Bohyun Kim

5 Investing in Artificial Intelligence: Considerations for Libraries and Archives

Abstract: This chapter highlights the relevance of artificial intelligence (AI) and machine learning (ML) to library and archive work through various pilot projects conducted in libraries and archives. It describes several projects that leveraged machine learning (ML) technologies, including computer vision, speech-to-text, named entity recognition, natural language processing NLP), and an AI chatbot powered by a large language model (LLM). This chapter presents examples of libraries' and archives' adopting and applying AI and ML to provide engaging and efficient information services and to generate richer metadata at scale, which allows users to discover, identify, access, navigate, cluster, analyze, and use materials more easily and effectively. Also discussed are where and how libraries and archives should invest in AI and ML to reap the most benefit and the implications of such investments in costs and prospects.

Keywords: Artificial intelligence – Economic aspects; Machine learning; Artificial intelligence – Library applications

Introduction

With the emergence of ChatGPT, artificial intelligence (AI) and machine learning (ML) have become popular topics in mass media. Although chatbots are not new and ChatGPT is not the only generative AI tool available, ChatGPT's much-improved performance as a chatbot surprised many and quickly captured people's imagination. AI and ML have been topics of interest among librarians and archivists well before ChatGPT. But the adoption and implementation of AI and ML in libraries and archives has been slow. Given that most libraries and archives do not have existing expertise in AI and ML, this slow adoption may be partly attributed to uncertainty about which ML technologies are relevant to the work of libraries and archives and unfamiliarity with a range of compelling use cases that apply ML technologies to library and archive-related tasks. To address such uncertainty and unfamiliarity, this chapter discusses several pilot projects that make use of specific ML technologies, highlights the relevance of AI and ML to library and archive work, and explores ways in which libraries and archives can invest in AI and ML and the costs and prospects associated with those potential investments.

https://doi.org/10.1515/9783111336435-006

Machine Learning Technologies Relevant to Libraries

In adopting AI and ML, the goal of libraries and archives is to enhance and improve their services and operations. If newly developed AI and ML capabilities can add value in that manner, that is a good reason for a library or an archive to investigate and invest in AI and ML. However, many libraries experience difficulties in determining where and how they should begin. Fortunately, previous ML pilots conducted in the library and archive setting can serve as good starting points for those libraries that are looking to build capacity and invest in AI and ML. The rest of this section summarizes some pilot projects and highlights the ML technologies that are likely to be most relevant to libraries and archives.

Computer Vision

With the Project AIDA (Image Analysis for Archival Discovery), the University of Nebraska-Lincoln and the Library of Congress in the United States explored the use of ML in the library and archive context, in particular computer vision. Computer vision is an ML technology that enables computers to identify objects and people in digital images and videos and to derive meaningful information from them. Computer vision applications process a large volume of digital images and video data to perform tasks such as object identification, facial recognition, and image classification (IBM n.d.a). The AIDA project included several ML experiments that applied computer vision technology to archival images, such as developing an ML algorithm that identifies the region of graphical content in historical newspaper page images, determining whether an item is handwritten, printed, or a mix of both and classifying manuscript collection images as digitized from the original format or a microform reproduction (Lorang et al. 2020, 3–18). This pilot showed the potential of ML technology, specifically computer vision, to expedite image processing in archival materials.

Another ML pilot was run by the Frick Art Reference Library in New York in collaboration with researchers at Stanford University, Cornell University, and the University of Toronto. Also using computer vision, the research team developed an image classifier algorithm for the library's digitized Photoarchive. Based on visual elements, the image classifier algorithm applied a local hierarchical classification system and automatically assigned metadata to the digitized images of portrait paintings (Prokop et al. 2021, 15). The image classifier showed that an AI tool can help the library staff assign metadata to digitized images more quickly, thereby improving the ability of users to access and retrieve those images.

