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, Retesco, 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.

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
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].

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., an algorithmic journalist) using NLP, developed by Automated Insights, one of the leading commercial providers of NLG technologies, became

![Figure 1: I-T-O model of algorithmic journalism [7].](image-url)
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](image)

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 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
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