Algorithmic Journalism and Its Impacts on Work

Ayse Dalgali†
School of Information Studies
Syracuse University
USA
ayocal@syr.edu

Kevin Crowston
School of Information Studies
Syracuse University
USA
crowston@syr.edu

ABSTRACT

In the artificial intelligence era, algorithmic journalists can produce news reports in natural language from structured data thanks to natural language generation (NLG) algorithms. This paper presents several algorithmic content generation models and discusses the impacts of algorithmic journalism on work within a framework consisting of three levels: replacing tasks of journalists, increasing efficiency, and developing new capabilities within journalism. The findings indicate that algorithmic journalism technology may lead some changes in journalism by enabling individual users to produce their own stories. This paper may contribute to an understanding of how algorithmic news is created and how algorithmic journalism technology impacts work.

KEYWORDS

Algorithmic journalism, Automated journalism, Robot journalism, AI journalism

1 Introduction

In the present era, artificial intelligence (AI) applications address diverse tasks including image recognition, machine translation, and guidance for automated vehicles. And in journalism, they are now used for much harder cognitive tasks, such as content generation, content editing, combining databases with editor-created story templates to generate stories, and editing reports, all of which journalists used to perform before. Companies such as Arria, AX Semantics, Retresco, Automated Insights, Narrative Science, Associated Press, and Gannett [8,13] utilize algorithmic journalism. For instance, the Associated Press generates around 3,700 earnings reports on US and Canadian companies using AI technologies [13]. Narrative Science produces algorithmic news from economic indicators and game reports [11]. Yahoo uses Wordsmith to prepare texts for fantasy sports games [8]. In this paper, we define algorithmic news, discuss the underlying technology and discuss the impacts on the work of journalists.

C+J 2020,20 March 2020, Boston, MA USA

2 Algorithmic News

To describe the use of AI technologies in journalism, different terms are used such as *AI journalism, automated journalism* and *robot journalism* [5]. In [9], the term *automated journalism* is defined as "any process or system of news production under the control of media or electronic devices, with little or no external influence" (p. 6). Ref [3] also uses the term *automated journalism* and the definition used in that source is cited by many articles: "journalism in which a program turns data into a news narrative, made possible with limited—or even zero—human input" (p. 416). On the other hand, the term *robot journalism* might be understood to mean "physical robots" that are envisioned as replacing newscasters in newsrooms [10], making it less appropriate for our purposes. In this paper, we use the term *algorithmic journalism*, following Dörr's definition [7]:

> the (semi)-automated process of NLG (natural language generation) by the selection of electronic data from private or public databases (input), the assignment of relevance of pre-selected or non-selected data characteristics, the processing and structuring of the relevant data-sets to a semantic structure (throughput), and the publishing of the final text on an online or offline platform with a certain reach (output). (p.702).

To define the news created by these processes, we will use the term *algorithmic news*, and to define the algorithms that generate algorithmic news, we will use the term *algorithmic journalists*. *Algorithmic news* refers news reports generated by algorithmic journalism. According to Dörr's definition, the output is described as the final text published, and the throughput is described as a semantic structure.

Algorithmic news reports are publishable texts that have a semantic structure created by algorithms from data. The main technology used in algorithmic journalism [1,4,7] is NLG, a subfield of natural language processing (NLP), which describes a software process in which structured data are converted into human (natural) language. Other technologies may also be used to generate news content. For example, Narrative Science and Automated Insights add graphics and pictures to their generated texts [7]. These additions are different from NLG [7], but these visualization tools, which improve the diversity of the content, may help to generate more

enjoyable and attractive news. In many applications, human intervention is still needed; that is, the process may not be fully automated.

Dörr's algorithmic journalism model, an input-throughput-output (I-T-O) model (shown in Figure 1), is based on the model proposed by Latzer, Just, and Saurwein. Dörr notes that algorithms behind algorithmic journalism apply the rules of NLG [7]. Hence, in the I-T-O algorithmic journalism model, NLG also carries out algorithmic selection. The process of generating news algorithmically starts with a database, such as sports, financial, weather, or traffic data (input). Next, the data are converted according to predefined linguistic and statistical rules (throughput) into a text (output) in natural language [7].

