Scientific and Social Research, 2025, Volume 7, Issue 7
https://ojs.bbwpublisher.com/index.php/SSR
Online ISSN: 2981-9946
Print ISSN: 2661-4332

The Impact of Big Data on News Selection, Production, and Distribution

Jilin Li*

King's College London, London WC2R2LS, United Kingdom

**Author to whom correspondence should be addressed.*

Abstract: Big data is profoundly changing the news industry's production logic and dissemination mode. Nowadays, the media increasingly rely on data mining, machine learning, and user behavior analysis to accurately predict audience concerns and optimize content presentation. The rise of data-driven journalism has enabled media outlets to adjust their topic selection strategies based on real-time data, increase the appeal of their stories, and improve production efficiency and scale through automation. However, this model is not without its drawbacks. Traffic orientation may lead media to cater to algorithmic recommendations overly, undermine in-depth reporting, and even contribute to the proliferation of emotional and vulgar content. In contrast, the role of big data in news dissemination is more intuitive. Precision communication allows media to customize news distribution based on audience reading habits and interests, improve user engagement and information reach efficiency, and enhance the social impact of news. This paper will discuss how big data reshapes news selection and content production and analyze its challenges and pitfalls.

Keyword: Journalism; Content production; Precision communication; Algorithmic recommendation

Online publication: August 12, 2025

1. Introduction

The convergence of big data and generative AI is reshaping journalism from the cult of efficiency to cognitive reconstruction, and technological rationality is reconfiguring the nature of news production. The core challenge lies in protecting the "human core" of news in the wave of intelligence, maintaining public openness, avoiding inaccurate communication, and adhering to professional values. The future news ecology should move towards value symbiosis rather than human-machine substitution—algorithms expanding cognitive breadth, humans guarding thinking depth, machines enhancing distribution efficiency, journalists defending verification accuracy, systems mining data correlation, and editors shaping value consensus. Only by establishing a dynamic and balanced symbiosis mechanism can journalism complete its paradigm shift and become the cognitive infrastructure of the digital era.

2. Reconstruction of the whole process of news production based on big data analysis

While big data technology revolutionizes topic selection and content production in the news industry, its dual impact—enhancing efficiency while introducing risks—demands critical scrutiny. With data mining, automated writing, and data visualization, news organizations optimize their topic selection strategies and improve the efficiency of content production to better meet audience needs and market competition. As a new model that integrates data analysis and news reporting, data journalism is changing the traditional way of news selection and content production [1]. It emphasizes optimizing topic selection strategies through data mining, automated writing, and visualization, accurately grasps audience needs, and provides more targeted news reports. Through large-scale data analysis, data journalism can accurately capture the hotspots of public concern, uncover hidden trends, and improve the science and diversity of topic selection [1]. According to data Journalism Editor E, although data journalism enables trend analysis and complex topic summarization, its reliance on third-party data raises concerns about journalistic independence and coverage breadth. For example, this approach has been demonstrated in the statistical and word-frequency analysis of the topic "involution" [6]. Big data broadens the perspective of news selection and enables news organizations to accurately identify hotspots, capture potential social trends, and enhance the foresight and innovation of selected topics. In terms of news content production, automation and innovative technologies accelerate the news production process and increase the efficiency and innovation of content generation. Big data drives news production automation, enabling news editors to quickly capture the latest information and analyze and innovate content based on the data [9, 12]. Data mining techniques help media to process massive amounts of data, quickly identify news leads, and automatically generate portions of content [9, 12]. Meanwhile, Artificial Intelligence (AI) optimizes the process, enabling news outlets to leverage data analytics, cross-validation, and content summarization to increase the efficiency of information screening and reduce manual processing costs [12]. For example, Tencent's Story Forest system intelligently clusters massive news content into events and organizes related events into a news story tree to help users understand news logic more clearly [5]. Bloomberg's NSTM system uses semantic clustering and summarization methods to compress news information into easy-to-understand points, helping users quickly grasp key news events [2]. Big data technology has improved the efficiency and precision of news production, enabling news organizations to respond quickly to hot spots, optimize topic selection, and enhance the depth and innovation of content. With the development of technology, news production will become more intelligent and precise, providing the public with more transparent and more intuitive reports and exerting a more excellent social value.

