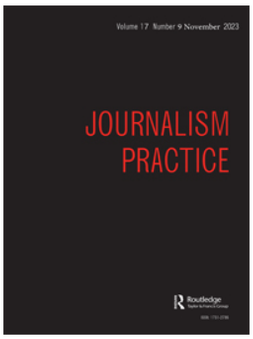

Routledge
Taylor & Francis Group

Journal of Journalism Practice

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rjop20

How (not to) Run an AI Project in Investigative Journalism

M. Fridman, R. Krøvel & F. Palumbo

To cite this article: M. Fridman, R. Krøvel & F. Palumbo (04 Sep 2023): How (not to) Run an AI Project in Investigative Journalism, Journalism Practice, DOI: 10.1080/17512786.2023.2253797

To link to this article: https://doi.org/10.1080/17512786.2023.2253797

Published online: 04 Sep 2023.

JOURNALISM PRACTICE
https://doi.org/10.1080/17512786.2023.2253797

How (not to) Run an AI Project in Investigative Journalism

M. Fridman [a], R. Krøvel [b] and F. Palumbo [a,b]

[a]Artificial Intelligence Lab (AI Lab), Institutt for informasjonsteknologi, Oslo Metropolitan University, Oslo, Norway; [b]Fakultet for samfunnsvitenskap, Institutt for journalistikk og mediefag, Oslo Metropolitan University, Oslo, Norway

ABSTRACT

Data journalists are increasingly reliant on automation and artificial intelligence (AI) to process and analyse massive datasets. AI can contribute to journalism by creating visualizations, verifying accuracy of information, analysing historical data, monitoring social media, finding patterns and outliers, generating text and much more. However, the integration of AI into the newsroom comes with its own challenges. In this article, we take a practice-based approach to develop a deeper understanding of how to overcome such challenges. Our teams of data scientists, AI experts and journalists took on four projects incorporating data science and machine learning into investigative journalism. From those experiences, we found that access to data at scale, data quality and reworking the concept of "newsworthy" as a machine learning question were the most significant obstacles to deploying AI in the newsroom. We recommend closer collaborations between team members of different disciplines to create a truly trans-disciplinary approach, as well as some practical considerations for choosing projects to facilitate successful AI-assisted investigations.

ARTICLE HISTORY
Received 21 March 2023
Accepted 27 August 2023

KEYWORDS
Data journalism; investigative journalism; machine learning; data science; artificial intelligence; trans-disciplinary journalism

Introduction

Artificial Intelligence (AI) is a broad field of computer science that aims to develop intelligent machines capable of performing tasks that typically require human intelligence. Machine learning is a subset of AI that develops algorithms and statistical models enabling computers to learn and make predictions or decisions without being explicitly programmed by humans. AI also includes other approaches such as natural language processing, computer vision, expert systems, and robotics, which collectively enable AI systems to understand, reason, learn, and interact with humans and their environments.

Artificial intelligence (AI) is being rapidly adopted by news media around the world, to the point that both the public and the journalists themselves start to wonder whether "robots will replace journalists" (Miroshnichenko 2018). However, while the adoption of AI in journalism is accelerating, experiential knowledge about AI applications in

CONTACT R. Krøvel royk@oslomet.no

journalism is lagging. Research on AI in journalism has been mostly qualitative and focused on a few topics such as data journalism, robotic writing, and news review (Parratt-Fernández, Mayoral-Sánchez, and Mera-Fernández 2021).

With the advent of digital technology and the explosion of data, it is becoming increasingly important to use AI to support investigative journalism, as traditional methods are no longer practical. The development of new, often open-source, tools makes solutions easier and faster to implement and requires fewer specialized resources. In this context, computers can play a central role in automating repetitive and computationally intensive processes, enabling journalists to extract information that would otherwise be inaccessible (Beckett 2019). To visualize the growth in automated data handling it´s enough to consider that the "Pandora Papers", released by ICIJ and composed of 2.94TB of documents, is 1700 times as large as the "2010 Wikileaks", which was 1.7GB (Infographic 2021; Pandora Papers 2021).

Upon examining the current research landscape, it is evident that two main research strands are prominently featured in the field: reports by journalists and/or developers describing a specific use case, and research articles based on interviews, surveys, and literature reviews (Ausserhofer et al. 2017; de-Lima-Santos and Ceron 2022; Stray 2019a). However, the complex nature of investigative journalism requires alternative and experimental research methodologies that enable researchers to understand and analyse often novel investigative methodologies. One of the biggest challenges for multidisciplinary teams working on AI in investigative journalism is to be able to communicate and collaborate effectively (Santos and de-Lima-Santos 2022).

This article draws on the insights and experiences gained from participation in four interdisciplinary teams in which data scientists and journalism researchers collaborated with investigative journalists on various projects. Our goal is to analyse how interdisciplinary teams can overcome the problems and limitations identified by researchers such as Stray (Stray 2019a) and lead to a better inclusion of AI methods in investigative journalism. By understanding the problems that need to be addressed, this paper aims to develop better methods for incorporating AI into investigative journalism.

Data Journalism and AI in Journalism

Data journalism is a field that incorporates data analysis, visualization, and database use with traditional journalism practices to uncover and tell stories. Data journalism has made substantial contributions to society, offering information and insights that have led to increased transparency, accountability, informed decision-making, and public understanding (Bounegru and Grey 2021). Data visualization is a critical aspect, as it helps journalists present complex information in a simple format (Rodríguez, Nunes, and Devezas 2015). Using data and interactive visualization, data journalism has revealed injustices, held powerful individuals accountable, and improved the functioning of society and the lives of citizens (Ausserhofer et al. 2017; Borges-Rey 2016; Bounegru and Grey 2021; Santos and de-Lima-Santos 2022; Young, Hermida, and Fulda 2018).

