5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16440

AI in the newsroom: A data quality assessment framework for employing machine learning in journalistic workflows

Laurence Dierickx[1], Carl-Gustav Lindén[1], Andreas L Opdahl[1], Sohail Ahmed Khan[1], Diana Carolina Guerrero Rojas[1]

[1]Department of Information Science and Media Studies, University of Bergen, Norway

Abstract

AI-driven journalism refers to various methods and tools for gathering, verifying, producing, and distributing news information. Their potential is to extend human capabilities and create new forms of augmented journalism. Although scholars agreed on the necessity to embed journalistic values in these systems to make AI-driven systems accountable, less attention was paid to data quality, while the results' accuracy and efficiency depend on high-quality data. However, defining data quality remains complex as it is a multidimensional and highly domain-dependent concept. Assessing data quality in AI-driven journalism requires a broader and interdisciplinary approach, considering journalists as end-users. It means meeting the challenges of data quality in machine learning and the ethical challenges of using machine learning in journalism. These considerations ground a conceptual data quality assessment framework that aims to support the collection and pre-processing stages in machine learning. It aims to strengthen data literacy in journalism by emphasizing limitations and possible biases related to data and making a bridge between journalism studies and scientific disciplines that should be viewed through the lenses of their complementarity.

Keywords: *data quality assessment, journalism, ethics, machine learning, artificial intelligence*

This research was funded by EU CEF grant number 2394203.

Editorial Universitat Politècnica de València

1. Introduction

AI-driven journalism refers to various methods and tools for news gathering, verification, production, and distribution (Thurman et al., 2019). They aim to support professional practices to help speed up time-consuming tasks, publish automated content, identify trends, or provide insights into large numeric or textual datasets. Hence, their potential is to extend human capabilities and augment journalism practices (Lindén, 2018). Although AI-driven systems are often considered opaque and not bias-free (Guidotti et al., 2019), they depend on high-quality data to avoid inaccurate analytics and unreliable decisions (Gupta et al., 2021). Explaining how data is collected, organised, cleaned, annotated, and processed participates in establishing a relationship of trust between the journalist as the end-user and the tool. It implies understanding the challenges of data quality that appear upstream and downstream of a machine learning process (Gudivada et al., 2017).

The “garbage in, garbage out” principle also applies in journalism, whereas quality information requires quality data to ensure the accuracy and reliability of the news (e.g., Anderson, 2018; Diakopoulos, 2020; Dierickx, 2017; Dörr & Hollbuchner, 2017; Lowrey et al., 2019). However, less attention was paid to this critical aspect. The conceptual framework presented in this paper intends to fill this gap, considering that assessing data quality is context and use dependent (Tayi & Ballou, 1998; Boydens & Van Hooland, 2011).

2. Theoretical backdrops

Data quality encompasses several complementary dimensions referring to a set of attributes in which dimensions – such as accuracy, completeness, and consistency – were refined over time. However, research agreed that data quality refers to data that adapts to the uses of data consumers, especially in terms of accuracy, relevance, and understandability (Wang & Strong, 1996). The emergence of big data brought new challenges, such as believability, verifiability, and the reputation of the data (Batini et al., 2015). The level of trustability of the data was also underlined, as various data sources challenge their interoperability and the contexts where data are used (e.g., Cai & Zhu, 2015; Liu et al., 2016; Saha & Srivastava, 2014). Big data quality issues are also related to incomplete, inaccurate, inconsistent, or ambiguous structured and unstructured data (Eberendu, 2016).

Approaching data quality in machine learning includes all these considerations but also encompasses several particularities insofar as the quality of the results is influenced by the data provided as input to the system (Gudivada et al., 2017; Gupta et al., 2021). Also, research emphasised that models trained on incomplete or biased datasets can produce discriminatory outputs and interfere with the accuracy of the tasks (Miceli et al., 2022; Shin et al., 2022).