Natural Language Processing

The AMPPD (Audiovisual Metadata Platform Pilot Development) project was run jointly by Indiana University (IU) Libraries, the University of Texas at Austin School of Information, the New York Public Library (NYPL), and AVP, an information innovation company. The goal of the project was to improve metadata workflows using ML technology. Using both proprietary and open-source ML tools, the AMPPD project team developed automated metadata generation mechanisms (MGMs) and integrated them with the human metadata generation process (Dunn et al. 2021, 5).

Various ML techniques such as optical character recognition (OCR), speech recognition also known as speech-to-text, and named entity recognition were used in the project to automate metadata work. OCR software converts an image of text in a printed format into a machine-readable text format for searching and further manipulation (AWS 2024c). Speech-to-text software enables computers to process human speech into a written format, thereby producing the text transcription of human speech in audio files (IBM, n.d.c). Named entity recognition (NER) is a component of a well-known ML technology, natural language processing (NLP). NLP techniques enable computers to recognize, interpret, generate, and respond to human language in voice and text data and are used for tasks such as translation, summarization sentiment analysis (IBM n.d.b). NER takes in a string of text and identifies and classifies entities that belong to specific categories, such as names of individuals, locations, organizations, and expressions of times.

The AMPPD project built a fully functioning platform with twenty-four MGMs that could be used to analyze, describe, and document the audiovisual materials in IU and NYPL's collections (Dunn et al. 2021, 13). The ML technologies mentioned above were used for detecting silence, speech, and music; speech-to-text and speaker diarization; named entity recognition for people, geographic locations, and topics; video and structured data OCR; and music genre detection. The experiment demonstrated how a library's traditional metadata creation workflow for audio and moving image materials can be augmented and improved by ML technologies.

The University of Notre Dame Libraries applied NLP to enhance metadata for its Catholic Pamphlets collection, which consisted of over 5,500 PDFs. MARC summary fields had been assigned to half of the collection, but most of the summaries were a few words at most, lacking sufficient metadata. The team used NLP automated summarization techniques to create more robust summaries by combining the summaries with Library of Congress subject headings (Flannery 2020, 23–24). The work undertaken provides another example of how libraries can make use of ML to enhance existing metadata for their collections.

Chatbot

Another ML pilot project using NLP is a custom chatbot developed by Zayed University Library in the United Arab Emirates. The library chatbot named Aisha was created with Python programming language and OpenAI's ChatGPT application programming interface (API) to provide reference and support services when students and faculty need help outside the library's operating hours (Lappalainen and Narayanan 2023, 38, 51). A chatbot is an ML computer program that simulates human conversation in text or voice (OCI 2024). A sophisticated AI chatbot such as ChatGPT utilizes NLP, one of the AI/ML technologies, to generate conversational responses to a user's input to mimic a human dialogue. Large language models (LLMs) underpin ChatGPT's NLP capability. ChatGPT API supports the creation of a customized version of ChatGPT chatbot powered by gpt-3.5-turbo and gpt-4, which are OpenAI's LLMs, with data specific to the purpose of the custom chatbot. An LLM is a type of mathematical ML model, which is a set of parameters and structure that allows a system to make predictions. It is built with neural networks, which consist of an encoder and a decoder, which extract meanings from a sequence of text and understand the relationships between words and phrases in it (Elastic 2024.). A neural network refers to an ML model that teaches computers to process data and makes decisions in a manner similar to the way interconnected biological neurons work together in a layered structure in the human brain weighing options and learning and improving by trial and error (AWS 2024b). Pre-trained on vast amounts of data and with hundreds of billions of parameters, an LLM processes a user's input given in human language and predicts and generates plausible language as a response (AWS 2024a).

Zayed University Library's chatbot, Aisha, was customized with the content of over 100 library guides and information in the library website and a list of 100 typical questions and answers regarding academic libraries generated by ChatGPT, which were further revised and updated with Zayed University-specific information found in previously asked questions and answers from LibAnswers (Lappalainen and Narayanan 2023, 45). Aisha suggests ways in which libraries can provide users with more personalized and engaging information services by adopting and experimenting with AI/ML chatbot technology.