There has been a push to include more data science and AI in data journalism. Data journalism and AI journalism are two different approaches to journalism that can overlap. Data journalism involves the use of statistical methods, data visualization tools, and other techniques to identify patterns and trends in large datasets, and to present

this information in a way that is accessible to the general public. AI in journalism additionally involves using AI technology to analyze data and discover, develop and publish new stories. Whereas the most used software by data journalists are Microsoft Excel and Google Sheets (The State of Data Journalism 2023), data scientists regularly use programming to wrangle and scrape data, improve data cleaning, and create custom analysis and dynamic interfaces to promote data interaction. They can also use machine learning and recent AI advances to automate repetitive processes, extract relationships that are hard to see otherwise and detect newsworthy anomalies (Amazon Mining Watch n.d.). Data journalists are now more reliant on automation and AI to process and analyse massive datasets, and even generate new stories (Challenges and Opportunities – Survey – State of Data Journalism 2022, n.d.). Florian Stalph identifies four main categories of data journalism: explanatory data journalism, investigative data journalism, interactive data journalism, and advocacy data journalism (Stalph 2018). A similar approach is followed by Konstantin Nicholas Dörr mapping the field of AI journalism into automated news writing, data-driven journalism, personalized news recommendation, and algorithmic curation (Dörr 2016). In addition, AI can contribute to journalism by creating visualizations, verifying the accuracy of statements, analysing historical data, and monitoring social media (AI and the Future of Journalism n.d.; Hacks/Hackers LDN 2019; Miroshnichenko 2018; Weber 2021). However, it is important to view AI as an aid to journalists, not a replacement. The best outcomes are achieved by combining AI with human expertise and ensuring unbiased and diverse data is used for training.

Data collection is a critical aspect of artificial intelligence (AI) because AI algorithms rely on large and diverse datasets to make accurate predictions and decisions. However, data collection for AI can be problematic due to several reasons. First, it can introduce bias and discrimination if the collected data reflects societal biases or discriminates against certain groups, leading to unfair outcomes (Buolamwini and Gebru 2018; Zhao et al. 2018). Second, limited or incomplete data can result in models lacking accuracy and reliability (Jain et al. 2020). Moreover, privacy concerns arise when personal data is collected, stored, and potentially used without consent, posing risks to privacy and data security (Crawford and Schultz 2014). Ethical considerations are also important, especially when sensitive information is involved, requiring transparency and adherence to ethical guidelines (O'Neil 2016). Data quality and pre-processing challenges, including errors and inconsistencies, can undermine AI model effectiveness. Data imbalance, where certain groups are underrepresented, can lead to biased results. Additionally, data ownership and access rights can create barriers to research and fair competition.

The research into data journalism is an ever-evolving field, but a few trends are emerging. Several studies highlight the importance of multidisciplinary teams (Borges-Rey 2016; de-Lima-Santos & Salaverría 2021; Santos and de-Lima-Santos 2022). Collaborations between journalists, data scientists, and developers to create tailored analyses and data-wrangling solutions are key to an investigations' success. Teams that instead have to rely on readily available free online tools struggle with the lack of customization (Young, Hermida, and Fulda 2018).

Despite the potential benefits, the adoption of AI in the field of investigative journalism has not been widespread. The cost of implementing new technologies, like AI, in various industries may be a factor. However, there is potential for AI to reduce costs as well, in the

scope of automation and lead generation. In addition, given its potential to recognize patterns and find stories that would otherwise stay buried, the cost-benefit analysis cannot be calculated purely in financial terms. This value to society allows such projects to turn to public funds, philanthropy, and crowdfunding.

Apart from the economical hurdles to incorporating AI and data science into data journalism, there are also the obstacles inherited from data journalism itself. These challenges are well explained in a comparative study between the United States and North European countries (Fink and Anderson 2015). Fink and Anderson summarize the main limiting factor for data journalism across newsrooms in the lack of: Time, Tools, Manpower and Legal Resources. They also point out that the role of a data journalist within an organization often lacks clear definition, frequently resulting in them either working in isolation or being burdened with an overwhelming amount of tasks.

An overview of the pros and cons of AI integration within newsrooms is given by Wu et al. after conducting interviews with professionals. In their work, they discuss the emergence of automation technologies, such as artificial intelligence, machine learning, and natural language processing, and their potential applications in journalism. They explore how these technologies are being integrated into newsrooms to streamline workflows, generate news content, and personalize news delivery. Some of the potentials of AI consist of increased efficiency, improved accuracy, and enhanced audience engagement through personalized news experiences. At the same time, concerns are related to job displacement, ethical considerations, and the need for human oversight in the automated news production (Wu, Tandoc, and Salmon 2019).

Nevertheless, AI journalism can play a crucial role in supporting the role of journalists as "watchdogs", to quote Tom Felle (Felle 2016), in providing the public with valuable information and holding those in power accountable for their actions. Of course, this comes not without challenges, but by focusing on disclosing sources, methodologies, and limitations journalists can enable audiences to assess the credibility and reliability of data-driven news stories (Anderson 2018; Zamith 2019).

Methodology

Parratt-Fernández et al. observe that 60% of academic work on AI applications in journalism utilizes qualitative methods, despite the numerical nature of the subject. While digital methods are prevalent in digital humanities, they are less common in journalism studies (Parratt-Fernández, Mayoral-Sánchez, and Mera-Fernández 2021). Sjøvaag and Karlsson attribute this to a higher threshold for journalism scholars, who often lack the necessary skills and knowledge to perform automated analysis on large datasets (Karlsson and Sjøvaag 2016). We have consequently chosen to develop and employ alternative methodologies to enrich the existing literature from complementary methodological perspectives. The methodology used in this research is practice-based (Biggs and Büchler 2007; Vear 2022).

Practice-based research emphasizes the study of real-world problems and practices, rather than solely theoretical or abstract concepts. In contrast to other methodologies, practice-based research focuses on understanding how people actually do things, rather than how they should theoretically do them. This allows us to focus on how investigative journalists do their job. It is an empirical research method, which is based on the

collection of data through observation, discussion, and other forms of direct engagement: practitioners and researchers work together to identify research questions, collect data, and analyse results.

Practice-based research allows journalists and scholars to better understand the practical implications of new technologies and changing newsroom practices. It is necessary to bridge the gap between academic research and industry practice and can lead to the development of more innovative and effective journalism (Barroca et al. 2018; Biesta 2007). In addition to its benefits for practitioners, practice-based research also profits academia by helping educators and researchers stay current with the latest trends and developments in the field. This results in the development of more relevant and effective journalism curricula (Niblock 2007; 2012; Robie 2015).

In designing the methodology of this study, we chose to conduct practice-based research by collaborating with the Norwegian Association for Investigative Journalism (Skup.No n.d.). SKUP is a non-profit organization that promotes investigative journalism in Norway. Together with SKUP, we published a call in May 2021 inviting investigative journalists to submit projects where we could assist using data science and AI techniques.

By opening a public call, we allocate financial resources to support academic staff dedicated to 4 projects for a period of 6 months. We encouraged all the Norwegian newsrooms to submit a project proposal and a committee of both AI and Journalism academics was appointed to select the 3 best ones. By doing so we encourage Norwegian newsrooms to explore avenues with which they are not familiar and that in normal circumstances they would have not pursued We received several applications and for the three selected projects, we dedicated an Associate Professor of Artificial Intelligence. Each team was then composed of one academic from OsloMet, one Investigative Journalist daily collaborating with the AI expert, and a staff of supporting journalists.

