Case Study: Using AI in the retention and disposition of records at the New South Wales State Archives

Kaila Fewster[1]

Educational applications: This case is particularly useful for exploring how AI tools can be integrated into records management workflows and demonstrates how tool experimentation and adaptation is necessary when working with AI on archival projects. In this sense, it is also useful for highlighting the importance of basic data science skills and algorithmic thinking for archivists and records managers. Furthermore, this case study also illustrates the importance of evaluating different off-the-shelf AI tools and algorithms for their effectiveness and meaningful integration into existing archival and records management workflows.

Educational topics: AI for retention and disposition, records as data for AI, evaluating and adjusting AI/ML models for archives, collaborative management of AI projects, digital literacy for AI in archives[2].

About: This case study is part of a series of learning materials developed by InterPARES Trust AI[3] researchers and educators to train archival professionals and students to effectively leverage artificial intelligence in their archival work. The final draft was completed on October 21st, 2023.

This case study describes a 2017 internal pilot project investigating the application of machine learning to records management. The New South Wales State Archives (NSW) Digital Archives team initially began in early 2017 by publishing a research paper exploring what machine learning is capable of and how to apply it to records management. They found that although machine learning has the potential to improve the classification and disposition of digital records, there has been minimal adoption of the technology. As a result, Humphries and his colleagues came to undertake a series of internal and external projects to explore further and demonstrate machine learning in records management. The pilot project's goal was to take off-the-shelf machine learning software and apply it to the problem of classifying a corpus of

---

[1] InterPARES Trust AI Graduate Academic Assistant, University of British Columbia.

[2] Educational applications map to a Body of Knowledge proposed by InterPARES researchers for AI/ML for the archival professionals. https://docs.google.com/document/d/1UsjkkkGeSJrgCDJGASCAy5q0uo_ZkQpzi_Ch8XUcqYw/edit?usp=sharing

[3] This case study is an outcome of InterPARES Trust AI, an international research partnership led by Drs. Luciana Duranti and Muhammad Abdul-Mageed, University of British Columbia, and funded by the Social Sciences and Humanities Research Council of Canada (SSHRC). https://interparestrustai.org/

[4] Case Study: Using AI in the retention and disposition of records at the New South Wales State Archives © 2024 by Fewster, Kaila is licensed under CC BY-NC-SA 4.0. To view a copy of this license, visit https://creativecommons.org/licenses/by-nc-sa/4.0/

unstructured data against a retention and disposition authority. The corpus had previously been processed against a disposition authority manually, and thus, the project aimed to test the accuracy of the machine-learning algorithms in automatically classifying the corpus to match these disposal classes.

Due to limited resources and the nature of the project, the team investigated low-cost, off-the-shelf technological solutions and ultimately chose scikit-learn, a free and open-source machine learning library that can be used in Python. The corpus of records selected had been transferred from a central government department to the State Archives in 2016 and boasted a complete corporate file structure of over 42,500 files. The corpus was reduced to only include easily text extractable, required state records, leaving 8,784 files in the final sample. The sample then went through data cleaning and a text vectorization process following the Bag-of-Words model, which disregards the location of a word in the document and instead focuses on the frequency of each word. This process creates a document-term matrix that illustrates a term's frequency among documents and helps the algorithm's classification process.

Once the data was prepared, the project team chose to compare two widely used machine learning classification algorithms, the Multinomial Naïve Bayes (MLB) model and the Multi-Layer Perceptron (MLP) network. Both a cleaned and uncleaned version of the corpus was tested through both algorithms. 75% of the data was used to train the algorithms, and the remaining 25% was used to test the accuracy of their classifications. The findings show that the MLP algorithm was the most successful, with an average accuracy rate of 80.4%, compared to the MLB's 66.6% accuracy. It was also determined that while the corpus had been manually sentenced against the disposition authority to the folder level, the MLP model could sentence directly at the document level much quicker. Given the size and complexity of this project, the researchers determined the algorithm's accuracy was an acceptable result and more broadly demonstrates the potential for machine learning in records management.

Potential Discussion Questions:

1. What are the potential benefits and drawbacks of using off-the-shelf machine learning and AI tools like scikit-learn in archives and records management projects?
2. In what ways will the integration of AI tools into archives and records management redefine the role of human information professionals in these spaces, and what kinds of new competencies might be required as datafication becomes more prevalent?
3. What role does resource limitation play in the selecting AI tools and technologies for projects like this one? How might resource constraints influence the outcomes and scalability of AI initiatives for archives and records management long-term?

References

Humphries, G. (2018, March 20). Case Study – Internal Pilot – Machine Learning and Records Management. *Future Proof - Protecting Our Digital Future*. https://futureproof.records.nsw.gov.au/case-study-internal-pilot-machine-learning-and-records-management/

Humphries, G. (2018). *Machine-learning-pilot* [Python]. NSW State Archives. https://github.com/srnsw/machine-learning-pilot (Original work published 2018)

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