Promising Artificial Intelligence and Machine Learning Technologies

Although there are many other AI/ML pilots, the five examples described in this chapter point to areas of ML technologies that are likely to be relevant and promising for libraries and archives to explore and in which to invest. Computer vision is pertinent to libraries and archives as it can create, enhance, and augment item-level metadata with efficiency and accuracy. It is particularly useful for still or moving image materials as shown in pilot projects described earlier. Another area of ML technologies that holds great potential for libraries and archives is NLP, the most sophisticated example of which is a chatbot powered by an LLM. NLP techniques can be used to analyze and summarize text, assign subjects, generate summaries, and extract main claims from text materials held in libraries and archives. The work of creating and enhancing metadata for materials can significantly benefit from the tasks NLP techniques enable such as named entity recognition for people, geographic locations, and topics that appear in library materials, as the AMPPD project and the Notre Dame University Libraries project both demonstrated.

AI/ML technologies make it possible for libraries and archives to not only extract content residing in their collections as machine-readable and interpretable data but also create new or enhance existing metadata with more detailed and accurate information. By adopting and leveraging AI/ML technologies, libraries and archives can generate richer metadata at scale to be verified and further augmented by human catalogers if necessary to improve search options and to make it easier for users to discover, identify, access, navigate, cluster, analyze, and use library and archive materials. Breakthroughs in chatbot technology with LLMs point to innovative and engaging ways in which libraries and archives can provide their information and reference services tailored to their own users.

Where and How to Invest in Artificial Intelligence and Machine Learning

The previous section of this chapter highlighted areas of library and archive work that appear to be a good match for AI/ML technologies. Based on the specific areas chosen and local needs, individual libraries can determine which type of ML use would make most sense. The library's respective collections, services, current strengths, and future directions should also be taken into consideration.

Libraries and archives will also benefit by directing their efforts to address problems that emerged in previous ML pilot projects. For example, several projects discovered limitations in custom-built ML models due to available training data being too small in amount or poor in quality (Flannery 2020, 26; Lorang et al. 2020, 32–34). The greater the amount of data fed into an ML model, the better it performs. For this reason, libraries and archives can take full advantage of ML only when a sufficiently large volume of data is amassed. To ensure that an ML model performs as expected, it is also critical for libraries and archives to build reliable ground truth sets, which serve as the expected result against which the performance of an ML model is measured. Without reliable ground truth data sets, it is difficult to estimate how well an ML model meets its intended purpose. If libraries and archives decide to build more custom ML models, they should also make efforts to collaborate across institutions to create a sufficiently large training data and ground truth sets to support the creation of ML algorithms appropriate for libraries and archives to use.

Libraries and archives can also more actively explore the use of pre-trained off-the-shelf ML models. Pre-trained ML models, such as AWS Rekognition and Google Cloud Vision AI, can be useful for text recognition, face and object detection, and image labeling. It is to be noted that the pre-trained ML models have some limitations. For example, the generic ML models trained mostly with color images are likely to perform poorly with historic black-and-white photographs. For this reason, their usefulness varies depending on the type, age, and other conditions of library materials, to which those generic models are applied. If the models are applied indiscriminately, pre-trained off-the-shelf ML models may impose social and historical biases and harmful assumptions to library and archive materials due to their opaqueness. However, pre-trained off-the-shelf ML models can successfully tackle certain tasks, presenting fewer barriers for adoption, and libraries and archives can also work on better informing users regarding the use of ML models in the description of materials (Craig 2021, 203–205).

Growing generative AI cloud platforms and also something for libraries and archives to explore further. Amazon, Google, and Microsoft all offer various generative AI services and products through their cloud platforms, called Amazon Bedrock, Vertex AI, and Azure Open AI respectively. These platforms and services make it easier for AI/ML developers to access, customize, fine-tune, and deploy ML models, including many foundational models. Foundational models are large neural networks trained on massive amounts of unlabeled broad data. They are designed to produce a variety of outputs and can perform a wide range of general tasks, such as understanding languages, generating text and images, and mimicking human conversations (Jones 2023). Foundational models such as GPT-4, Imagen, and PaLM can serve as standalone systems or foundations on which other AI/ML applications with more specific purposes are built.