In our research approach, the investigative journalists were responsible for deciding the topic, finding the initial data sources, formulating the research questions, and interpreting the results. The teams had variable compositions in terms of the number of journalists and experience with data handling. A smaller team of data journalists assisted with data-related issues. Regular meetings ensured alignment between the data-focused teams and the journalists. This collaborative approach enabled a targeted use of data science and AI techniques in investigative journalism.

In 2022, we again selected three new projects and provided similar support.

The collaboration with SKUP not only provided access to experienced investigative journalists but also ensured that the research had a direct impact on the field.

The insights drawn from applying data science and AI techniques to investigative journalistic questions form the backbone of the present study. The journalistic results were published in various outlets. Here, we have anonymized identities to maintain confidentiality. Team meetings and Discord channels were utilized to ensure efficient collaboration among stakeholders. Group meetings were held to identify significant learning experiences and analyse findings considering the existing literature. For the purpose of this article, we draw on notes and discussions in team meetings and on Discord channel to reconstruct the research processes.

Description of Projects

Project 1

Exploration reimbursement scheme in the petroleum industry. The main goal of this project was to bring attention to the exploration reimbursement tax scheme in the petroleum industry. The journalists proposed to investigate and combine two publicly available data sources: the Petroleum Tax Lists (The Norwegian Tax Administration 2020) and the Norwegian Petroleum Directorate (NPD) Fact pages (NPD Fact Pages n.d.). The goal of the project was to better understand how companies were accessing and using the exploration reimbursement scheme, which has been in place since 2005. Additionally, the project aimed to provide some oversight to the program by mapping oil exploration activities, documented in the NPD fact pages, to the reimbursement claims in annual taxes. In the process of this investigation, the idea emerged to deploy graph databases to represent connections between petroleum companies, reimbursement payments and drilling licenses.

Project 2

Adverse events in elderly care. The main goal of this project was to better understand how the quality of elderly care varies across municipalities. The journalists proposed to correlate publicly available data recording "non-conformity reports" (NCR) and statistics from SSB. NCR are gathered at the municipality level and are regulated by the government guidelines "Norsk kodeverk for uønskede pasienthendelser" (Helsedirektoratet 2021). The statistics available from SSB describe a wide set of municipality parameters. By correlating these datasets, the project aimed to identify critical factors that contribute to the occurrence of adverse events in elderly care.

Project 3

Eating disorders in professional skiers. The main goal of this project was to investigate the occurrence of eating disorders among top athletes. The journalists proposed an innovative approach to solve this problem by using computer vision. The core idea was to take advantage of the mediatic exposure of the athletes. By collecting public images of the athletes over time it is possible to infer the body mass index and body fat percentage and consequently investigate drastic changes over the years.

Project 4

Media landscape analysis. The main goal of this project was to better understand the Norwegian media landscape. The project started with an exploratory data analysis of a dataset of Norwegian news and blog stories. The dataset collected information about the article itself, including title, publisher, authors, and links, as well as information about its social media engagement. In addition to exploratory analysis, this project grew to track how stories move across both traditional and social media.

Results

In the field of data journalism, the deployment of machine learning (ML) and artificial intelligence (AI) algorithms requires a thorough understanding of the necessary data and the feasibility of obtaining and processing that data. Based on existing literature, we expected this process to be time-consuming – particularly in investigative journalism projects where there is significant complexity in both the questions and societal structures involved. However, as we experienced, one should carefully consider the quality and availability of the required data to correctly estimate the feasibility of the project and its potential outcomes. It is crucial to note that input data of poor quality will result in equally poor output, regardless of the algorithm's computational power.

We found similar difficulties and challenges across all of three projects, which we could broadly categorize using the concepts introduced by Stray (Stray 2019b) as follows:

- Data availability: Data relevant to a story may not be publicly accessible, data collection results to be challenging, or the available dataset is incomplete/scattered
- Data quality: Journalistic inference requires high-accuracy data
- Newsworthiness: The concept of "newsworthy" is difficult to encode computationally

Data Collection is a Critical Aspect of AI in Journalism

We found that the availability of data was a recurring challenge across all projects. Despite identifying data sources before the project's start and therefore expecting that data was readily accessible, each project encountered difficulties in obtaining the necessary data.

The project exploring the reimbursement scheme in the petroleum industry stumbled across a lack of consistent historical records. Despite the presence of publicly available data from sources such as the Brønnøysund registry (brreg.no n.d.), historical data for defunct companies was often missing. This required the purchase of data or the use of alternative methods for collection. This project also suffered from a lack of transparent data from the side of the tax authorities. While the oil exploration reimbursement amounts are ostensibly available and published every year by the Norwegian Tax Office, these reimbursements are not split between exploration and termination reimbursements. The Norwegian Tax Office refused to release this information on request, without which it is impossible to evaluate the success or failure of the exploration reimbursement policy. Moreover, the life cycle of a company can be complex and involves merges and splits, takeovers and name changes. This makes it complicated to track licenses and understand financial flows over time. Petroleum extraction licenses are typically shared between several companies in alliances that might also change over time. It is consequently very difficult to reconstruct the timeline of company history and reimbursements without support from experts.

The project which sought to investigate the quality of elderly care in Norway also encountered challenges in obtaining data. Despite the requirement for municipalities to keep detailed records of non-conformity reports (NCR) in the elderly care sector, a comprehensive analysis of these reports had never been conducted. We discovered that often the data was missing, unstructured or in hard-to-access formats. There were different

reasons for this, ranging from a lack of human resources in the municipality to a lack of digital reporting systems. Privacy issues also prevented data sharing in the case of small municipalities. Even when the municipality invested significant resources to collect and anonymize the records, the data analysis was not trivial due to the lack of standardization across municipalities. In the best-case scenario, municipalities sent their data in the shape of Microsoft Excel spreadsheets. This required extensive effort from the team to restructure the data in a standard format and sometimes led to data loss. Other municipalities delivered data in unstructured formats such as PDFs, Word files or just printed out on A4 paper, leading to additional challenges wrangling the data into a usable format. At a later stage, the project intended to correlate the NCR statistics with publicly available data from the Norwegian Central Bureau of Statistics (SSB) (Statistisk sentralbyrå 2023) to better understand factors contributing to adverse events in elderly care facilities. SSB maintains updated statistics of many societal parameters; however, historical analysis of the data was challenging given the complex and frequent changes to the structure of Norwegian municipalities. Upon request, SSB supported our investigation by allowing our team to purchase restructured datasets, but these datasets were not provided freely through their platform. Since then, they have integrated our proposed indexing of the data in their report system for current and future data. This example shows how the structure and shareability of the data improved through the interaction between journalists and data collectors.

