Application of Large Language Models for Digital Libraries

Xiaotong Hu
Informatics, University of Illinois at Urbana-Champaign
Champaign, Illinois, USA
xh9@illinois.edu

ABSTRACT

Digital library collections are valuable documents, which contain vast amounts of knowledge. The rise of artificial intelligence (AI) technologies, especially Large Language Models (LLM), has made an impact on various domains, and may also shift the focus of digital library services from storing and preserving information to utilizing and extracting knowledge from that information. We expect to add value created by LLMs to existing digital library methods so that users can better leverage the information stored in digital library collections. This research aims to improve access to digital library collections by providing users with classification labels and summaries to help easily find their works of interest leveraging large language models. Two expected major benefits are: (1) Enhancing metadata in the form of classification labels, which will help users discover and use digital library collections more easily. (2) Providing summaries of works stored in the digital library to aid researchers in finding works of interest without having to read the entire content, which will help remediate the information overload problem.

CCS CONCEPTS

• Applied computing → Digital libraries and archives; • Computing methodologies → Natural language processing.

KEYWORDS

Digital libraries, natural language processing, machine learning, artificial intelligence

ACM Reference format:

Xiaotong Hu. 2024. Application of Large Language Models for Digital Libraries. In *The 2024 ACM/IEEE Joint Conference on Digital Libraries (JCDL'24), December 16–20, 2024, Hong Kong, China. ACM, New York, NY, USA, 2 pages.* https://doi.org/10.1145/3677389.3702617

1 Introduction and Research Statement

With the advancement of information and communication technologies, digital libraries have been created to allow users to access and use digital information resources, which facilitate the consumption, creation, and sharing of knowledge in human society [3, 10]. The digitized texts in digital libraries allow users to interact with and study texts in completely new ways, providing opportunities for unprecedented research [11]. Scholars in digital humanities, cultural analytics, and computational social science are increasingly interested in leveraging digitized texts with computational techniques to answer literary, historical, and cultural questions. Meanwhile, it is important to find ways to improve the usability and accessibility of the datasets provided by digital libraries for research purposes [1, 4].

Researchers in digital libraries, digital humanities, and text mining communities have explored computational approaches to enable the use of digital library content for various downstream applications. Hu et al. [4] investigated the limitations and challenges of a curated literature dataset provided by digital libraries and improved its representativeness and usability in support of digital humanities research. Vandenbosch et al. [12] studied the identification of duplicate, variant, and partially overlapping copies of long text works in digital library collections. Jiang et al. [6] evaluated BERT embeddings for their encoding of semantic relevance and narrative coherence of OCRed digital library collections. Jiang et al. [7] explored the domain classification of book excerpts in digital libraries using BERT embeddings. Parulian et al. [9] piloted machine learning methods and word feature analysis in order to identify Black fantastic genre texts to assist finding materials of interest in digital library collections. Organisciak et al. [8] investigated fast comparison of long-format texts in a large-scale digital library. Choi et al. [2] conducted computational thematic analysis of poetry to enhance the accessibility of various poems within digital library collections. Zhang et al. [13] explored the potential of GPT models for correcting OCR outputs to enhance retrieval applications.

Noting the success of large language models on many other text-based tasks, this study will extend the aforementioned studies by applying them to authentic, real-world digital library content (i.e., unstructured, uncorrected texts that may still include OCR errors) in order to add value to existing digital library services and to assist scholars in using the digitized texts in a research context.

JCDL '24, December 16–20, 2024, Hong Kong, China
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1093-3/24/12
https://doi.org/10.1145/3677389.3702617

2 Research Questions

I aim to answer two sets of research questions in this proposed research:
RQ1: To what extent can LLMs be leveraged to automatically classify real-world digital library content?
1.1 To what extent can BERT-based classifiers be leveraged to automatically classify real-world digital library content?
1.2 To what extent can GPT-based classifiers be leveraged to automatically classify real-world digital library content?

RQ2: To what extent can LLMs be leveraged to automatically summarize real-world digital library content?
2.1 To what extent can BERT-based methods be leveraged to automatically summarize real-world digital library content?
2.2 To what extent can GPT-based methods be leveraged to automatically summarize real-world digital library content?

3 Methodology

3.1 Applications of Large Language Models for Digital Library Content Classification

The ability to effectively classify works is important for productive research in various domains. In this case study, I will empirically evaluate large language models for automatically classifying digital library content in an authentic, digitized scenario. I utilise a corpus of curated digitized library collections, which consist of 189 sampled books from the HathiTrust Digital Library [5]. The documents are selected from four domains, including agriculture, fiction, social science, and world war history [5]. I will make use of this dataset to evaluate BERT-based and GPT-based classifiers. I will use the F1 score, Precision and Recall to evaluate the performance of the models. This first case study will provide a different yet practical perspective to understand how large language models fit in the digital library contexts to improve the accessibility of the datasets provided by digital libraries.