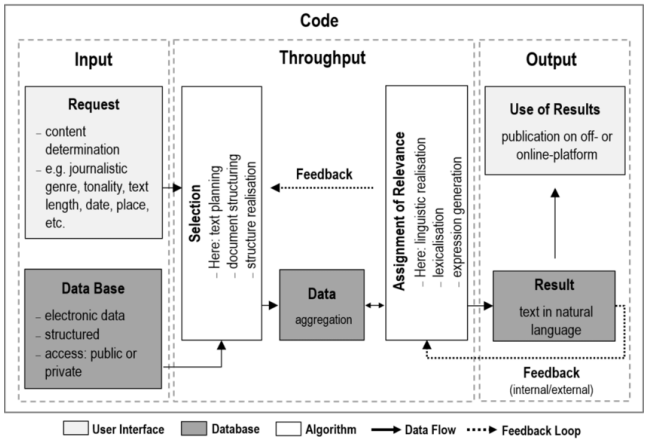

Figure 1: I-T-O model of algorithmic journalism [7].

In Dörr's model, the content is generated in three stages: (1) text planning, (2) document structuring, and (3) structure realization. The purpose of content determination (input level) is to decide which information is helpful to the user or important for the expected output. The input to the text planner (i.e., structured data) is the input to the entire process of NLG. This structured data is accessible through public application programming interfaces (APIs) or through private databases (e.g., commercial data). Because algorithmic news, based on NLG, consists of content-related texts, to generate these texts, specific codes, rules, and dictionaries are used and adapted. Hence, NLG operates based on pre-set special rules regarding the linguistic creation process and criteria for identifying and selecting facts in the data to be processed and transformed into natural language [7].

In the planning stage, at the input level (request), features such as text length, journalistic genre, tonality, and the time and place of publication are determined. In the throughput stage, based on the criteria arrived at in the planning stage, the NLG algorithm selects components from the data set, and aggregates and assigns relevance to them. After this process, the algorithm identifies the linguistic structures (words, syntax, sentences) to be used to achieve the desired information, the forms of words to use, and their order of appearance. More specifically, the NLG algorithm makes lexical choices and decides which content and words should be used to explain domain concepts; makes syntactic choices and decides which syntactic structures should be used in generating sentences; and aggregates data and decides how many messages should be included in each sentence. After the text in natural language (i.e., the output result) is generated, humans intervene in the process through the feedback loops to optimize this generation until the intended result is accomplished. Finally, after the text-generation process is completed, the texts are often published automatically on online or offline news outlets [7].

Although Dörr's definition includes the term *semantic structure*, previous NLG models that generate news were only simple descriptions of routine sports and financial news [7]. The I-T-O model and other previous NLG algorithms for generating news did not contain semantic features. Therefore, they usually produce only short, simple descriptive news in limited domains, rather than more complex news like that generated by humans, such as event-driven narratives [4]. In [4], Caswell and Dörr relate this problem to the absence of semantic elements in data and the absence of appropriate data models (methods for processing data) from the production of more complex algorithmic news. The common previous method for algorithmic journalism used trees and templates, for example, if the data field 1 is X, then write Y, which creates story templates. After these story templates are created, when new data are collected, the templates are filled out by the new data based on the trees.

To solve these problems, Caswell and Dörr propose a model also using semantic features to generate event-driven narratives. Caswell and Dörr's model is based on a "story database." Journalists enter events and narratives into this database, which uses the semantics, or meanings, of journalistic events to categorize news stories in the form of structured data. This model merges NLG with structured data that represent stories semantically in a story database to generate complex journalistic narratives as illustrated in Figure 2.

To contribute to the story database, journalists enter data according to the semantic features of actual news reports. For example, if the news report is related to commerce, it is stored under the "commerce" semantic frame, which entails certain roles, including buyer, seller, etc., and certain actions, including buying, selling, paying, etc. In Caswell and Dörr's model, first groups of related semantic frames from structured stories are chosen. Then, for each group of semantic frames, an appropriate template created by Wordsmith with blanks is filled out with the relevant structured data obtained from the story database, based on the semantic features and context of the news report to be generated. Then, these completed templates (i.e., text blocks) are combined to generate a complete news report.

In 2015, Wordsmith, an artificial writer (i.e., and algorithmic journalist) using NLP, developed by Automated Insights, one of the leading commercial providers of NLG technologies, became

available [12]. Wordsmith branches paths (i.e., creates trees) by adding words, sections, or phrases, or modifying or removing them [12]. A Wordsmith user enters data, such as criminal records, and then Wordsmith builds branches around that data. This process constitutes a story structure, which is used as a template for numerous further articles. A sample report concerning crime trends produced and shared by Automated Insights is shown in Figure 3 [12].