These experiences highlight the need for increased transparency and accessibility of data, particularly regarding business and healthcare. The lack of labeled data when looking for suspicious activity and the unstructured nature of business annual reports also presented significant obstacles in utilizing AI and ML algorithms. Considering these challenges, it is crucial to consider the feasibility of obtaining a complete and high-quality dataset before embarking on investigative journalism projects that involve data analysis.

Journalism would Benefit from Greater Transparency in Company Structure

In recent years, there has been a growing recognition of the need for greater transparency in the financial flows of companies, particularly in the extractive and financial sectors (Stiglitz 2002). This is because these sectors are often characterized by complex and opaque ownership structures (Sachs and Warner 2001). One of the main challenges in uncovering these financial flows is the lack of comprehensive and publicly available data on company ownership and beneficial ownership (Cobham and Janský 2020). While some countries have made progress in this area, for example by joining the Extractive Industries Transparency Initiative (EITI) (Extractive Industries Transparency Initiative n.d.), many countries still lack comprehensive and publicly available registers.

Additionally, even when data is publicly available, it is often unstructured, on paper and in a customized format, which makes it challenging to extract information even with state-of-the-art Object Character Recognition (OCR) algorithms. Furthermore, in some cases, the companies themselves may be unwilling to disclose information about their ownership and financial flows, making it difficult for investigative journalists to access the necessary data (Making Transparency Possible 2019). Even when a company is willing to disclose information, it might be difficult to access it as datasets might be

scattered across multiple platforms and in multiple countries, not to mention privacy legislations limiting data access (EU Court of Justice Delivers Blow to Beneficial Ownership … 2022).

While investigating the reimbursement scheme in the Norwegian petroleum industry we experienced how complex and opaque ownership structures are. This is a particularly illustrative example given that Norway, as a member of EITI, commits to disclose information regarding the extractive (Extractive Industries Transparency Initiative n.d.). There are numerous reasons why investigative journalists, and the news media, might want to investigate complex and opaque ownership structures in the extractive and financial sectors. First, opaque ownership structure makes it difficult to hold companies accountable for their actions, especially when it comes to taxes and royalties for the resources they extract. Several studies suggest that investigative journalism, and other forms of transparency-promoting activities, can play a critical role in exposing complex and opaque ownership structures in the extractive and financial sectors (Beckett 2019; Radon and Achuthan 2017). Based on our research, we believe data journalism in this field depends on initiatives such as OpenCorporates (OpenCorporates :: The Open Database Of The Corporate World n.d.) and OpenOwnership (Open Ownership n.d.) to move the field forward. These and other organizations are taking the lead in utilizing advanced Machine Learning and Graph Databases to analyse complex company structures and beneficial ownership.

Information is Often not Publicly Accessible

In the case of the investigation of accidents in elderly care, Norwegian municipalities are obliged to provide data to journalists. However, a significant number of municipalities did not provide the requested information. Various factors can influence whether municipalities provide requested information to journalists or not. However, as the context can vary from country to country and region to region, it is important to consider the specific circumstances and legal framework of each inquiry. Among the municipalities that did not provide the data in our investigation, the main reasons cited were a lack of digital records, on-going lawsuits that prevented the sharing of data and insufficient human resources to anonymize the data for privacy reasons. In addition, some municipalities simply failed to respond to the journalist's requests.

Additionally, there were also challenges in obtaining data from SSB (Statistisk sentralbyrå 2023) due to the frequent rearrangements of municipalities by the government. Frequent re-organization of municipalities leads to several challenges in terms of data analysis (Kommunereformen 2020, 2021). Administrative boundary changes can impact data quality, as data collection and reporting procedures may change, resulting in inconsistent or inaccurate data. This makes it difficult to conduct accurate and reliable analyses of parameters over the years.

In the project employing computer vision for BMI estimation, we negotiated with other researchers to share their datasets, some of which had been scraped from public sources like Reddit. While at first glance there were multiple approaches, public and shared datasets available, we quickly realized that these were not sufficient for the uses of the project. It is therefore important to highlight that it's not only the quantity of data that is crucial, but also their relevance to the research question.

Given the challenges of limited access to data, it is likely that journalists will need to invest significant resources to obtain the necessary data, as demonstrated by the examples presented in this article. Strategies such as collaborating with organizations or individuals who have access to relevant data sets, using publicly available data sources, or using alternative data sources can help obtain relevant datasets.

Journalistic Inference Requires Very High Accuracy

In the context of our practice-based research, it became clear that the process of journalistic inference demands high accuracy. Machine learning algorithms have been developed to identify patterns and common characteristics within datasets. However, when it comes to investigative journalism, particularly in the realm of fraud detection, the goal is to uncover the unusual, events that were not expected to occur. These subtle variations are crucial to consider, as they do not align with the primary purpose for which ML and AI algorithms were developed.

Defining what constitutes suspicious or unethical practices within the available datasets is a difficult task, due to obscure business practices and opaqueness in the law and interpretation of it. This is perhaps unsurprising, as even cases with extensive evidence have been ruled legal by the courts (Skattemotiverte transaksjoner – opplysningsplikt og fradragsrett 2020). More fundamentally, even defining what is a "company" over the years within a landscape of multiple organizations and leadership structures can be prohibitively complex, as we described in the previous chapter. This makes it precarious to associate any company or private entity with an accusation of misconduct.

Our research on the care of the elderly highlighted the difficulty in categorizing events into different classes. After evaluating multiple language models, we determined that accurate quantification of all categories could not be guaranteed. Thus, we decided to focus solely on adverse events related to medicine, which were successfully classified. It is important to note that we were not investigating causality in this project, but rather exploring correlations between variables, which can provide insight into relevant variables for optimizing the services offered by the municipality. Although it was not possible to quantify issues and problems related to elderly care within this landscape, our research aimed to guide the journalist's investigation towards addressing these issues, supported by data when the administration or leaders of the municipality were questioned.

