Khizanah al-Hikmah : Jurnal Ilmu Perpustakaan, Informasi, & Kearsipan
p-ISSN: 2354-9629
e-ISSN: 2549-1334
Volume 13 Issue 1, June, 2025

# Artificial Intelligence in Recordkeeping: A Systematic Review of Machine Learning Applications for Automated Records Classification

**Sigit Sumarsono[1] & Muhamad Prabu Wibowo[2]**

[1,2]Universitas Indonesia, Indonesia
Correspondence email: sigitsumarsono@gmail.com

Notes

*Submitted: 20-12-2024*
*Revised: 19-03-2025*
*Accepted: 24-03-2025*

**How to cite:** Sumarsono, S., & Wibowo, M. P. (2025). Artificial Intelligence in Recordkeeping: A Systematic Review of Machine Learning Applications for Automated Records Classification. Khizanah Al-Hikmah : Jurnal Ilmu Perpustakaan, Informasi, Dan Kearsipan, 13(1). https://doi.org/10.24252/v13i1a12

DOI: 10.24252/v13i1a12



**ABSTRACT**

This study presents a systematic literature review (SLR) of scholarly research published between 2014 and 2023, with the aim of identifying prevailing trends, methodological approaches, and contextual factors surrounding the use of Machine Learning (ML) models for records classification within Records Management and Archival Science. Employing the PRISMA framework, the review analyzes a curated selection of studies to assess the scope and maturity of ML applications in this domain. The findings revealed that while ML has been increasingly explored for tasks such as classification and appraisal, its application remains geographically skewed, with the majority of studies originating from Global North countries. The models employed range from probabilistic and regression-based algorithms to decision tree classifiers, reflecting diverse but largely traditional methodological approaches. The adoption of more sophisticated techniques, including deep learning and large language models, was still limited. The study underscores a critical research gap concerning the implementation of advanced ML models, particularly in the context of Global South institutions, where such technologies could significantly enhance recordkeeping efficiency and scalability. This review highlights the need for further empirical studies that develop and evaluate cutting-edge ML models in diverse archival contexts, promoting more inclusive and globally representative innovation in archival automation.

**Keywords:** Records Management; Artificial Intelligence; Machine Learning; Archival Science

## 1. INTRODUCTION

The incorporation of Artificial Intelligence (AI) into automatic document classification is one of the major areas of interest in information science, particularly in records management and archival science. Automatic document classification has become a significant technique for managing digital information, including document retrieval, filtering, and summarization (Kowsari et al., 2019). From a records management perspective, records classification is not only essential for document retrieval purposes, but also plays crucial role in the development of classification system by providing unlabeled records with context, content, and structure, which is vital for ensuring authenticity and reliability of records (Mokhtar & Yusof, 2015).

One of the most promising applications of AI in records management and archival science is the automation of recordkeeping workflows previously performed manually, such as description, appraisal, and access (Colavizza et al., 2022). In practical terms, a significant portion of records metadata required to support recordkeeping quality can be derived from context and content of records themselves (Makhlouf Shabou, 2015), automatic classification can be employed to address gaps in essential metadata, including records classification of the document (Büttner, 2019). Such automation can enable organizations to develop more objective recordkeeping workflow by utilizing computers as an assistant in appraisal decision-making (Oladejo & Hadžidedić, 2021).

Machine learning has emerged as a key AI-based approach for automatic classification. While various definitions of machine learning exist, this study conceptualizes it within the context of textual records classification as the development of computational models utilizing Natural Language Processing (NLP) techniques. These models are trained on human-labeled documents to systematically predict the most probable labels for unlabeled documents, thereby enhancing the efficiency and accuracy of classification processes (Alsmadi & Gan, 2019; Colavizza et al., 2022; Meng et al., 2020). This process leverages the computational power of computers to rapidly analyze textual content or contextual metadata through syntax and semantic-level analysis, matching the results with potential records classification schemas. In practice, this automation is categorized as supervised machine learning, wherein the model is trained using a dataset of documents that have been accurately labeled with the appropriate records classification subjects by human experts. The labeled dataset is then utilized for training, testing, and evaluating the model's accuracy (Hutchinson, 2020).