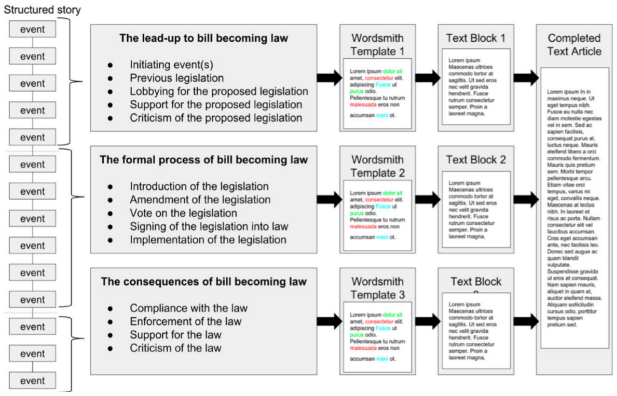

Figure 2: Generating algorithmic news from structured stories [4].

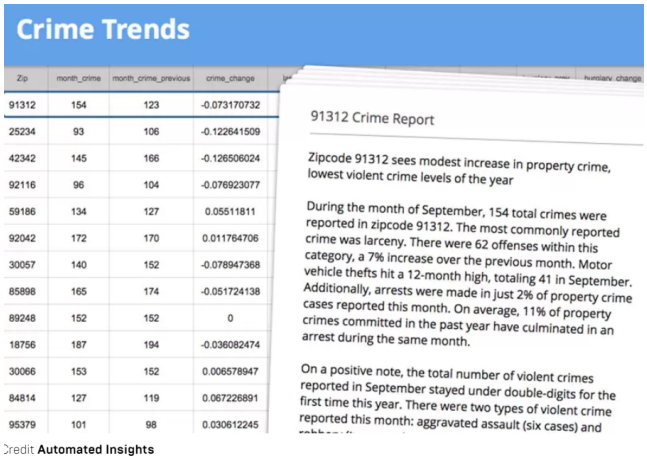

Figure 3: Example crime report shared by Automated Insights [12].

3 Impacts of Algorithmic Journalism on Work

In this section, we will discuss the impacts of algorithmic journalism on work. We will analyze these impacts within a framework consisting of three levels: replacing tasks of journalists, increasing efficiency, and developing new capabilities within the journalism.

Algorithmic journalism can take over some tasks like producing weather reports and sports stories. This taking over, however, does not correspond to job losses of journalists. In the short term, the concern about unemployment seems unrealistic because these algorithms are used by humans entering data, processing models (i.e., NLG rules or statistical rules, etc.), or checking models; therefore, they are not independent of humans. The Associated Press also noted that "algorithmic journalists" have not caused any human job losses so far.

As for increasing efficiency, algorithmic journalism may reduce the time and costs for the creation of news reports. Moreover, algorithms can analyze a huge amount of data to generate news; therefore, they can produce a lot of news, such as sport news, financial news reports, crime reports. As result, it is possible to publish stories that otherwise would not have been written. On the other hand, journalists sharing their workload with the algorithmic journalists may have more time for more creative tasks, or for tasks like checking the news generated by algorithms. Thus, collaboration of journalists with algorithmic journalists may increase the quantity of news reports.

Finally, algorithmic journalism may bring with it some new capabilities into current journalism practices. For example, if Wordsmith or similar tools to be developed are open to public use, this may lead to a transformation in journalism from news written by journalists in newsrooms and published to everyone to more individual preferences-oriented news. Robbie Allen, the CEO of Automated Insights, indicated that Wordsmith can generate articles much faster than even the fastest writer: it generated more than 1.5 billion pieces of content in 2015, up from 300 million in 2013 [8]. Allen added that instead of writing one story to present it to a million people, Wordsmith provides the opportunity to create individual stories for a million users according to their specific preferences through their participation in the process of news generation [12]. Each story is specific to the user because it is powered by their data [12].