Lastly, libraries and archives interested in AI and ML must continue digitization. Digitization has been taking place for many years and has lost the allure of being new and exciting. However, vast numbers of physical items held within libraries and archives are still waiting to be properly digitized. Only when more of them are converted into digital objects can libraries and archives take full advantage of ML, which relies on a large amount of data rather than a set of rules to train and build an intelligent system.

Considerations for Libraries and Archives

Although interest in AI and MI has grown steadily in recent years, the use of AI and ML in libraries and archives remains experimental. Few ML applications have been developed and deployed in production by libraries and archives and led to significant improvement in their services. The situation is not vastly different in commercial library products. As indicated in Elsevier's Scopus AI, however, vendors are in the process of developing new AI products or adding new AI features to existing products by applying ML technologies. AI-enhanced products driven by ML that are marketed to libraries are increasing in number. For example, Consensus is an AI-powered search tool that answers research questions by extracting findings from scientific research papers in the Semantic Scholar database. Elicit is an AI research assistance tool that aims to expedite the literature review process by providing a list of relevant articles for a user's query in addition to summaries of content and syntheses of findings. Scite is a chatbot that allows users to find answers from the full texts of research articles with 1.2 billion citation statements extracted and analyzed from 187 million articles, book chapters, preprints, and datasets. Scopus AI offers expanded and enhanced summaries of academic articles and more refined search capabilities. Talpa Search provides users with a way to ask about and find books in a library catalog in natural language. The pace of the adoption of AL and ML by library vendors is likely to quicken, and more AI-powered products that aim to either enhance or replace the library's existing services and systems will follow.

Before the rise of generative AI, libraries and archives have taken a boutique-like approach to AI and ML. They have been experimenting more with building highly customized ML models than with testing and using generic off-the-shelf ML models. Libraries and archives have been primarily interested in training ML models with their unique and relatively small datasets and developing ML applications with functionalities related to highly specialized work ML models and techniques are most powerful when trained and applied at scale. Given that, a boutique-like approach may become a limitation to successful future applications of AI and ML in libraries and archives. It remains to be seen whether libraries and archives will begin to use more generic ML models and applications made available through growing AI cloud platforms and services.

Experimenting with AI and ML requires the preparation and assembly of large data sets. It also means that libraries and archives need new staff expertise. Cleaning up data and preparing data sets for ML work is time-consuming and labor-intensive. Building staff expertise in AI and ML can be tricky and may require hiring staff with new skills and expertise. AI and ML projects will also require additional funding to obtain appropriate ML tools and set up a robust computing environment and a necessary digital infrastructure. While it is natural for any novel endeavor to incur additional cost, the cost implications of adopting AI and ML must be given careful consideration because many libraries and archives face flat budgets or budget cuts year after year.

While investing in AI and ML may pose some logistical challenges, such investments can allow libraries and archives to promote specific areas of their services and operations where AI and ML can make significant improvements. By directing AI/ML-related work towards the areas where the impact will be high, libraries and archives can also signal to the vendors the problems that they deem most critical to be addressed in developing AI/ML applications for enhanced user experience. However, ML models and techniques are most powerful when trained and applied at scale. This will allow libraries and archives to influence the future directions of commercially available products, particularly when vendors are actively looking for new product ideas or features using AI. However, if libraries and archives miss the narrow window of time and opportunity currently open, the prospects of enhancing the activities of libraries and archives through the effective use of AI and ML may not be realized. System vendors may fail to leverage and integrate AI and ML to add value to existing products that will meet the rising expectations of both libraries and library users. Libraries and archives may end up as dissatisfied consumers of inadequate commercial ML products. And users may find libraries and archives less useful and relevant in meeting their needs.

Conclusion

This chapter has highlighted aspects of library and archive work that appear to be a good match for AI/ML technologies, described specific AI/ML technologies that are particularly relevant, and suggested areas in which libraries and archives looking to build capacity in AI and ML may invest efforts to maximize their impact. When a library or an archive decides to make an investment in AI and ML, administrators and decision-makers should carefully consider whether the project ideas connect AI/ML with significant needs in their organization, how the needed data for the project is to be obtained and prepared, whether the existing staffing and the digital infrastructure can sufficiently support the new ML project work, and whether the project team has the right mix of skills and knowledge in areas relevant to the project, including knowledge of specific library workflows, metadata, information and communications technology, data science, and software development.