Machine learning for textual classification offers numerous benefits to the recordkeeping process. Unlike alternative methods such as rule-based expert systems, as observed in some earlier studies, machine learning produces classification results that are more stable, accurate, and meaningful (Li et al., 2022). Additionally, machine learning models are easier to develop and less costly to maintain compared to rule-based classification systems (Payne & Baron, 2017). However, these benefits come with certain limitations. One major concern is the human involvement required during the process, particularly the subjectivity and potential inaccuracies of training data labeled by human experts, who often rely on subjective judgment (Hjørland, 2023). Another challenge is the lack of standardized subject taxonomies in records management, unlike the bibliographic control systems in libraries that utilize standardized classification schemes (e.g., Dewey Decimal Classification or Library of Congress Subject Headings). These standardized schemes facilitate the development of common classification models applicable across organizations. In records management, however, records and archives are not arranged as individual documents but as *fonds*, i.e., groups of related documents organized based on the principles of provenance and original order. This requires organizations to maintain functional records classification schemes tailored to their specific recordkeeping needs, rather than adhering to uniform standards across organizations (Mokhtar & Yusof, 2015).

Over the past two decades, there has seen a rapid surge of literature exploring the use of machine learning in textual classification (Palanivinayagam et al., 2023). Consequently, several literature reviews have been conducted on this topic. Some reviews have focused on describing machine-learning-based auto-classification as a state-of-the-art approach in electronic records management system research (Colavizza et al., 2022; Oladejo & Hadžidedić, 2021). Another review provided a brief overview of text classification algorithms, including machine learning, across various implementation domains beyond records management and archival science (Kowsari et al., 2019). In the field of archival science, one literature review examined the use of machine learning to support archival processing, though it was not specifically focused on textual classification (Hutchinson, 2020). Meanwhile, in other disciplines such as health informatics and computer science, numerous systematic literature reviews have explored the application of machine learning in text classification. Some of these reviews evaluate machine learning generally (Alsmadi & Gan, 2019; Palanivinayagam et al., 2023; Riduan et al., 2021) while the others focus on one specific machine learning algorithm (Minaee et al., 2022; Pintas et al., 2021). To the best of the author's knowledge, the usage of machine learning in textual records classification in records management and archival science has garnered minimal attention from the scholarly community and has not been systematically analyzed.

This study aims to examine current research trends on the utilization of machine learning models for records classification tasks within the fields of Records Management and Archival Science through a systematic literature review. The formulation of research questions and the critical appraisal were guided by the Theory, Context, Characteristics, and Method (TCCM) framework of systematics review (Paul et al., 2021, 2024).

This study seeks two research questions: first, what research contexts form the basis for classifying textual records using machine learning? And second, which machine learning models have been applied in such classification efforts? By engaging with these questions, the study aims to fill notable gaps in the current body of literature and to advance the scholarly conversation on the integration of machine learning in records management. A systematic examination of recent research trends not only maps the trajectory of current developments but also outlines promising avenues for future inquiry. Moreover, by identifying the models most commonly employed in the classification of textual records, the study offers practical insights that may inform best practices and stimulate the application of more sophisticated techniques. These findings hold particular relevance for policymakers, archival practitioners, and researchers invested in enhancing the effectiveness and efficiency of textual records management in the era of artificial intelligence.

## 2. METHODS

To address the research questions mentioned in the previous part, this study employs a framework-based systematic literature review (SLR) technique following the procedures outlined by Petticrew and Robert (2006). These procedures include identifying research questions, conducting a literature search, screening the search results, critically appraising and synthesizing the literature, and reporting the findings. The findings are reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. The advantage of this approach lies in its systematic nature, which enhances the quality of the review and its replicability (Page et al., 2021).

A significant challenge in identifying relevant literature on artificial intelligence is the tendency of researchers to avoid using general terms such as "*artificial intelligence,*" instead opting for more specific and technical terminology to prevent unnecessary marginalization of their findings (Toosi et al., 2021). To mitigate this issue, the literature search in this study employed not only specific terms like "*machine learning,*" "*deep learning,*" "*supervised learning,*" and "*natural language processing,*" but also combined these terms with discipline-specific keywords such as "*text classification,*" "*automatic classification,*" "*records management,*" and "*archival science.*"

Several criteria were applied in the literature identification developed from population, intervention, comparison, outcomes, and context (PICOC) model (Sampaio, 2015). The inclusion criteria for this study encompassed research publications on automatic textual records or archive classification using machine learning, published between 2014 and 2023, and indexed in Scopus, ACM Digital Library, IEEE Xplore, and Taylor & Francis. The use of multiple scientific databases for sourcing electronic literature is justified by the fact that most high-quality, relevant studies are indexed within these platforms. Articles that were duplicates or lacked full-text availability were excluded from the study.

