Metadata generation and extracting keywords from unstructured text in manuscripts

Shiva Kanaujia Sukula
Dy. Librarian,
Dr B R Ambedkar Central Library
Jawaharlal Nehru University
New Delhi-110067

Parveen Babbar
Dy. Librarian
Dr B R Ambedkar Central Library
Jawaharlal Nehru University
New Delhi-110067

Mamta Rani*
Assistant Librarian
Dr B R Ambedkar Central Library
Jawaharlal Nehru University, New Delhi-110067
Email: mamta728@gmail.com

Bibhuti Bhushan Pattanaik
Assistant Librarian
Dr B R Ambedkar Central Library
Jawaharlal Nehru University, New Delhi-110067
Email: bibhutipattanaik@gmail.com

Corresponding author

Abstract

This study presents a scoping review of metadata generation and keyword extraction from unstructured text, with a particular focus on ancient Indian and multi-lingual manuscript collections. The following study examines a total of 126 publications from 2015 to 2025 to assess the progress of computational methods, spanning rule-based approaches, traditional machine learning, deep learning frameworks, modern artificial intelligence, and other advances using LLMs (large language models). The review highlights significant advancements in Optical Character Recognition (OCR) for Indic scripts, including Sanskrit, Devanāgari, Tamil, and Malayalam, as well as improvements in Natural Language Processing (NLP) for summarization, translation, and knowledge extraction from classical literature. Despite progress, persistent research challenges remain, including the limitations of large annotated datasets, script and orthographic variation across regions and historical timelines, and document degradation in terms of image quality. The study highlights the crucial role of computational approaches in preserving cultural heritage and advocates for the development of standardized benchmark datasets, scalable

processing tools, and inclusive AI systems to support future scholarly research, discovery, and user accessibility of historical manuscripts worldwide.

Keywords: Metadata generation; Keyword extraction; Digital preservation; Indic scripts; Optical Character Recognition (OCR); Natural Language Processing (NLP); Ancient manuscripts; Deep learning; Large Language Models (LLMs)

Introduction: Generative Metadata and Keyword Extraction

The extraction of metadata shifted from earlier approaches based on rules and structures to more advanced systems based on AI. Many basic tools, such as CERMINE (Content ExtRactor and MINEr) demonstrated that automatic extraction extracted from scientific PDFs was adequately performance (Tkaczyk et al., 2015), and subsequent proposed, rule-based approaches demonstrate improvements in accuracy of metadata extraction across document types (Azimjonov& Alikhanov, 2018; Alghamdi et al., 2022). Eventually, the automatic extraction of complicated components such as multicultural metadata and algorithmic representations was achieved through deep learning models (Choi et al., 2023; Safder et al., 2020). Neural architectures that involved embeddings to grasp semantic richness were at the core of the context-aware extraction of relevant keywords (Zhang et al., 2020; Khan et al., 2022). Sci-tech-scale distributed and serverless (Skluzacek et al., 2021) processing for scholarly collections are example of comparable research that is going on to be able to scale. Recent research suggests that AI in conjunction with large language models, will be a game-changer in research reproducibility, metadata organization, and discovery (Yang et al., 2025; Formanek, 2025). Essentially, these works opened the way for scholarly communication and made data more accessible.

Major insights in Metadata extraction:

- Research on metadata extraction in scholarly communication has evolved beyond simple rule-based systems to incorporate deep learning and multi-lingual techniques for higher accuracy.
- Foundational tools like CERMINE and PDFMEF (PDF Multi-Entity Framework) proved automated metadata extraction is feasible, which enabled more specialized systems for legal and algorithm-related metadata.
- Present innovations primarily concentrate on scalability through serverless and bulk processing methods, in addition to feature-oriented frameworks like *FLAGS: Federated Learning AlGorithms Simulation*-PDFe.
- AI improvements are turning metadata into a more valuable resource for reproducibility, recommendations, and contextual keyword extraction.
- The cutting-edge movement heavily features generative AI and the matching of metadata quality to real user requirements in areas such as diversity and digital image collections.

Keywords extraction from unstructured text in manuscript

Research in unstructured text analytics has notably evolved in concurrence with the utilization of data mining, NLP, and AI-powered techniques. The initial research emphasized the use of rule-based and analytical frameworks for supporting the finding and recognizing of new patterns in large text corpora (Al-Barhamtoshy& Eassa, 2015; King et al., 2017). Besides, the current models rely on supervised as well as unsupervised learning for summarization, clustering, and invisible pattern revelation (Lydia et al., 2020; Saeed et al., 2020). Moreover, the advent of contextual embeddings has brought considerable progress in keyword getting and knowledge of the functional area (Khan et al., 2022; Cheong et al., 2017). The present document AI techniques have extended the frontier of the medical, safety, and clinical areas, among others, to include reporting, documentation, and violence detection, respectively (Malashin et al., 2024; Mirończuk, 2020; Botelle et al., 2022). The most recent trends have, among other things, efficiency and future scalability as the core of their concern, and feature cutting-edge extraction (Yeghiazaryan et al., 2022; Mahadevkar et al., 2024) systems accordingly

Keywords extraction from manuscripts

Research in academic text processing keeps on changing through the stages of automation and the betterment of scholarly discovery. The main focus of the initial work was on the direct extraction of abstracts and keywords from the article context for the purposes of retrieval and analysis (Müngen& Kaya, 2018; Salloum et al., 2017). Research on manuscripts propelled the invention of historical document search methods that required no learning-based detection (Mohammed et al., 2021) and the development of transcription technology like Transkribus (Muehlberger et al., 2019) to aid the facilitation of archival scholarship. The progress in the field of deep learning has led to the rapid development of abstractive summarization through the use of hybrid Convolutional Neural Network Long Short-Term Memory Network Architectures (CNN-LSTM) (Song et al., 2019; Zaman et al., 2020). Current innovations turn the spotlight on generative AI as a research tool while also drawing attention to the effect of dataset construction on the ethical side of the matter (Glickman & Zhang, 2024; Scheuerman et al., 2021).