A final aspect worth mentioning is the variation in data reporting among municipalities. Our analysis of collected databases revealed that each municipality, depending on its size and structure, may report data on finances and human resources with varying degrees of resolution. This makes the analysis process significantly more challenging, as municipalities may report a single budget for their entire health department while others may report individual budgets for different health departments (home assistance, hospitals, nursing homes, etc.), making comparisons difficult.

In our study on eating disorders in sports, we ultimately had to abandon the model due to insufficient accuracy. This serves as a crucial example of why expertise in AI is important. A team without sufficient knowledge of AI may not have been able to understand that the model was unreliable and unsuitable for use.

Finally, in the research we did with the fact-checking organization Faktisk.no we did succeed in identifying certain claims that could be confidently made and relayed, mainly because they were of a generalized, descriptive and qualitative nature.

Discussion

Data Preparation Tasks

According to Stray (Stray 2019a), data preparation tasks represent a significant opportunity for AI to benefit investigative journalism in the short term.

In the context of the petroleum exploration reimbursement investigation, we found that linking databases and transforming into Graph databases can reveal connections that were previously unknown, by identifying patterns and relationships within large amounts of data. This enhances the ability to uncover hidden links between entities and to build a comprehensive picture of the issue at hand.

In the care for the elderly project, we faced the challenge of merging information from various data types, including excel sheets, pdf files, and printed papers. Our research indicates that public bodies should play a role in standardizing data dissemination, similar to the example set by VG during the COVID-19 pandemic, where the government was weekly providing updated statistics to the media.

Finally, in our study with the fact-checking organization Faktisk.no, we discovered that most of the "AI", specifically unsupervised machine learning, was used in the pre-processing steps to cluster similar stories. This step greatly assisted in the later analysis, demonstrating the potential for AI in improving the efficiency of investigative journalism.

Our research supports the findings of Stray that data preparation tasks are the area where AI seems to have the most immediate impact on investigative journalism. However, much research and development are still needed to fully realize the potential of AI in this field and to ensure its ethical and effective application.

Cross-Database Record Linkage

The use of AI in investigative journalism presents significant potential for cross-database record linkage. The ability to link records across databases has the potential to greatly reduce the time, effort, and costs associated with many investigations while producing more robust results.

Referential integrity across databases refers to the consistency of relationships between records in different databases. Ensuring referential integrity is important in ensuring the accuracy of the information being analysed and can greatly assist in connecting records that would have been difficult to link otherwise. However, referential integrity can be difficult to achieve due to differences in keys and naming conventions across databases.

In the context of investigative journalism, referential integrity is crucial when linking records from various sources. For example, in the investigation of petroleum tax reimbursements, linking tax data with petroleum discovery data can help to hold companies accountable. However, in some cases we found that companies used different organization numbers for taxation and licensing purposes, which made linking records

difficult. Similarly, in the investigation of care for the elderly, referential integrity across databases is essential to ensure that information is accurate and correctly linked.

Our practice-based research highlights the potential of AI in facilitating cross-database record linkage, which could greatly enhance the effectiveness of investigative journalism.

Using the Right Tools Saves Time and Money and Enhances the Results' Quality

In our investigation, we identified a multitude of tools that can be adapted and leveraged for the purpose of investigation. These tools, including Pandas (Pandas - Python Data Analysis Library n.d.), Seaborn (Waskom 2021), Norwegian language models such as NorBERT or NoTraM (Web64 2016/2023), zero-shot classification (NbAiLab/Nb-Bert-Base-Mnli Hugging Face 2023), Neo4j (Neo4j Graph Data Platform – The Leader in Graph Databases n.d.), NetworkX (NetworkX — NetworkX Documentation n.d.), and HuggingFace (Hugging Face – The AI Community Building the Future. n.d.), represent only a selection of the numerous available options that can be utilized to fulfill specific needs. It is of utmost importance to have a comprehensive understanding of recent advancements in AI to make informed decisions when selecting the most appropriate tools and platforms for a given investigation.

Pandas, a data analysis library in Python, provides efficient storage and manipulation of tabular data through its data structures. Seaborn, another Python library, offers a high-level interface for generating informative and visually appealing statistical graphics. NB-NERT, a Norwegian language model for named entity recognition (NER), can be utilized to identify named entities in text. Zero-shot classification, a machine learning technique, enables categorization of unseen categories without the need for any additional training data. Neo4j is a graph database management system designed for the storage and querying of complex networked data. NetworkX, a Python library, enables the creation, manipulation, and analysis of the structure, dynamics, and functions of complex networks. Lastly, HuggingFace is a natural language processing platform providing access to a wide range of pre-trained language models, including NER models.

These tools can be employed in a variety of ways to support journalistic investigations. They can be used to analyze data and generate visualizations to gain a deeper understanding of trends and patterns, identify named entities in text, categorize text, store and query complex networked data, and access pre-trained language models for NER tasks. In our research project, we invested significant resources to support the four projects. Two associate professors were dedicated to the projects and worked almost full-time. Three masters students also contributed to the work. Additionally, two other professors provided support, including time spent on funding and administrative tasks. All in all, we estimate that the total cost of the projects was close to $200,000.

The sizes of the projects varied greatly. Two of the projects were large-scale efforts that involved a substantial team of journalists and developers. At times, each team could consist of over ten individuals. The cost related to data science and AI was a small portion of the overall budget for the largest projects, estimated to be less than 10%. On the other hand, one of the smaller projects had a significantly higher proportion of its budget allocated to data science and AI, approximately half of the total. However, it should be noted that it is challenging to provide precise estimates as several journalists and developers worked on multiple projects simultaneously.

The projects presented significant challenges in terms of budgeting and resource allocation. The extent of the data pre-processing required was unforeseen and caused unforeseen expenses. When evaluating the costs of investment in data science and AI, it is important to weigh them against the potential benefits. In one instance, having machine learning experts working on the project prevented the publication of unreliable results. In another instance, the potential benefits could be measured in terms of improvements to quality of life and longevity. While it is difficult to determine the financial returns of these investigations to the news media, we believe that they hold great promise for long-term benefit to society.

Importantly, the advent of AI in the media industry is already changing the professional profile of the journalist, since they have to manage constantly developing technologies, an increasing amount of accessible data, fake news generation, and last but not least, the ethical implications introduced by the use of AI in modern society (Túñez-López, Fieiras-Ceide, and Vaz-Álvarez 2021). Therefore, appropriate training of journalists, as well as a productive collaboration with experts in the fields of AI, is necessary to allow them to automate repetitive tasks and to process large amount of data efficiently so that they can focus on creating high-quality human-crafted journalism (Noain-Sánchez 2022).