The literature search was conducted on October 28, 2024, focusing on the occurrence of one or more specified terms in the titles, abstracts, and keywords. The search terms used in this study were as follows:

("CLASSIFICATION (OF INFORMATION)" OR "AUTO-CLASSIFICATION" OR "AUTOMATIC CLASSIFICATION" OR “TEXT CLASSIFICATION”) AND ("ARCHIVAL SCIENCE" OR "RECORDS MANAGEMENT") AND ("MACHINE LEARNING" OR "NATURAL LANGUAGE PROCESSING" OR "ARTIFICIAL INTELLIGENCE" OR "DEEP LEARNING" OR "SUPERVISED LEARNING").

Following the literature search, the retrieved literature was qualitatively screened based on its relevance and ability to address the research questions within records management and archival science. The number of studies identified at each stage of the search process is summarized in Figure 1 below.

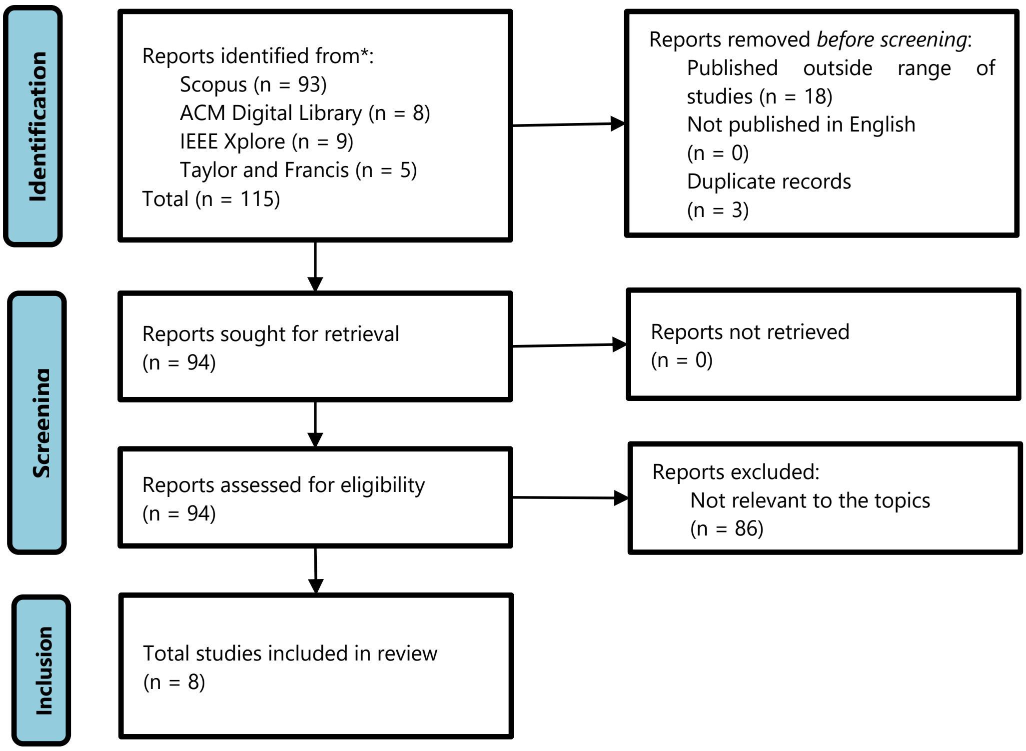


**Figure 1.** PRISMA Flowchart summarizing literature search

This study has certain limitations, including the small number of articles reviewed and the lack of detailed case-specific analyses. As modern records increasingly encompass formats beyond texts such as audio and visual materials. Future research should address these gaps by broadening the scope of systematic reviews and employing more rigorous comparative evaluation methods, such as meta-analysis.