Data quality issues are likely to appear since the data acquisition stage: data availability does not equal data quality (Elouataoui et al., 2022), especially when working with open data, user-generated data, or data coming from multiple sources (Hair & Sarstedt, 2021). Data pre-processing involves addressing classical data quality issues, such as missing data, duplicates, strongly correlated variables, abnormal or inconsistent values, normalization, and standardization (Polyzotis et al., 2018; Foidl & Felderer, 2019; Elouataoui et al., 2022).

Training datasets, which refer to the process of adapting the model to the data, are needed to evaluate the suitability of the data for machine learning tasks –in terms of efficiency, accuracy and complexity (Gupta et al., 2021). In this context, the validation process aims to ensure that data does not contain errors that can propagate into the model. These errors will likely be introduced during the collection, aggregation or annotation stage (Polyzotis et al., 2018; Gupta et al., 2021). However, it is practically impossible to achieve it exhaustively, even though evaluating the risk of poor data quality is possible. At the same time, there is a lack of discussion on methods to define the level of validation in each step of a machine learning process (Foidl & Felderer, 2019). Furthermore, corpus annotations for supervised tasks are problematic because they are inherently error-prone, either if they rely on automation or crowdsourcing (Gupta et al., 2021).

Because the relationship between users and AI systems lies on trust (Rai, 2020), data quality should follow three fundamental principles: prevention, detection, and correction to ensure the trustworthiness and reliability of machine learning applications (Ehrlinger et al., 2019). A good understanding of the data provides correct analyses and reliable decisions (Gupta et al., 2021). It should also reflect the knowledge of the domain experts. Furthermore, selecting or creating a dataset for an AI-driven system involves human decisions beyond technical aspects, thus requiring empirical considerations (Miceli et al., 2021).

Ethical journalism practices join these concerns. They refer to the rules, routines and institutionalised procedures to produce knowledge (Ekström, 2002). Although ethical journalism is a question of practice, providing truthful information is not dissociable from the news's credibility (or believability) (van Dalen, 2019). Hence, ethical principles of journalism can be summarised according to the main principle of respecting the truth with accuracy and objectivity (Ward, 2018). The development of data-driven practices focused specifically on the data source's reliability, accuracy, the right to extract and use the data, and the right to privacy (Craig et al., 2017). At the same time, transparency has become a motto, viewed as an instrument to increase credibility and trust toward audiences (Koliska, 2022).

In AI-driven journalism, the ethical challenges of transparency concern the data, the algorithms at work, and the outcomes (Dörr & Hollbuchner, 2017). Nonetheless, transparency is not always easy to implement in journalism, where practitioners often lack data and algorithm literacy to grasp how algorithms work (Porlezza & Eberwein, 2022), no more than it is easy to implement in deep learning models where even their creators need to learn how they operate because of the multiplicity of their parameters (Burkart & Huber, 2021). While a recognised need exists to blend AI-driven systems with journalistic values to fit professional practices (Broussard et al., 2019; Gutierrez Lopez, 2022), it should start with the data. If they are biased or contain errors, the system will likely reproduce these biases and errors (Hansen et al., 2019). Considering that accuracy and reliability are two prerequisites of ethical journalism practices, trusting the system is also about trusting the data it relies on.

3. Building the conceptual assessment framework

Data quality assessment is critical and gives rise to operations that aim to improve the overall data quality by identifying erroneous data elements and understanding their impact on the processes at work (Cichy & Rass, 2019). From an end-user perspective, assessing data quality indicators refer to their fitting to human needs or user requirements through the aggregation of different information on data quality (Cappiello et al., 2004). The assessment framework we have developed in the context of AI-driven journalism is a part of this data quality assessment tradition. It is based on the learnings from the scientific literature (e.g., Batini et al., 2009; Cichy & Rass, 2019; Fox et al., 1994; Pipino et al., 2002; Shanks, 1999) and on the core ethical principles in journalism acknowledged by professionals.