Conclusion

In this study, we applied an interdisciplinary approach to enhance investigative journalism with advanced machine learning and data science techniques. We utilized a variety of tools, including Pandas, Seaborn, Norwegian language models like NB-BERT, zero-shot classification, Neo4j, NetworkX, and HuggingFace to build applications that made the investigative process more cost-effective. During the course of this project, we observed that the landscape of large language models and computer vision was rapidly changing, with the release of four major neural networks in the latter half of 2022. This trend is likely to continue, with the speed of innovation and openness in the field leading to falling costs of investigative projects.

Based on our experiences, we believe that it is crucial to move beyond interdisciplinary projects and towards true trans-disciplinary projects. “Trans-disciplinary” refers to a collaborative approach that integrates knowledge and skills from multiple disciplines to solve complex problems and address real-world challenges. It differs from interdisciplinary teamwork in that it involves active participation and collaboration from all stakeholders including those outside the field of expertise, to ensure that the solutions produced are holistic and relevant to the real-world context. Additionally, it requires participants to develop joint theoretical and methodological frameworks to guide the teamwork.

In order to maximize the chances of success we have developed a recommended workflow summarized in Figure 1. We recommend prioritizing projects with well-stated research questions and a clear hypothesis, while also considering the potential value of the database itself, even in the absence of a “smoking gun”. Moreover, it is important to keep in mind that historical data can be missing, incomplete and difficult to interpret, therefore, we recommend focusing on projects based on recent or current time periods. Happily, data are becoming increasingly available, thanks to the active contribution of both private and public bodies, such that hopefully this restriction will diminish with

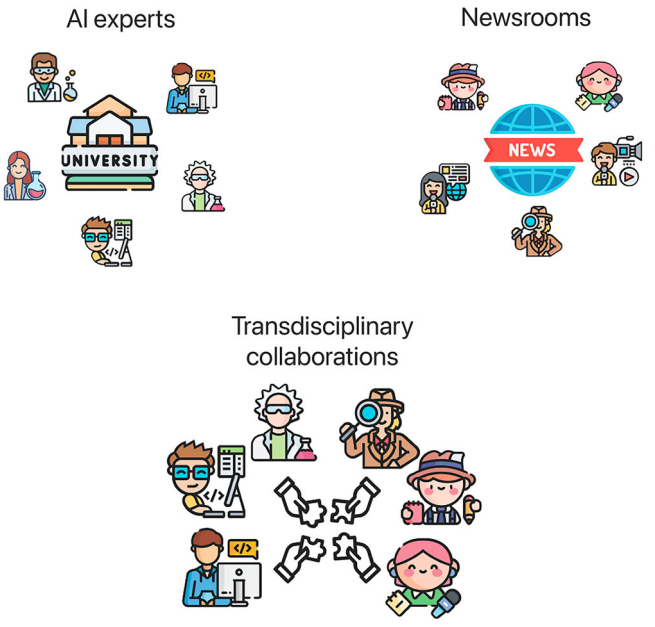

Figure 1. Recommended workflow to successfully implement and develop an AI Journalism project. This recommendation is based on the practice and experience of the projects we describe in this article, and it aims to guide other journalists on how to implement AI in their newsrooms.

time. Additionally, we suggest using explainable methods and visualization techniques that are easily understandable to journalists, editors, and the general public.

In addition to the tangible outcomes of these projects in the form of news reports and documentaries, we also gained valuable insights into the challenges and complexities of these types of collaborations. We found that the investigations were more productive when both journalists and AI specialists became literate in each other's fields and

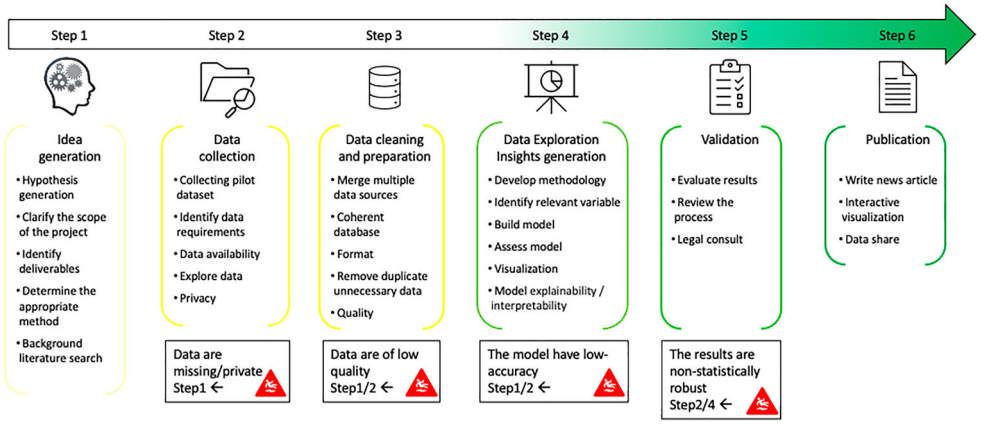

Figure 2. Schematic representation of what we envision as trans-disciplinary collaborations between AI experts and Investigative Journalists. These two professional figures do not work independently, but actively collaborate with each other integrating their skills to solve complex problems.

engaged in a mutual learning process, as represented in Figure 2. This highlights the importance of using existing algorithms, programs, and models, as well as developing an understanding of the broader range of techniques in AI and data analysis.

Data journalism has emerged as an important field in contemporary journalism. The rise of big data, open data and data visualization technologies have enabled journalists to leverage data in innovative ways to tell more compelling stories. Data journalism offers a new way of reporting that is grounded in the analysis of large data sets, and that allows for the creation of new insights that would not be possible through traditional reporting methods. In this paper, we explore the role of data journalism in contemporary journalism and examine the ways in which it is transforming the field of journalism.

Acknowledgment

We thank Gustavo Borges Moreno e Mello for the main contribution in establishing the "The AI Journalism Resource Center" and supporting the group in obtaining the necessary fundings supporting the work presented in this article. We also thank Morten Goodwin for contributing to the development of the projects with stimulating discussions and insight.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by Artificial Intelligence Lab (AI Lab), Institutt for informasjonsteknologi, Oslo Metropolitan University, Oslo, Norway and Fritt Ord Foundation and the Norwegian Directorate for Higher Education and Skills.

ORCID

M. Fridman http://orcid.org/0000-0002-3065-8888
R. Krøvel http://orcid.org/0000-0003-2231-7714
F. Palumbo http://orcid.org/0000-0002-5571-5420

References

AI and the Future of Journalism. n.d. Accessed September 3, 2021. http://ulam.ai/ai-and-the-future-of-journalism/.