4 Discussion

In this paper, several adopted models were presented that illustrate the steps that are used to generate news content by algorithms, such as decision trees and NLP, and potential effects on algorithmic journalism on work was discussed. It is controversial whether the use of algorithmic journalism is beneficial or not. Whereas some have advocated that algorithmic journalists will augment journalists by helping them to generate news at higher speed using Big Data, others have argued the news generated by algorithms will not be as effective as news written by human journalists because algorithms do not have emotions, values, creativity, and so forth. Jon Bernstein, an editor, writer, and digital media consultant, endorses this statement asserting that journalism is not just about presenting information, it is actually explaining what the presented information means [8]. Bernstein adds that automatically generated story content fails to explain the "why" of the story, which requires the ability to analyze and infer. Furthermore, when presented with two articles in an informal poll, NPR listeners preferred the one

written by the NPR White House correspondent to the one written by Wordsmith, emphasizing that while Wordsmith is faster than a human journalist, the human-written article was richer and more engaging [12]. Furthermore, some claim that an algorithmic journalist could never beat a human journalist's style or insight [12]. Thus, even if algorithmic journalism is getting more advanced, it is ambiguous whether algorithmic journalists will be as effective as human journalists.

Even though Wordsmith enables the opportunity users to enter their own data and create their own stories [12], the stories from the input data and the statistical or NLG rules and templates will be monotypic or very similar (because as input data and rules change, the output changes) and lack human creativity, different human emotions, different human perspectives, and different humans' different writing styles, which make the news more engaging and richer. Although Wordsmith adds emotive languages with appropriate syntax and diction to generate more readable news [8], it still lacks more complex emotions that are specific to humans and vary from person to person.

Moreover, other problems with algorithmic journalism concern authorship, credibility, quality, unemployment that affects journalists, and ethical concerns, such as risks of violation of journalistic ethics, the lack of interlocutors to take responsibility for the violation of ethical rules, fake news, news with errors or with mistakes, etc. However, these potential problems are not unsolvable. In the example of Wordsmith, if a user enters data and creates a story, in this case, the user may be considered the author of the story. The quality and credibility of the stories may depend on the data and data models.

As for journalistic ethics, there are concerns regarding the potential emergence of problems concerning transparency, privacy, bias, etc. Another main concern is that if ethical rules are violated, who will be responsible for this violation? Thurman et al. examined what journalists think about algorithmic journalism by conducting workshops and semi-structured interviews with ten professional journalists from the BBC, CNN, and Thomson Reuters [13]. A journalist stated that if the readers do not know how the news is generated, whether by a human journalist or by an algorithmic journalist, this may cause problems, such as not being able to determine who or what should be credited and held responsible for the output [13]. There could be violations of policy-related rules relating to privacy, security, misinformation, disinformation, fake news, etc., or mistakes contained in news stories due to errors in data. This question concerns liability. To solve this problem, recommends providing algorithmic transparency by providing data, the model used to process the data, and the results, including errors [6]. However, if this transparency reveals source code and a detailed-enough methodology, may be used to violate anonymity or privacy. Thus, Diakopoulos suggests that transparency should be provided to some extent, not full transparency nor lack of transparency, but balanced transparency for appropriate people.

Concerning bias, some journalists think that algorithmic journalism reduces bias that stems from human subjectivity [13]. Nevertheless, some of them argue that it might increase bias because of possible data manipulation by humans desiring to generate biased news by using biased data or biased models that process the data in automatic news generation. Regarding verification, although algorithmic journalism can eliminate human error, it hinders readers from verifying whether the data source is valid and reliable [13].

Algorithms may generate biased news because of biased data obtained from biased sources with which the models are fed. In addition, these algorithms may be misused, or accidents may occur while using them. For example, data to be used as input may be tainted with misinformation (false or inaccurate information which unintentionally deceives others), which may result in generating misinformation; input data may not correspond 100% with the templates, or with the NLG rules to be used, which may produce misinformation. Furthermore, input data may be created to manipulate readers according to the data providers' purposes. Also, these news reports containing manipulated data would be perceived by readers as objective, since they are generated by algorithms, not people. Therefore, this perception might be dangerous in terms of changing peoples' minds and attitudes towards current events.

Present concerns related to journalistic ethics, may be decreased by assigning interlocutors during the generation of algorithmic journalism. For example, the person who inputs the data, chooses the algorithm, and checks the story may be considered the interlocutor who is responsible for any violations of the ethical rules. After the news is generated and shared, for each part of the content of the news, individuals should apply a verification framework to examine the news, its source, and its context [2].

In sum, if algorithmic journalism technology is used under the control of journalists checking results and following ethical rules, we can envision that this technology can be useful for creating a huge amount of content at higher speed with fewer costs. However, further comprehensive studies are needed to explore the impacts of algorithmic journalism on journalists, newsroom managers, and other newsroom workers who control the creation of algorithmic news, and the individuals who read the algorithmic news.