The ethical principle of telling the truth relates to respecting facts. It refers to the syntactic and semantics levels and the dimensions of the data's accuracy, consistency, correctness, and understandability. It requires the application domain knowledge to deal, for instance, with incorrect values or duplicates. Because objectivity is a disputed concept in journalism due to its intrinsic subjective nature, we privileged the one of fairness related to the elements that guarantee to report honestly, avoiding bias or unbalanced information. It concerns the context of producing, validating, disseminating, and using the data for a journalistic purpose. Hence, it is connected to the pragmatic level and relates to the dimensions of timeliness, completeness, accessibility, objectivity, relevance, and usability. Transparency refers to the trustability of information, but it is not the only constituent of trust. The broader concept of trust can be thus understood through the social semiotic level. It encompasses the dimensions of credibility, reliability, and verifiability.

The assessment framework encompasses formal and empirical indicators, inducing that the overall assessment includes a human perspective. Its application can be objective or subjective (Pipino et al., 2002) to detect data quality issues and challenges likely to appear upstream of the processes, either generally or more granularly. It can be applied to the data collection and pre-processing stages from which the training, the test, and the validation datasets are derived for developing machine learning systems in a journalistic context.

Table 1. Data quality assessment framework.

Ethical	Semiotic	Dimension	Verification
Truth	**Syntactic**	Accuracy	- Level of interoperability, standardisation - Measure of erroneous data (ratio accurate values/total values) - Uniqueness (duplicate entries and redundancies) - Encoding problems and information overload
	Semantic	Consistency	- Well-defined data structure (percentage of data with consistent format and values) - Homogeneity vs heterogeneity (format, structure, values) when data come from multiple sources - Unambiguous and explicit labelling
		Correctness	- Identifying abnormal values - Identifying the causes of NULL values - Evaluation of the spelling coherence - Data documented/compliant with metadata
		Understandability	- The extent to which data are comprehensible (feedback from the end-user)
Fairness	**Pragmatic**	Timeliness	- Currentness (percentage of updated data)
		Completeness	- Appropriate amount of data (ratio missing values/total values, ratio NULL values/total values)
		Accessibility	- Right to use the data (terms of use) - Level of retrievability of the data
		Objectivity	- Unbiased data (size and representativity)
			- Identification of human bias (annotation incl.)
		Relevance	- The extent to which the data are relevant for the purpose (feedback from the end-user)
			- Newsworthiness (journalistic added values and expected impact, feedback from the end-user)
			- Data scarcity (measurement of the fraction of data containing relevant information)
		Usability	- Fitness for use (to assess globally through the formal and empirical indicators of the frameworks, = making sense of AI in a journalistic context)
			- How automation structures and presents the data
Trust	**Social**	Reliability	- Authenticity (source) - Authority (source, annotators) - Reputation (source, annotators)
		Credibility	- Degree of the believability of the data source
			- Degree of the believability of the data
			- Degree of the believability of the annotation process and of the annotators
		Verifiability	- Verification of the source and the data
			- Verification of the annotation process

This framework was applied to a sample of datasets used for automated fact-checking. While the syntactic level did not have particular issues, on the semantic level, a cross-domain approach and a strong language dependency challenged the understandability and correctness of the datasets. The pragmatic level appeared problematic due to NULL values, no attached licence, and no mention of the last update. The dimension of completeness was more difficult to assess because of the content's domain and language dependency. A lack of harmonization in the classification was also detected, from "true" to "false", "half true", "contradiction", or "unrelated". On the social level, datasets collected from Wikipedia raised questions about their reliability and credibility, due to the participative nature of this platform.