3. Empirical evidence of technological vulnerability in the smart news production chain

While big data improves the efficiency of news production, it also raises quality issues. Despite relying on massive datasets for analysis, data bias, algorithmic errors, or source limitations may affect content accuracy. Meanwhile, the selection of topics for data journalism is limited by data availability, and many important topics are overlooked due to the lack of open, structured data support [1]. Such data dependency not only narrows the breadth of coverage but fundamentally contradicts the core promise of data journalism—to reduce reliance on official sources. Ironically, media outlets remain trapped in passive dependence on third-party data. Although data journalism is seen as a means to reduce reliance on official sources and thus enhance journalistic independence, in reality, third-party data provided by governments, research institutions, and corporations are still the primary source of information for data journalism, with a low percentage of media-autonomous investigations [3]. Although

data journalism aims to reduce reliance on official sources, governments, research institutions, and corporations are still the primary sources of data, with a low percentage of media-autonomous surveys [3]. Although self-collected data can reduce external reliance, it is not easy to become a routine practice due to its high cost and time-consuming nature [1]. For example, to report on the "dating corner", a news team arranged for seven journalists to go out six times to collect data and go through a long process of data entry [1]. The high cost of data journalism limits the range of topics that can be covered, favoring quantitative issues, while important issues that lack data support are often overlooked [1]. This may lead to "data bias", where news is centered around available data only, ignoring social realities that are difficult to quantify. The media should expand data sources, such as civic data, user-generated data, or interdisciplinary collaborations, to remedy this problem. In addition, data journalism faces challenges of authenticity and objectivity. The diversity of data sources does not equal accuracy, and the collection process may be subject to bias, computational errors, and even data manipulation [4]. Specific political and corporate organizations may selectively disclose data, affecting journalistic integrity [3]. Data visualization may mislead the public, and an oversimplified presentation of information may blur context and even affect perception [10]. The limitations of data analysis methods should not be overlooked, as faulty data modeling or statistical methods may cause news to cater to a particular narrative rather than a true reflection of the facts [13]. The high cost of data integration affects news quality, and issues such as incompatible formats and inconsistent variable naming make data cleaning complex and time-consuming [4]. In addition, the separation of data journalism teams from traditional editorial teams and the lack of close collaboration between analysts, journalists, and designers affect news innovation and depth [3]. Data are not entirely objective but are influenced by how they are collected, the logic of processing, and the means of presentation. If news organizations rely only on authoritative data and lack independent investigation and multidimensional analysis, data news may become a tool for information manipulation rather than a guarantee of authenticity. Therefore, the media should strengthen data transparency, verification mechanisms, and ethical norms to ensure the credibility and social value of reporting.

4. Thinking and reflective learning on AI and the media industry

Big data-driven news push and search optimization improve information access efficiency and promote the personalized development of news dissemination. Accurate matching of user needs makes content distribution intelligent and optimizes the user reading experience. The application of big data algorithms covers personalized recommendations, information filtering, user profile construction, and news consumption analysis. Based on user behavioral data, news push is more accurate, effectively matches interests, and improves distribution efficiency [7–8]. Google News employs a recommendation system based on user clicking behavior, predicting user interests through a Bayesian framework, and combining collaborative filtering to optimize recommendations and improve user engagement [8]. Information filtering combines users' historical behaviors and news trends to adjust the pushed content to make personalized recommendations more accurate while optimizing content retrieval from search engines and social media to reduce active search time [7]. The optimized recommendation system increases the click-through rate of Google News recommendations by 30.9% and the frequency of user visits by 14.1% [8]. Similarly, "Today's Headlines" can quickly update the user interest model to better match the news content with reading preferences [7]. These technologies improve the efficiency of news dissemination and enhance the user experience, making news access more intelligent and precise [7–8]. Personalized algorithms may lead to homogenization of information reception, exposing users to similar content for an extended period and forming

an "information cocoon" [7, 11]. Algorithms build a profile based on user browsing data and make recommendations based on predicted interests, narrowing the scope of information. In addition, algorithms tend to push content with a high click-through rate, which makes users fall into "traffic traps" and are attracted to low-quality or harmful information, solidifying content consumption patterns [7, 11]. Under the influence of information cocoon and traffic orientation, quality content is marginalized, news producers cater to traffic at the expense of content quality, and in-depth information is challenging to enter the public eye [7, 11]. Meanwhile, algorithmic agenda setting has limitations, and it is challenging to recognize users' potential needs, failing the effectively disseminate policies and international news, and the passive acceptance of recommendations by the audience [7, 11]. In addition, algorithmic pushing reduces users' willingness to search and think deeply and actively. Information overload triggers thinking fatigue and weakens the ability to filter and make judgments, making users more inclined to accept the recommended content in its entirety instead of screening the authenticity and depth of information. Fragmented information consumption patterns weaken critical thinking and may lead to frequent opinion reversals [7, 11]. Although big data recommendation systems have improved the convenience of information access, their mechanisms have exacerbated the problems of information cocooning and traffic orientation, trapping users in a closed loop and weakening their ability to think independently. Algorithmic agenda-setting restricts diverse information presentation and marginalizes critical content, potentially eroding public discourse diversity—a paradox in an era claiming to prioritize "personalization" [11].

