for Slavic languages. On one hand these week algorithms deprive journalists and analysts from a wide range of topics they can investigate, but on the other hand they create unique opportunities for further growth and innovations.

## Opportunities for Future Research

New automation and machine learning technologies are gradually becoming a part of the newsroom arsenal. They provide a unique possibility for deeper data analysis, better productivity, and more accurate reporting. As Larry Fenn accurately summarized it, “It’s important to bring science into newsrooms because the standards of good science — transparency and reproducibility — fit right at home in journalism” (Marconi & Siegman, 2017). More importantly, these techniques can be used for a wide variety of tasks on different stages of news production in newsrooms of almost any scale. Their implementation in small newsrooms is particularly interesting, since it can naturally solve the eternal problem of small teams and limited resources working on big projects.

Further research on the results of automation practices will pave the way for better understanding of how journalists can use these technologies to their advantage, and which tools and algorithms are particularly promising and should receive more attention from media managers, developers, scientists and the broader startup community.

## References


*The Future of Augmented Journalism: A Guide for Newsrooms in Age of Smart Machines,* AP, 2017, https://insights.ap.org/industry-trends/report-how-artificial-intelligence-will-impact-journalism Accessed 5 April, 2017.

*Guide to Automated Journalism*, Tow Center For Digital Journalism, 2016, http://towcenter.org/research/guide-to-automated-journalism/. Accessed 1 November 2016.

Lynn, Mary. and Alfred Hermida. From Mr. and Mrs. Outlier to Central Tendencies. *Digital Journalism*. 3(3), 2014, pp.381–397.

Zuboff, Shoshana. *In the Age of the Smart Machine: The Future of Work and Power*. New York: Basic Books, 1988.