4. Conclusion

Acknowledging that the relationship between journalists and AI-driven systems is built on trust, the data that feed these systems must also be trusted. However, the definition of “good” data in journalism remains challenging due to the multidimensionality of the concept of quality. This concept is intrinsically related to the expertise of a given application domain and to the understanding of how data are collected, validated, and disseminated. It should also be considered through its relevance to be used in a journalistic context and the overall purpose of the AI-driven system. Also, approaching data quality through normative lenses consists of a practical solution to address the recognized need for embedding journalistic values and ethical principles in AI-driven systems. Hence, the conceptual assessment framework presented in this communication was designed as an adaptive and flexible tool that can be used in various forms that AI-driven journalism tools can take. It shows that data quality issues are far from trivial, as the quality of the data at every stage of the process will directly influence machine learning outcomes. Nevertheless, the main limitation of this framework is that it is only applicable for common machine learning tasks because the provenience and nature of the vast amounts of data used in the most complex systems remain mostly uncertain.

References

Anderson, C. W. (2018). *Apostles of certainty: Data journalism and the politics of doubt*. Oxford University Press.

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, *41*(3), 1–52.

Batini, C., Rula, A., Scannapieco, M., & Viscusi, G. (2015). From data quality to big data quality. *Journal of database management*, *26*(1), 60–82.

Boydens, I., & van Hooland, S. (2011). Hermeneutics applied to the quality of empirical databases. *The Journal of Documentation; Devoted to the Recording, Organization and Dissemination of Specialized Knowledge*, *67*(2), 279–289.

Burkart, N., & Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, *70*, 245-317.

Broussard, M., Diakopoulos, N., Guzman, A. L., Abebe, R., Dupagne, M., & Chuan, C.-H. (2019). Artificial intelligence and journalism. *Journalism & Mass Communication Quarterly*, *96*(3), 673–695.

Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the Big Data era. *Data Science Journal*, *14*(0), 2.

Cappiello, C., Francalanci, C., & Pernici, B. (2004). Data quality assessment from the user's perspective. In *Proceedings of the 2004 international workshop on Information quality in information systems* (pp. 68–73).

Cichy, C., & Rass, S. (2019). An overview of data quality frameworks. *IEEE Access: Practical Innovations, Open Solutions*, *7*, 24634–24648.

Craig, D., Ketterer, S., & Yousuf, M. (2017). To post or not to post: Online discussion of gun permit mapping and the development of ethical standards in data journalism. *Journalism & Mass Communication Quarterly*, *94*(1), 168–188.

Diakopoulos, N. (2019). *Automating the News: How Algorithms Are Rewriting the Media*. Harvard University Press.

Dierickx, L. (2017). News bot for the newsroom: How building data quality indicators can support journalistic projects relying on real-time open data. In *Global Investigative Journalism Conference 2017, Academic Track.*

Dörr, K. N., & Hollnbuchner, K. (2017). Ethical challenges of algorithmic journalism. *Digital Journalism*, *5*(4), 404–419.

Eberendu, A. C.. (2016). Unstructured Data: an overview of the data of big data. *International Journal of Computer Trends and Technology*, *38*(1), 46–50.

Ehrlinger, L., Haunschmid, V., Palazzini, D., & Lettner, C. (2019). A DaQL to monitor data quality in machine learning applications. In *Lecture Notes in Computer Science* (pp. 227–237). Springer.

Ekström, M. (2002). Epistemologies of TV journalism: A theoretical framework. *Journalism*, *3*(3), 259-282.

Elouataoui, W., Alaoui, I. E., & Gahi, Y. (2022). Data quality in the era of big data: A global review. In *Big Data Intelligence for Smart Applications* (pp. 1–25). Springer.

Foidl, H., & Felderer, M. (2019). Risk-based data validation in machine learning-based software systems. *Proceedings of the 3rd ACM SIGSOFT International Workshop on Machine Learning Techniques for Software Quality Evaluation*.

Fox, C., Levitin, A., & Redman, T. (1994). The notion of data and its quality dimensions. *Information Processing & Management*, *30*(1), 9–19.

Gudivada, V., Apon, A., & Ding, J. (2017). Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. *International Journal on Advances in Software*, *10*(1), 1–20.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), 1–42.