3.2 Applications of Large Language Models for Digital Library Content Summarization

Concise and coherent summaries of digital library collections can guide researchers to quickly identify works in large-scale digital libraries. This simplifies the exploration process, and allows researchers and practitioners to quickly gain insight into digital library works, which can alleviate the information overload problem. In the second case study, I will use the subset of literary works curated from the HathiTrust Digital Library [5] to empirically evaluate large language models for automatically summarizing works in the digital library collections. I will examine the proficiency of BERT-based and GPT-based models in summarizing long literary works in an authentic, digitized scenario. I will use the Rouge score to evaluate machine generated summaries compared with gold standard human written summaries.

4 Expected Contributions

Incorporating large language models into digital library services may lead to robust solutions that enhance the exploration of real-world (i.e., mass-digitized, uncorrected, uncategorized, etc.) digital library objects, and contribute to deeper insights and a broader understanding of the digital library in general. My proposed case studies would serve as exemplars to investigate the effectiveness of large language models for processing digital library content for the classification and summarization of long texts in these authentic digital library contexts. Various language model based methods will be examined, and the methods for classification and summarization can be utilized to improve the accessibility and usability of original works within collections. The case studies provided by the project will aid scholars in exploring and using the digitized texts in digital libraries, and also pave the way for future work to improve the accessibility and curation of research data provided by digital libraries and to unlock the full potential of the digitized texts for various research questions.

REFERENCES

[1] Alex Byrne. 2003. Digital libraries: barriers or gateways to scholarly information? The electronic library 21, 5 (2003), 414–421.

[2] Kahyun Choi. 2023. Computational thematic analysis of poetry via bimodal large language models. Proceedings of the Association for Information Science and Technology 60, 1 (2023), 538–542.

[3] Gobinda G Chowdhury and Sudatta Chowdhury. 2003. Introduction to digital libraries. Facet publishing.

[4] Yuerong Hu, Ming Jiang, Ted Underwood, and J Stephen Downie. 2020. Improving digital libraries’ provision of digital humanities datasets: A case study of htrc literature dataset. In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020. 405–408.

[5] Ming Jiang, Ryan C Dubnicek, Glen Worthey, Ted Underwood, and J Stephen Downie. 2022. A prototype Gutenberg-Hathitrust sentence-level parallel corpus for OCR error analysis: Pilot investigations. In Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries. 1–5.

[6] Ming Jiang, Yuerong Hu, Glen Worthey, Ryan C Dubnicek, Ted Underwood, and J Stephen Downie. 2021. Evaluating BERT’s Encoding of Intrinsic Semantic Features of OCR’d Digital Library Collections. In 2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL). IEEE, 308–309.

[7] Ming Jiang, Yuerong Hu, Glen Worthey, Ryan C Dubnicek, Ted Underwood, and J Stephen Downie. 2021. Impact of OCR quality on BERT embeddings in the domain classification of book excerpts. Proceedings http://ceur-ws. org ISSN 1613 (2021), 0073.

[8] Peter Organisciak, Benjamin M Schmidt, and Matthew Durward. 2023. Approximate nearest neighbor for long document relationship labeling in digital libraries. International Journal on Digital Libraries 24, 4 (2023), 311–325.

[9] Nikolaus Nova Parulian, Ryan Dubnicek, Glen Worthey, Daniel J Evans, John A Walsh, and J Stephen Downie. 2022. Uncovering black fantastic: Piloting a word feature analysis and machine learning approach for genre classification. Proceedings of the Association for Information Science and Technology 59, 1 (2022), 242–250.

[10] Aaron D Purcell. 2016. Digital library programs for libraries and archives: Developing, managing, and sustaining unique digital collections. American Library Association.

[11] Jeffrey Rydberg-Cox. 2005. Digital libraries and the challenges of digital humanities. Elsevier.

[12] Adrienne VandenBosch, Benjamin M Schmidt, Krystyna K Matusiak, and Peter Organisciak. 2021. Moving Past Metadata: Improving Digital Libraries with Content-Based Methods. Proceedings of the Association for Information Science and Technology 58, 1 (2021), 849–851.

[13] James Zhang, Wouter Haverals, Mary Naydan, and Brian W Kernighan. 2024. Post-OCR Correction with OpenAI’s GPT Models on Challenging English Prosody Texts. In Proceedings of the ACM Symposium on Document Engineering 2024. 1–4.