## 3. RESULTS AND DISCUSSION

The literature search identified 115 studies from Scopus, ACM Digital Library, IEEE Xplore, and Taylor & Francis. Of these, 21 studies were excluded due to publication dates falling outside the research scope or duplication resulting from the use of multiple scientific databases. Further qualitative screening excluded additional 83 studies that did not meet the inclusion criteria. After this process, 8 studies were selected, consisting of 5 journal articles, 2 conference papers, and 1 book chapter. A general overview of the selected studies is summarized in Table 1 below.

**Table 1.** General descriptions of selected literature

| ID | Author (Year) | Title | Source | Type |
|---|---|---|---|---|
| A1 | Vellino & Alberts (2016) | Assisting the appraisal of e-mail records with automatic classification | Records Management Journal | Article |
| A2 | Rolan et al. (2019) | More human than human? Artificial intelligence in the archive | Archives and Manuscripts | Article |
| A3 | Bardelli et al. (2020) | Automatic Electronic Invoice Classification Using Machine Learning Models | Machine Learning and Knowledge Extraction | Article |
| A4 | Goodrum et al. (2020) | Automatic classification of scanned electronic health record documents | International Journal of Medical Informatics | Article |
| A5 | Franks (2022) | Text Classification for Records Management | Journal on Computing and Cultural Heritage | Article |
| C1 | Lei et al. (2017) | Automatically Classify Chinese Judgment Documents Utilizing Machine Learning Algorithms | Database Systems for Advanced Applications | Book Chapter |
| P1 | Payne (2023) | An Intelligent Class - The Sequel: The Development of a Novel Context Capturing Method for The Functional Auto Classification of Records | Proceedings - 2023 IEEE International Conference on Big Data, BigData 2023 | Conference paper |
| P2 | Triantafyllou (2023) | Thematic categorization on university records | Proceedings of the 2023 IEEE 11th International Conference on Systems and Control | Conference paper |

### Context Behind Implementations

The context of studies related to machine learning can be analyzed based on the location of the study, the objectives of model implementation, the type of dataset used, the classification targets, and evaluation metrics used. Regarding the study locations, the selected literature indicates that research has been conducted across six different countries in the Americas, Asia, Australasia, and Europe. Two countries, Australia and the United States, accounted for more than one study, with two studies each. Other studies originated from Canada, Italy, Greece, and China.

The scope of research varied based on the type of organization. Several studies were conducted within public organizations, including companies (A1 and A3), hospitals (A4), and universities (P2). Two studies from Australia focused on government agencies: Study A5 examined current records actively managed by the organization, while Study A2 analyzed archival records held by archival institutions. In contrast, Studies C1 and P1 were categorized as desk research, with the development of models aimed purely at research purposes rather than practical application within specific organizations.

Based on their objectives, five out of eight models were developed for general purposes of classification and knowledge organization. For instance, Study A3 focused on classifying invoice documents into classification codes to facilitate financial reporting, while Study A4 aimed to assign scanned documents to the appropriate records classification. Three studies pursued more specific objectives related to archival appraisal. For example, Study A1 sought to classify records based on their utility value without further classification. Meanwhile, two studies conducted in Australia, A2 and A5, had more practical aims, classifying documents into categories that already included metadata for retention schedule attached in classification categories.

The datasets used in the studies varied significantly. While Study P1 utilized a publicly available dataset—the Enron Email Dataset—most studies, particularly those at the proof-of-concept stage, relied on datasets derived from organizational records of varying scopes. For instance, the dataset in Study A1 originated from two employees whose records had been acquired. The datasets in Studies A4 and A5 were obtained from an EDMS (Electronic Document Management System) used by a single organization, while the dataset in Study A3 was sourced from two different companies.

More complex yet similar settings were observed in Studies C1 and P2. The datasets in these studies were metadata captured from information systems used to publish publicly accessible documents, specifically the Chinese Judgement Online System in Study C1 and the Greek "Diavgeia" Portal in Study P2. The most comprehensive dataset was reported in Study A2, which sampled "The full corpus comprised 30 GB of data, in 7,561 folders, containing 42,653 files [with] no disposal rules applied to the files" (Rolan et al., 2019).