Investing extensively in AI and ML may not be realistic or appropriate for all libraries and archives. But for most, AI and ML present opportunities to improve their services in new and creative ways. To seize the opportunities, library and archives staff must become familiar with how AI and ML models and techniques can be used in the library and archive context. They will also need to develop the ability to determine which library and archive tasks may most benefit from ML and to evaluate ML applications in a meaningful way. Library administrators and decision-makers must tackle the problem of how libraries and archives can develop new knowledge and skills to build long-term capacity in AI and ML to achieve desired outcomes and benefits with limited resources. They will also have to develop a compelling vision, acquire the resources needed, gain support from all levels of the organization, and build necessary staff buy-in, skills, expertise, and participation.

References

Amazon Web Services (AWS). 2024a. “What Are Large Language Models (LLM)?” https://aws.amazon.com/what-is/large-language-model/.

Amazon Web Services (AWS). 2024b. “What Is a Neural Network?” https://aws.amazon.com/what-is/neural-network/.

Amazon Web Services (AWS). 2024c. “What Is OCR?” https://aws.amazon.com/what-is/ocr/.

Craig, Jessica. 2021. “Computer Vision for Visual Arts Collections: Looking at Algorithmic Bias, Transparency, and Labor.” *Art Documentation: Journal of the Art Libraries Society of North America* 40, no.2: 198–208. https://doi.org/10.1086/716730.

Dunn, Jon W., Ying Feng, Juliet L. Hardesty, Brian Wheeler, Maria Whitaker, Thomas Whittaker, Shawn Averkamp, et al. 2021. “Audiovisual Metadata Platform Pilot Development (AMPPD) Final Project Report.” Indiana University. https://scholarworks.iu.edu/dspace/handle/2022/26989.

Elastic. 2024. “What Is a Large Language Model (LLM)?” https://www.elastic.co/what-is/large-language-models.

Flannery, Jeremiah. 2020. “Using NLP to Generate MARC Summary Fields for Notre Dame’s Catholic Pamphlets.” *International Journal of Librarianship* 5, no. 1: 20–35. https://doi.org/10.23974/ijol.2020.vol5.1.158.

IBM. n.d.a. “What Is Computer Vision?” https://www.ibm.com/topics/computer-vision.

IBM. n.d.b. “What Is Natural Language Processing (NLP)?” https://www.ibm.com/topics/natural-language-processing.

IBM. n.d.c. “What Is Speech Recognition?” https://www.ibm.com/topics/speech-recognition.

Jones, Elliot. 2023. “Explainer: What Is a Foundation Model?” Ada Lovelace Institute. July 17, 2023. https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/.

Lappalainen, Yrjo, and Nikesh Narayanan. 2023. “Aisha: A Custom AI Library Chatbot Using the ChatGPT API.” *Journal of Web Librarianship* 17, no. 3: 37–58. https://doi.org/10.1080/19322909.2023.2221477.

Lorang, Elizabeth, Leen-Kiat Soh, Yi Liu, and Chulwoo Pack. 2020. “Digital Libraries, Intelligent Data Analytics, and Augmented Description: A Demonstration Project.” *University of Nebraska-Lincoln.* https://digitalcommons.unl.edu/libraryscience/396.

Oracle Cloud Infrastructure (OCI). 2024. “What Is a Chatbot?” https://www.oracle.com/chatbots/what-is-a-chatbot/.

Prokop, Ellen, X. Y. Han, Vardan Papyan, David L. Donoho, and C. Richard Johnson jr. 2021. “AI and the Digitized Photoarchive: Promoting Access and Discoverability.” *Art Documentation: Journal of the Art Libraries Society of North America* 40, no. 1: 1–20. https://doi.org/10.1086/714604.