Amazon Mining Watch. n.d. Accessed February 27, 2023. https://amazonminingwatch.org/en.

Anderson, C. W. 2018. *Apostles of Certainty*. Vol. 1. Oxford University Press. https://doi.org/10.1093/oso/9780190492335.001.0001.

Ausserhofer, J., R. Gutounig, M. Oppermann, M. Oppermann, S. Matiasek, and E. Goldgruber. 2017. "The Datafication of Data Journalism Scholarship: Focal Points, Methods, and Research Propositions for the Investigation of Data-Intensive Newswork." *Journalism: Theory, Practice & Criticism* 21: 950–973. https://doi.org/10.1177/1464884917700667.

Barroca, L., H. Sharp, D. Salah, K. Taylor, and P. Gregory. 2018. "Bridging the Gap between Research and Agile Practice: An Evolutionary Model." *International Journal of System Assurance Engineering and Management* 9 (2): 323–334. https://doi.org/10.1007/s13198-015-0355-5.

Beckett, P. 2019. *Ownership, Financial Accountability and the law: Transparency Strategies and Counter-Initiatives*. London: Routledge, Taylor & Francis Group.

Biesta, G. 2007. "Bridging the Gap between Educational Research and Educational Practice: The Need for Critical Distance." *Educational Research and Evaluation* 13 (3): 295–301. https://doi.org/10.1080/13803610701640227.

Biggs, M. A. R., and D. Büchler. 2007. "Rigor and Practice-Based Research." *Design Issues* 23 (3): 62–69. https://doi.org/10.1162/desi.2007.23.3.62.

Borges-Rey, E. 2016. "Unravelling Data Journalism a Study of Data Journalism Practice in British Newsrooms." *Journalism Practice* 10 (7): 833–843. https://doi.org/10.1080/17512786.2016.1159921.

Bounegru, L., and J. Grey, eds. 2021. *The Data Journalism Handbook: Towards a Critical Data Practice*. Amsterdam: Amsterdam University Press.

Brreg.no. n.d. Brønnøysundregistrene. Accessed February 28, 2023. https://www.brreg.no.

Buolamwini, J., and T. Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, edited by S. A. Friedler, and C. Wilson, 77–91. Vol. 81. PMLR. https://proceedings.mlr.press/v81/buolamwini18a.html.

Challenges and Opportunities – Survey – .State of .Data .Journalism 2022. n.d. Accessed February 28, 2023. https://datajournalism.com/survey/2022/challenges-and-opportunities/.

Cobham, A., and P. Janský. 2020. *Estimating Illicit Financial Flows: A Critical Guide to the Data, Methodologies, and Findings*. 1st ed. Oxford University PressOxford. https://doi.org/10.1093/oso/9780198854418.001.0001

Crawford, Kate, and Jason Schultz. 2014. "Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms." *Boston College Law Review* 55: 93. https://ssrn.com/abstract=2325784.

de-Lima-Santos, M.-F., and W. Ceron. 2022. "Artificial Intelligence in News Media: Current Perceptions and Future Outlook." *Journalism and Media* 3 (1): 13–26. https://doi.org/10.3390/journalmedia3010002.

De-Lima-Santos, M. F., and R. Salaverría. 2021. "From data journalism to artificial intelligence: challenges faced by La Nación in implementing computer vision in news reporting." *Palabra Clave* 24 (3).

Dörr, K. N. 2016. "Mapping the Field of Algorithmic Journalism." *Digital Journalism* 4 (6): 700–722. https://doi.org/10.1080/21670811.2015.1096748.

EU Court of Justice delivers blow to beneficial ownership … . 2022, November 22. Transparency.Org. https://www.transparency.org/en/press/eu-court-of-justice-delivers-blow-to-beneficial-ownership-transparency.

Extractive Industries Transparency Initiative. n.d. EITI. Accessed February 28, 2023. https://eiti.org/.

Felle, T. 2016. "Digital Watchdogs? Data Reporting and the News Media's Traditional 'Fourth Estate' Function." *Journalism* 17 (1): 85–96. https://doi.org/10.1177/1464884915593246.

Fink, K., and C. W. Anderson. 2015. "Data Journalism in the United States: Beyond the "Usual Suspects"." *Journalism Studies* 16 (4): 467–481. https://doi.org/10.1080/1461670X.2014.939852.

Hacks/Hackers LDN (Director). 2019, November 29. AI & Journalism: New powers, new responsibilities \textbar Charlie Beckett. https://www.youtube.com/watch?feature=youtu.be&v=L-qgP14TK8U&app=desktop.

Helsedirektoratet. 2021. *Norsk kodeverk for uønskede pasienthendelser*. https://www.helsedirektoratet.no/rapporter/norsk-kodeverk-for-uonskede-pasienthendelser/Norsk%20kodeverk%20for%20uønskede%20pasienthendelser.pdf/_/attachment/inline/e95247b1-bdb4-463b-b730-5a09398db917:88e99f1e911c29fd8101025ad12f685eef995b9c/Norsk%20kodeverk%20for%20uønskede%20pasienthendelser.pdf.

Hugging Face – The AI community building the future. n.d. Accessed March 1, 2023. https://huggingface.co/.

Infographic: The Scale Of The Pandora Papers Leak. 2021, October 4. Statista Infographics. https://www.statista.com/chart/11698/the-scale-of-the-paradise-papers-leak.

Jain, A., H. Patel, L. Nagalapatti, N. Gupta, S. Mehta, S. Guttula, S. Mujumdar, S. Afzal, R. Sharma Mittal, and V. Munigala. 2020. "Overview and Importance of Data Quality for Machine Learning Tasks." *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3561–3562. https://doi.org/10.1145/3394486.3406477.

Karlsson, M., and H. Sjøvaag. 2016. "Content Analysis and Online News: Epistemologies of Analysing the Ephemeral Web." *Digital Journalism* 4 (1): 177–192. https://doi.org/10.1080/21670811.2015.1096619.

Kommunereformen 2020. 2021, March 15. ssb.no. https://www.ssb.no/offentlig-sektor/kommune-stat-rapportering/kommunereformen-2020.

Making Transparency Possible. 2019. Cappelen Damm Akademisk/NOASP. https://doi.org/10.23865/noasp.64

Miroshnichenko, A. 2018. "AI to Bypass Creativity. Will Robots Replace Journalists? (The Answer Is "Yes")." *Information* 9 (7): 183. https://doi.org/10.3390/info9070183.