In terms of classification targets, four studies (A2, A3, P1, and P2) utilized publicly established classification standards. For instance, Study A3 used hierarchical structure of the Corporate Chart of Accounts in Italy as the basis for document labeling, study P1 on their function based on the Operational Records Classification System (ORCS) employed by the Government of Canadian province of British Columbia, while Study A2 adopted The General retention and disposal authority: administrative records (GA28) classification standard set by the State Records of New South Wales, Australia. Three other studies (A4, A5, and C1) used classification standards specific to internal organizational practices. Meanwhile, study A1 employed a simpler classification model by categorizing records solely based on their business value (i.e. important or not).

### Method and Characteristics

All studies share a similar workflow in developing and testing automatic classification models. This workflow typically includes the following steps: (i) document acquisition, (ii) text preprocessing, (iii) feature selection, and (iv) training and testing the machine learning classifier. The training and testing phases are often preceded by splitting the dataset into training data, used for model training, and testing data, used to evaluate the model's performance. After testing, some studies, such as A5 performed parameter tuning by resampling selected variables to further enhance model accuracy.

The data preprocessing was tailored to the characteristics of the datasets in each study. Three studies utilizing records in Extensible Markup Language (XML) format—two on emails (A1 and P1) and one on invoice documents (A3)—required minimal preprocessing, as XML is inherently machine-readable. In contrast, studies involving PDF-based records (A2, A5, C1, and P2) had to extract text into a more accessible format. Study A4, which used image PDFs, employed Optical Character Recognition (OCR) to convert images into text before further processing.

Once the data was in a machine-readable format, common preprocessing methods included tokenization to divide corpora into meaningful units and vectorization to transform text into numerical representations for computational analysis. Additional steps varied by study: for example, Studies A2 and C1 performed stop word removal, while Studies A4 and A5 eliminated categories with insufficient records to mitigate chronic class imbalance issues.

While most studies employed a vectorized corpus as the sole feature for classification, several studies incorporated feature extraction to select only the most relevant features from the text corpus. For instance, Study A1 applied feature extraction by identifying email attributes (e.g., main, forward, reply) as classification features. Similarly, Studies C1 and P2 extracted key terms based on internally developed term dictionaries. The rationale for this approach is that not all content within a document or judgment is relevant for achieving classification objectives. Extracting only the necessary features reduces noise, thereby enhancing the model's performance (Lei et al., 2017). Other studies modified the vectorized corpus during feature extraction. For example, Study A2 limited the number of terms included in the vectorized corpus, while study A3 removed terms with an occurrence rate of less than 0.1%.

The training and testing phases reflected the unique characteristics of each study. During the training stage, most studies employed a train-test split ratio of 80/20. However, studies without significant class imbalance issues tended to use a lower ratio of training data, such as Study A2 (75/25) and Study C1 (70/30). The choice of classification algorithms varied across the studies. Classification models can be broadly categorized into five major groups:

1) *Probabilistic Models*, based on probability theory, such as Naive Bayes (NB).
2) *Regression Models*, relying on linear regression to classify documents based on features, such as Support Vector Machines (SVM) and Logistic Regression (LR).
3) *Decision Trees utilizing*, hierarchical classification representations, such as Decision Trees (DT) and Random Forest (RF).
4) *Deep Learning Models*, based on artificial neural networks, such as Convolutional Neural Networks (CNN), Long Short-Term Memory networks (LSTM), and Multilayer Perceptrons (MLP). And
5) *Language Models*. Advanced neural networks trained to solve linguistic problems by converting text into vector representations, often referred to as transformer models. Examples include Bidirectional Encoder Representations from Transformers (BERT) and Robustly Optimized BERT Pre-training Approach (RoBERTa) (Alsmadi & Gan, 2019; Franks, 2022).

Regarding the number of models tested, two studies (A1 and P1) applied only one classification model, while five studies (A2, A3, A4, C1, and P2) compared between two and four classification models. A more comprehensive approach was observed in Study A5, which compared seven different models spanning three classification types.

Another commonality lies in the evaluation metrics, with seven studies utilizing the F1 Score as a primary measure. The F1 Score, an information retrieval metric defined as the harmonic mean of precision and accuracy, is particularly robust for document classification tasks involving unbalanced categories (Franks, 2022). Although variations in study settings prevent direct comparisons of F1 scores, the reported scores remain useful for assessing the accuracy of the models employed, particularly when multiple models are evaluated within a single study. Reported evaluation metrics of the selected studies are summarized in Table 2 below.