5 Conclusion

For years, journalists have been benefitting from many technological tools, such as bullhorns, tablets, computers, cameras, and phones, to gather, produce, present, and distribute information to the public [14]. Current presentational formats, including hooks, listicles, gifs, podcasts, virtual and augmented reality, conversational interfaces, and data visualization, are utilized to produce more attractive news [14]. Today, more advanced technologies, such as NLG (natural language generation) based on AI, are used to generate news content. Moreover, because of ongoing improvements in AI – thanks to Big Data, advanced

algorithms, and more powerful computers – new programs may emerge that can enhance algorithmic journalism [13].

Algorithmic journalism may be spreading out more because of its speed and its power to deal with huge amounts of data, which provide deeper, more specific, and immediately available information, which can benefit society, as long as ethical rules are followed and necessary measures are taken, such as checking input, output, and models regularly to eliminate ethical concerns, such as violations of transparency, verification, privacy, bias, etc. Moreover, automation may call for human skills, such as judgment, curiosity, and skepticism, so that we can continue to access succinct, comprehensive, and accurate news.

REFERENCES

[1] Lucas Vieira de Araujo. 2018. Algorithms, artificial intelligence and NLG in the production of Brazilian journalism. *Set International Journal of Broadcast Engineering* (2018), 9.

[2] Hazel Baker. 2019. Making a "deepfake": How creating our own synthetic video helped us learn to spot one - Press Gazette. Retrieved November 27, 2019 from https://www.pressgazette.co.uk/making-a-deepfake-how-creating-our-own-synthetic-video-helped-us-learn-to-spot-one/

[3] Matt Carlson. 2015. The Robotic Reporter: Automated journalism and the redefinition of labor, compositional forms, and journalistic authority. *Digital Journalism* 3, 3 (May 2015), 416–431. DOI:https://doi.org/10.1080/21670811.2014.976412

[4] David Caswell and Konstantin Dörr. 2018. Automated Journalism 2.0: Event-driven narratives: From simple descriptions to real stories. *Journalism Practice* 12, 4 (April 2018), 477–496. DOI:https://doi.org/10.1080/17512786.2017.1320773

[5] Arjen van Dalen. 2012. The algorithms behind the headlines: How machine-written news redefines the core skills of human journalists. *Journalism Practice* 6, 5–6 (October 2012), 648–658. DOI:https://doi.org/10.1080/17512786.2012.667268

[6] Nicholas Diakopoulos. 2016. BuzzFeed's pro tennis investigation displays ethical dilemmas of data journalism - Columbia Journalism Review. Retrieved November 27, 2019 from https://www.cjr.org/tow_center/transparency_algorithms_buzzfeed.php

[7] Konstantin Nicholas Dörr. 2016. Mapping the field of Algorithmic Journalism. *Digital Journalism* 4, 6 (August 2016), 700–722. DOI:https://doi.org/10.1080/21670811.2015.1096748

[8] Matthew Jenkin. 2016. Written out of the story: the robots capable of making the news. *The Guardian*. Retrieved November 7, 2019 from https://www.theguardian.com/small-business-network/2016/jul/22/written-out-of-story-robots-capable-making-the-news

[9] Seth C. Lewis, Andrea L. Guzman, and Thomas R. Schmidt. 2019. Automation, Journalism, and Human–Machine Communication: Rethinking Roles and Relationships of Humans and Machines in News. *Digital Journalism* 7, 4 (April 2019), 409–427. DOI:https://doi.org/10.1080/21670811.2019.1577147

[10] Ekaterina Pashevich. 2018. Automation of news production in Norway: Augmenting newsroom with artificial intelligence.

[11] Alex Primo and Gabriela Zago. 2015. Who And What Do Journalism?: An actor-network perspective. *Digital Journalism* 3, 1 (January 2015), 38–52. DOI:https://doi.org/10.1080/21670811.2014.927987

[12] Emily Reynolds. 2015. Wordsmith's "robot journalist" has been unleashed. (2015), 3.

[13] Neil Thurman, Konstantin Dörr, and Jessica Kunert. 2017. When Reporters Get Hands-on with Robo-Writing: Professionals consider automated journalism's capabilities and consequences. *Digital Journalism* 5, 10 (November 2017), 1240–1259. DOI:https://doi.org/10.1080/21670811.2017.1289819

[14] Barbie Zelizer. 2019. Why Journalism Is About More Than Digital Technology. *Digital Journalism* 7, 3 (March 2019), 343–350. DOI:https://doi.org/10.1080/21670811.2019.1571932