NbAiLab/nb-bert-base-mnli Hugging Face. 2023, January 25. https://huggingface.co/NbAiLab/nb-bert-base-mnli.

Neo4j Graph Data Platform – The Leader in Graph Databases. n.d. Neo4j Graph Data Platform. Accessed March 1, 2023. https://neo4j.com/.

NetworkX — NetworkX documentation. n.d. Accessed March 1, 2023. https://networkx.org/.

Niblock, S. 2007. "From "Knowing How" to "Being Able": Negotiating the Meanings of Reflective Practice and Reflexive Research in Journalism Studies." *Journalism Practice* 1 (1): 20–32. https://doi.org/10.1080/17512780601078829.

Niblock, S. 2012. "Envisioning Journalism Practice as Research." *Journalism Practice* 6 (4): 497–512. https://doi.org/10.1080/17512786.2011.650922.

Noain-Sánchez, A. 2022. "Addressing the Impact of Artificial Intelligence on Journalism: The Perception of Experts, Journalists and Academics." *Communication & Society* 35 (3): 105–121. https://doi.org/10.15581/003.35.3.105-121.

The Norwegian Tax Administration. 2020. *Petroleumsskatt på 116 milliarder kroner for 2019*. https://www.skatteetaten.no/en/presse/nyhetsrommet/petroleumsskatt-pa-116-milliarder-kroner-for-2019/.

NPD Fact Pages. n.d. Accessed May 1, 2022. https://factpages.npd.no/.

O'Neil, C. 2016. *Weapons of Math Destruction: How big Data Increases Inequality and Threatens Democracy*. 1st ed. Crown.

OpenCorporates: The Open Database Of The Corporate World. n.d. Accessed February 28, 2023. https://opencorporates.com/.

Open Ownership. n.d. Openownership.Org. Accessed February 28, 2023. https://www.openownership.org/en/.

pandas—Python Data Analysis Library. n.d. Accessed March 1, 2023. https://pandas.pydata.org/.

Pandora Papers: An offshore data tsunami - ICIJ. 2021, October 6. https://web.archive.org/web/20211006063105/https://www.icij.org/investigations/pandora-papers/about-pandora-papers-leak-dataset/.

Parratt-Fernández, S., J. Mayoral-Sánchez, and M. Mera-Fernández. 2021. "Aplicación de la inteligencia artificial al periodismo: Análisis de la producción académica." *El Profesional de La Información*, e300317. https://doi.org/10.3145/epi.2021.may.17.

Radon, J., and M. Achuthan. 2017. "Beneficial Ownership Disclosure." *Journal of International Affairs* 70 (2): 85–108. JSTOR.

Robie, D. 2015. "Advocating Journalism Practice-as-research: A Case for Recognition in the New Zealand PBRF Context." *Asia Pacific Media Educator* 25 (1): 62–73. https://doi.org/10.1177/1326365X15575591.

Rodríguez, M. T., S. Nunes, and T. Devezas. 2015. "Telling Stories with Data Visualization." *Proceedings of the 2015 Workshop on Narrative & Hypertext - NHT* 15: 7–11. https://doi.org/10.1145/2804565.2804567.

Sachs, J. D., and A. M. Warner. 2001. "The Curse of Natural Resources." *European Economic Review* 45 (4–6): 827–838. https://doi.org/10.1016/S0014-2921(01)00125-8.

Santos, M. F. D. L., and M.-F. de-Lima-Santos. 2022. "ProPublica's Data Journalism: How Multidisciplinary Teams and Hybrid Profiles Create Impactful Data Stories." *Media and Communication*, https://doi.org/10.17645/mac.v10i1.4433.

Skattemotiverte transaksjoner – opplysningsplikt og fradragsrett. November 13, 2020. HR-2020-2200-A (Høyesterett (Norwegian Supreme Court)). https://www.domstol.no/no/hoyesterett/avgjorelser/2020/hoyesterett-straff/hr-2020-2200-a/.

Skup.no. n.d. Accessed February 28, 2023. https://www.skup.no/.
Stalph, F. 2018. "Classifying Data Journalism: A Content Analysis of Daily Data-Driven Stories." *Journalism Practice* 12 (10): 1332–1350. https://doi.org/10.1080/17512786.2017.1386583.
The State of Data Journalism (No. 2022). 2023. datajournalism.com. https://datajournalism.com/survey/2022/.
Statistisk sentralbyrå. 2023, February 28. SSB. https://www.ssb.no/.
Stiglitz, J. 2002. "Transparency in Government." In *The Right to Tell: The Role of Mass Media in Economic Development*. 1st ed. World Bank Publications.
Stray, J. 2019a. "Making Artificial Intelligence Work for Investigative Journalism." *Digital Journalism* 7 (8): 1076–1097. https://doi.org/10.1080/21670811.2019.1630289.
Túñez-López, J.-M., C. Fieiras-Ceide, and M. Vaz-Álvarez. 2021. "Impact of Artificial Intelligence on Journalism: Transformations in the Company, Products, Contents and Professional Profile." *Communication & Society* 34 (1): 177–193. https://doi.org/10.15581/003.34.1.177-193.
Vear, C. ed. 2022. *The Routledge International Handbook of Practice-Based Research*. Routledge.
Waskom, M. 2021. "seaborn: Statistical Data Visualization." *Journal of Open Source Software* 6 (60): 3021. https://doi.org/10.21105/joss.03021.
Web64. 2023. Norwegian NLP Resources. https://github.com/web64/norwegian-nlp-resources (Original work published 2016).
Weber, M. 2021. "AI, Media and the Future of News on the Web." *13th ACM Web Science Conference* 2021: 10. https://doi.org/10.1145/3447535.3468474.
Wu, S., E. C. Tandoc, and C. T. Salmon. 2019. "When Journalism and Automation Intersect: Assessing the Influence of the Technological Field on Contemporary Newsrooms." *Journalism Practice* 13 (10): 1238–1254. https://doi.org/10.1080/17512786.2019.1585198.
Young, M. L., A. Hermida, and J. Fulda. 2018. *What Makes for Great Data Journalism. Journalism Practice* 12 (1): 115–135. https://doi.org/10.1080/17512786.2016.1270171.
Zamith, R. 2019. "Transparency, Interactivity, Diversity, and Information Provenance in Everyday Data Journalism." *Digital Journalism* 7 (4): 470–489. https://doi.org/10.1080/21670811.2018.1554409.
Zhao, J., T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. https://doi.org/10.48550/ARXIV.1804.06876.