The results of the literature review reveal several interesting trends worth discussing. There is a noticeable shift among researchers away from feature extraction processes. Of the six studies conducted since 2020, only one employed specific feature extraction techniques beyond term weighting. This shift can be attributed to advancements in commercial computing power, enabling the processing of increasingly complex vectorizations within realistic timeframes. Additionally, the emergence of language models equipped with transformer mechanisms has further reduced the need for feature extraction by transforming corpora into more efficient vectors without requiring extensive feature tuning. Nevertheless, the appropriate application of feature extraction remains important for creating more efficient models. For instance, the study P2 by Triantafyllou (2023) demonstrated that employing relevant feature extraction techniques can yield high-accuracy models, even when supported by relatively simple classification algorithms. This approach indirectly reduces computational load during data processing, thereby increasing the volume of records that can be managed effectively.

The evaluation results highlight that several simple models can serve as benchmarks for future research settings. Algorithms such as SVM and RF have demonstrated the ability to produce models with high levels of accuracy. For example, in Study C1, the SVM model outperformed more complex models, such as those based on deep learning. In contrast, in Study A5, the performance of SVM was surpassed only by language models with significantly higher complexity. Meanwhile, Study P2 showed that the RF model outperformed the SVM model. These findings corroborate the review by Alsmadi and Gan (2019), which observed that decision tree and SVM models occasionally achieve very high accuracy, especially compared to other simple algorithms. Therefore, both algorithms can be effectively utilized as benchmarks to assess the baseline performance of archival classification models. The use of more complex models should only be considered when they demonstrate a significant improvement in performance over simpler models.

**Table 2.** Reported evaluation metrics of selected literature

| ID | Preprocessing | Maximum F1 Score of Classification Models | | | | |
|---|---|---|---|---|---|---|
| | | Probabilistic | Regression | Decision Tree | Deep Learning | Large Language Model |
| A1 | TF | | **0.910(SVM)** | | | |
| A2 | TF | 0.648(NB) | | | **0.835(MLP)** | |
| A3 | TF | | | 0.918(RF)<br>0,928(Ada) | **0.972(MLP)** | |
| | Word2Vec | | | 0.946(RF)<br>0.949(Ada) | 0.938(MLP) | |
| A4 | TF | | | 0,883(RF) | | |
| | TF-IDF | | 0,877(LR) | | | |
| | LLM | | | | | **0,913(BERT)** |
| A5 | TF-IDF | | 0,732(SVM) | | 0.522(C-LSTM) | |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | 0,561(LSTM)<br>0,718(CNN) | |
| | LLM | | | | | **0,785(BERT)**<br>0,770(RoBERTa)<br>0,771(XLNET) |
| C1 | TF-IDF | 0,720(NB) | **0,860(SVM)** | 0,820(RF)<br>0,795(DT) | | |
| P1 | TF-IDF<br>LDA | | | 0,587(RF)<br>**0,646(RF)** | | |
| P2* | Prob-IDF | | 0,974(SVM) | **0,977(RF)**<br>0,967(DT) | | |

Note: * reported as P*F-Score

One significant gap to address is the potential application of deep learning and language models in automatic classification. Although trends in textual classification research indicate that text representation/vectorization methods based on LLM have gained popularity since 2019 (Palanivinayagam et al., 2023), only two of the eight studies reviewed have utilized LLM. On the other hand, CNN remain the most widely applied model for text classification problems (Riduan et al., 2021). However, most archival studies continue to focus on algorithms classified as "shallow learning," with only one study exploring CNN for archival classification tasks. Results from previous studies employing deep learning and language models such as Goodrum et al. (2020) and Franks (2022) of which CNN-based model or language model could easily beat the evaluation result of more simple methods, demonstrate the promising potential of these advanced methods, suggesting a need for further exploration in the field of study.

The geographic distribution and context of the studies indicate that automatic classification of textual records is a globally relevant research opportunity. The diverse research settings highlight opportunities for further investigation into various types of organizational records, whether in public or private sectors. However, the disproportionate representation between the Global North and Global South—with seven studies originating from the Global North and only one study from the Global South (China)—underscores the urgency of conducting research on records from the Global South, including those from Indonesia.