Gupta, N., Mujumdar, S., Patel, H., Masuda, S., Panwar, N., Bandyopadhyay, S., Mehta, S., Guttula, S., Afzal, S., Sharma Mittal, R., & Munigala, V. (2021). Data quality for machine learning tasks. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*.

Gutierrez Lopez, M., Porlezza, C., Cooper, G., Makri, S., MacFarlane, A., & Missaoui, S. (2022). A question of design: Strategies for embedding AI-driven tools into journalistic work routines. *Digital Journalism*, 1–20.

Hair, J. F., Jr, & Sarstedt, M. (2021). Data, measurement, and causal inferences in machine learning: opportunities and challenges for marketing. *The Journal of Marketing Theory and Practice*, *29*(1), 65–77.

Hansen, M., Roca-Sales, M., Keegan, J. M., & King, G. (2017). *Artificial intelligence: Practice and implications for journalism*. Columbia University.

Karlsen, J., & Stavelin, E. (2014). Computational journalism in Norwegian newsrooms. *Journalism Practice*, *8*(1), 34–48.

Keller, S., Korkmaz, G., Orr, M., Schroeder, A., & Shipp, S. (2017). The evolution of data quality: Understanding the transdisciplinary origins of data quality concepts and approaches. *Annual Review of Statistics and Its Application*, *4*(1), 85–108.

Koliska, M. (2022). Trust and journalistic transparency online. *Journalism Studies*, *23*(12), 1488–1509.

Lindén, C.-G. (2018). What Makes a Reporter Human? A Research Agenda for Augmented Journalism. *Questions de communication*, (37), 337–351.

Liu, J., Li, J., Li, W., & Wu, J. (2016). Rethinking big data: A review on the data quality and usage issues. *ISPRS Journal of Photogrammetry and Remote Sensing: Official Publication of the International Society for Photogrammetry and Remote Sensing (ISPRS)*, *115*, 134–142.

Lowrey, W., Broussard, R., & Sherrill, L. A. (2019). Data journalism and black-boxed data sets. *Newspaper Research Journal*, *40*(1), 69–82.

Miceli, M., Posada, J., & Yang, T. (2021). Studying up machine learning data: Why talk about bias when we mean power? *Proceedings of the ACM on Human-Computer Interaction*, *6*, 1–14.

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, *45*(4), 211–218.

Polyzotis, N., Roy, S., Whang, S. E., & Zinkevich, M. (2018). Data lifecycle challenges in production machine learning: A survey. *SIGMOD Record*, *47*(2), 17–28.

Porlezza, Colin, & Eberwein, T. (2022). Uncharted territory: Datafication as a challenge for journalism ethics. In *Media and Change Management* (pp. 343–361). Springer.

Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, *48*(1), 137–141.

Saha, B., & Srivastava, D. (2014). Data quality: The other face of big data. *2014 IEEE 30th International Conference on Data Engineering*.

Shanks, G. (1999). Semiotic approach to understanding representation in information systems. In *Proceedings of the Information Systems Foundations Workshop, Ontology, Semiotics and Practice*.

Shin, D., Hameleers, M., Park, Y. J., Kim, J. N., Trielli, D., & Diakopoulos, N. (2022). Countering algorithmic bias and disinformation and effectively harnessing the power of AI in media. *Journalism & Mass Communication Quarterly*, *99*(4), 887–907.

Tayi, G. K., & Ballou, D. P. (1998). Examining data quality. *Communications of the ACM*, *41*(2), 54–57.

Thurman, N., Lewis, S. C., & Kunert, J. (2019). Algorithms, automation, and news. *Digital Journalism*, *7*(8), 980–992.

van Dalen, A. (2019). Journalism, Trust, and Credibility. In *The Handbook of Journalism Studies* (pp. 356–371). Routledge.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, *12*(4), 5-33.

Ward, S. J. A. (2018). Reconstructing journalism ethics: Disrupt, invent, collaborate. *Media & Jornalismo*, *18*(32), 9–17.