5. The future technological landscape of intelligent news production and the construction of media ecology

With the development of big data, ChatGPT, and other artificial intelligence technologies, the news industry is experiencing a profound change in the logic of production and dissemination. AI technology enhances both the operational efficiency and analytical depth of news production, while refining language models to generate more engaging and context-rich narratives. In the future, news organizations may widely adopt AI-assisted topic selection, automatic news generation, and optimized user profiles in combination with deep learning to achieve accurate push while alleviating the problem of an information cocoon. Data transparency and news credibility will become key issues, and blockchain may be used for news traceability and information verification to reduce the risk of fake news. Regulators and media will also pay more attention to the transparency and ethics of algorithms and optimize the recommendation mechanism so that news dissemination can strike a balance between precision and diversity. To survive the digital intelligence era, traditional media must strategically integrate AI and big data without compromising journalistic ethics, thereby reclaiming competitiveness through authoritative and socially responsible reporting.

Disclosure statement

The author declares no conflict of interest.

Reference

[1] Bai H, Zhang T, 2022, Understanding Data Journalism as Knowledge: An Investigation Based on Journalism Epistemology. China News Review, 3(3): 74–82. https://cics.fudan.edu.cn/2e/49/c40201a667209/page.htm?utm_

source=chatgpt.com

[2] Bambrick J, Xu M, Almonte A, et al., 2020, NSTM: Real-Time Query-Driven News Overview Composition at Bloomberg. https://arxiv.org/pdf/2006.01117

[3] Chen Z, Wang P, 2016, Characteristics and Issues of Data Journalism in Chinese News Websites. Chinese Journal of Journalism & Communication, 38(6): 45–58.

[4] Kasica S, Berret C, Munzner T, 2020, Table Scraps: An Actionable Framework for Multi-Table Data Wrangling From An Artifact Study of Computational Journalism. IEEE Transactions on Visualization and Computer Graphics. https://arxiv.org/pdf/2009.02373

[5] Liu B, Niu D, Lai K, et al., 2018, Growing Story Forest Online from Massive Breaking News. https://arxiv.org/pdf/1803.00189

[6] Lan T, 2022, A New Way to Construct the Value of Data News: From Topic Selection Expansion to Visual Presentation — A Case Study of "Surging Beauty Mathematics Course. China News Review, 3(3): 74–82. https://doi.org/10.35534/cnr.0303008

[7] Li S, 2021, The Impact of Big Data Algorithms on Media Audiences. Advances in Social Sciences, 10(9): 2705–2709. https://doi.org/10.12677/ASS.2021.109371

[8] Liu J, Dolan P, Pedersen ER, 2010, Personalized News Recommendation Based on Click Behavior. Proceedings of the 15th International Conference on Intelligent User Interfaces, 31–40.

[9] Meng F, 2023, Research on Innovative Strategies for News Editing in the Era of Big Data. Electronics, Communication and Computer Science, 5(2): 88–90.

[10] Morini F, 2025, Different yet Complementary: A Systematic Literature Review on Data Journalism in Visualization Research and Journalism Studies. Journalism, 26(2): 425–444. https://doi.org/10.1177/14648849241237897

[11] Qiu D, Geng H, 2023, An Analysis of the Impact of Big Data Technology Application on Human Thinking Quality. Journal of Dialectics of Nature, 45(9): 81–88. https://doi.org/10.15994/j.1000-0763.2023.09.011

[12] Veglis A, Saridou T, Panagiotidis K, et al., 2022, Applications of Big Data in Media Organizations. Social Sciences, 11(9): 414. https://doi.org/10.3390/socsci11090414

[13] Zhou Q, Liang J, Xu X, 2021, A Comparative Study on Data Journalism Research and Practice at Home and Abroad—Based on Bibliometric and Content Analysis. Journal of Literature and Data Science, 3(2): 47–64. https://wxysjxb.ajcass.com/Admin/UploadFile/Issue/201902190001/2021/7//20210713020210WU_FILE_0.pdf?utm_source=chatgpt.com