The similarity in workflows for automatic classification is not only evident in the eight studies discussed but also reflects broader trends across other disciplines (e.g. Pintas et al., 2021; Riduan et al., 2021). Literature reviews from different fields reveal that, in principle, the application of machine learning follows comparable workflows regardless of the domain. To bridge interdisciplinary gaps, Records Management and Archival Science should leverage these similarities by adopting research trends from other disciplines and testing their applicability within the context of records management and archival science.

## 4. CONCLUSION

This study systematically reviews research trends in the application of machine learning for classifying textual records, using the PRISMA framework. From an initial set of 115 publications, eight studies were identified as relevant. The findings reveal that such applications are primarily concentrated in the Global North and are commonly used to replicate expert-driven categorizations or support archival appraisal processes. While text classification has been widely explored in other fields, records management and archival science have predominantly employed probabilistic, regression-based, and decision tree models. The integration of advanced approaches such as deep learning and large language models remains limited. Given their demonstrated effectiveness in related domains, future research should consider their potential to enhance records classification in archival contexts.

**ACKNOWLEDGEMENT**

-

**AUTHORS' CONTRIBUTIONS**

**Sigit Sumarsono:** Writing oiginal draft preparation. Ideas; formulation or evolution of overarching research goals and aims. **Muhamad Prabu Wibowo:** Ideas; formulation or evolution of overarching research goals and aims.

**CONFLICT OF INTERESTS**

We state that there are no known conflicts of interest linked with this publication, and that there has been no significant financial assistance for this work that could have influenced its outcome.

## REFERENCES

Alsmadi, I., & Gan, K. H. (2019). Review of short-text classification. *International Journal of Web Information Systems*, *15*(2), 155–182. https://doi.org/10.1108/IJWIS-12-2017-0083

Bardelli, C., Rondinelli, A., Vecchio, R., & Figini, S. (2020). Automatic Electronic Invoice Classification Using Machine Learning Models. *Machine Learning and Knowledge Extraction*, *2*(4), 617–629. Scopus. https://doi.org/10.3390/make2040033

Büttner, G. (2019). Auto-classification in an international organization: Report from a feasibility study. *Comma*, *2017*(2), 15–26. https://doi.org/10.3828/comma.2017.2.2

Colavizza, G., Blanke, T., Jeurgens, C., & Noordegraaf, J. (2022). Archives and AI: An Overview of Current Debates and Future Perspectives. *Journal on Computing and Cultural Heritage*, *15*(1), 1–15. https://doi.org/10.1145/3479010

Franks, J. (2022). Text Classification for Records Management. *Journal on Computing and Cultural Heritage*, *15*(3). Scopus. https://doi.org/10.1145/3485846

Goodrum, H., Roberts, K., & Bernstam, E. V. (2020). Automatic classification of scanned electronic health record documents. *International Journal of Medical Informatics*, *144*. Scopus. https://doi.org/10.1016/j.ijmedinf.2020.104302

Hjørland, B. (2023). Description: Its meaning, epistemology, and use with emphasis on information science. *Journal of the Association for Information Science and Technology*, *74*(13), 1532–1549. https://doi.org/10.1002/asi.24834

Hutchinson, T. (2020). Natural language processing and machine learning as practical toolsets for archival processing. *Records Management Journal*, *30*(2), 155–174. https://doi.org/10.1108/RMJ-09-2019-0055

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text Classification Algorithms: A Survey. *Information*, *10*(4), 150. https://doi.org/10.3390/info10040150

Lei, M., Ge, J., Li, Z., Li, C., Zhou, Y., Zhou, X., & Luo, B. (2017). Automatically Classify Chinese Judgment Documents Utilizing Machine Learning Algorithms. In Z. Bao, G. Trajcevski, L. Chang, & W. Hua (Eds.), *Database Systems for Advanced Applications* (Vol. 10179, pp. 3–17). Springer International Publishing. https://doi.org/10.1007/978-3-319-55705-2_1

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., & He, L. (2022). A Survey on Text Classification: From Traditional to Deep Learning. *ACM Transactions on Intelligent Systems and Technology*, *13*(2), 1–41. https://doi.org/10.1145/3495162

Makhlouf Shabou, B. (2015). Digital diplomatics and measurement of electronic public data qualities: What lessons should be learned? *Records Management Journal*, *25*(1), 56–77. https://doi.org/10.1108/RMJ-01-2015-0006

Meng, Y., Zhang, Y., Huang, J., Xiong, C., Ji, H., Zhang, C., & Han, J. (2020). *Text Classification Using Label Names Only: A Language Model Self-Training Approach* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2010.07245

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2022). Deep Learning-based Text Classification: A Comprehensive Review. *ACM Computing Surveys*, *54*(3), 1–40. https://doi.org/10.1145/3439726

Mokhtar, U. A., & Yusof, Z. M. (2015). Classification: The understudied concept. *International Journal of Information Management*, *35*(2), 176–182. https://doi.org/10.1016/j.ijinfomgt.2014.12.002

Oladejo, B., & Hadžidedić, S. (2021). Electronic records management – a state of the art review. *Records Management Journal*, *31*(1), 74–88. https://doi.org/10.1108/RMJ-09-2019-0059

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

Palanivinayagam, A., El-Bayeh, C. Z., & Damaševičius, R. (2023). Twenty Years of Machine-Learning-Based Text Classification: A Systematic Review. *Algorithms*, *16*(5), 236. https://doi.org/10.3390/a16050236

Paul, J., Khatri, P., & Kaur Duggal, H. (2024). Frameworks for developing impactful systematic literature reviews and theory building: What, Why and How? *Journal of Decision Systems*, *33*(4), 537–550. https://doi.org/10.1080/12460125.2023.2197700

Paul, J., Lim, W. M., O'Cass, A., Hao, A. W., & Bresciani, S. (2021). Scientific procedures and rationales for systematic literature reviews (SPAR-4-SLR). *International Journal of Consumer Studies*, *45*(4). https://doi.org/10.1111/ijcs.12695

Payne, N. (2023). An Intelligent Class – The Sequel: The Development Of A Novel Context Capturing Method For The Functional Auto Classification Of Records. *2023 IEEE International Conference on Big Data (BigData)*, 2071–2082. https://doi.org/10.1109/BigData59044.2023.10386255

Payne, N., & Baron, J. R. (2017). Auto-categorization methods for digital archives. *2017 IEEE International Conference on Big Data (Big Data)*, 2288–2298. https://doi.org/10.1109/BigData.2017.8258182

Petticrew, M., & Roberts, H. (2006). *Systematic Reviews in the Social Sciences: A Practical Guide* (1st ed.). Wiley. https://doi.org/10.1002/9780470754887

Pintas, J. T., Fernandes, L. A. F., & Garcia, A. C. B. (2021). Feature selection methods for text classification: A systematic literature review. *Artificial Intelligence Review*, *54*(8), 6149–6200. https://doi.org/10.1007/s10462-021-09970-6

Riduan, G. M., Soesanti, I., & Adji, T. B. (2021). A Systematic Literature Review of Text Classification: Datasets and Methods. *2021 IEEE 5th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 71–77. https://doi.org/10.1109/ICITISEE53823.2021.9655788

Rolan, G., Humphries, G., Jeffrey, L., Samaras, E., Antsoupova, T., & Stuart, K. (2019). More human than human? Artificial intelligence in the archive. *Archives and Manuscripts*, *47*(2), 179–203. https://doi.org/10.1080/01576895.2018.1502088

Sampaio, A. (2015). Improving Systematic Mapping Reviews. *ACM SIGSOFT Software Engineering Notes*, *40*(6), 1–8. https://doi.org/10.1145/2830719.2830732

Toosi, A., Bottino, A. G., Saboury, B., Siegel, E., & Rahmim, A. (2021). A Brief History of AI: How to Prevent Another Winter (A Critical Review). *PET Clinics*, *16*(4), 449–469. https://doi.org/10.1016/j.cpet.2021.07.001

Triantafyllou, I. (2023). Thematic Categorization on University Records. *2023 IEEE 11th International Conference on Systems and Control (ICSC)*, 384–389. https://doi.org/10.1109/ICSC58660.2023.10449857

Vellino, A., & Alberts, I. (2016). Assisting the appraisal of e-mail records with automatic classification. *Records Management Journal*, *26*(3), 293–313. Scopus. https://doi.org/10.1108/RMJ-02-2016-0006