Keywords extraction from Indian Sanskrit manuscripts

Research on Sanskrit and Indic manuscripts has quickly changed through the integration of OCR, NLP, and deep learning(Tomar et al., 2015; Narang et al., 2019). to make the texts more accessible. The main points are on the preparation of the text, feature recognition, and the identification of ancient documents in Sanskrit and Devanāgari scripts. The improvements in text summarization have been gradually moving from purely extractive methods (Barve et al., 2015; Sinha & Jha, 2020; Bhatnagar et al., 2023; Gupta & Shah, 2025) to hybrid deep learning architectures such as BERT(Bidirectional Encoder Representations from Transformers a Google-developed AI model that revolutionized Natural Language Processing (NLP) for understanding word context bidirectionally. The quality of speech-temporization has been enhanced by using attention-based frameworks. Different example of deep learning are CNN-BiLSTM(Convolutional Neural Network Bidirectional Long Short-Term Memory for severely palm-leaf manuscripts (Kataria & Jethva, 2019; Narang et al., 2020; Sudarsan & Sankar, 2022) as well as new character sets for Malayalam and Sanskrit (Dhruva et al., 2023; Krishnan et al., 2025; Nair & Rani, 2023). The scope of NLP applications in classical texts has been increased, thus allowing knowledge

extraction (Srivastava et al., 2018; Sujoy et al., 2023; Jain et al., 2025; Kumari & Malik, 2024) from such works as Charak Samhita. Machine translation, speech technologies offer new ways of accessing and interacting with the cultural heritage. (Koul & Manvi, 2021; Anoop & Ramakrishnan, 2019; Chand et al., 2023) The latest reviews show an increasing focus on the Vedic scripts, Pali preservation, and neural transliteration as a means of creating a sustainable digital infrastructure (Samantaray et al., 2025; Gudadhe et al., 2024; Pradeep & Mamidi, 2025) for Sanskrit scholarship with the following issues:

- Improved OCR for Indic scripts
- Growth in NLP applications for classical literature
- Progress in summarization and translation
- Stronger focus on digital preservation and cultural heritage

Data in table (Appendix 1) reflects various focus areas and developments.

Scoping Review: Information Extraction and Computational Methods for Ancient Indian and Multi-lingual Manuscripts (2015–2025)

This scoping review is a landmark in the field since it compiles in one comprehensive document all the research efforts related to information extraction, OCR, NLP, and AI-based processing of ancient Indian as well as multi-lingual manuscripts. The existing research works are quite varied and scattered across multiple fields like computer science, linguistics, digital humanities, and library science. The study by merging these different fields not only offers a broader but also a deeper understanding of the current digitization level of the heritage content with the benefits as well as the drawbacks involved. It brings out how essential computational methods are to save the knowledge that is at the risk of extinction which is not only in Sanskrit, Tamil, Malayalam but also in Devanāgari and other scripts of the Indic languages. In that way, it takes care of the problems of physical degradation, limited accessibility and linguistic complexity arising from the content

The review emphasizes that AI-powered methods are very important for performing such tasks as metadata extraction, transliteration, summarization, translation, and semantic understanding with the ultimate goal of making cultural materials more discoverable and available to scholars. The findings of this research work not only fuel up national digitization initiatives but also serve as a perfect map which guides libraries, archives, and policy makers in adopting viable solutions that offer a wide range of applications. It advocates the requirement for uniform datasets, benchmark evaluation, and gender-neutral language technology development to research more on low-resource scripts. By pinpointing research gaps and next-step priorities, this review becomes a direct agent for the long-term digital preservation and international sharing of India's intellectual heritage.

Data collection and search strategy

A systematic search was conducted in Google Scholar using the following search phrases:

1. "Generate metadata and extract keywords from manuscript"
2. "Extract keywords from unstructured text in manuscript"
3. "Extract keywords from manuscripts"
4. "Extract keywords from Indian Sanskrit manuscripts"

The searches focused on literature related to keyword extraction, metadata generation, and information processing for ancient and Indic manuscripts. In total, 129 references were initially retrieved. After screening for relevance and removing 3 duplicate records, 126 publications were included in the scoping review (Appendix 1).

Structured Thematic Synthesis

The present scoping review is based on 126 references that were identified through systematic searches in Google Scholar. These sources are the research articles, conference publications, and scholarly studies that address the computational processing of the manuscripts written in the Indic script. The collected materials reflect the cross-disciplinary contributions from the fields of computer science, linguistics, heritage studies, and library and information science

1. Information Extraction & Unstructured Big Data Analytics

Key Contributions

- Analytical studies highlight performance challenges in extracting information from heterogeneous and multidimensional data (Adnan & Akbar, 2019a, 2019b).
- Text mining and NLP-based knowledge extraction frameworks improved organizational insight and content navigation (Salloum et al., 2017; Lydia et al., 2020).
- Methods evolved from rule-based extraction (Azimjonov& Alikhanov, 2018) to deep-learning–based metadata understanding (Safder et al., 2020).

Metadata extraction remains crucial for:

- scientific publications (Ahmed & Afzal, 2020; Meng et al., 2018)
- legal documents (Sleimi et al., 2021)
- libraries and archives (Therrell, 2019; Leipzing et al., 2021)

2. OCR and Handwritten Text Recognition for Indic Manuscripts

Representative Works are given as following:

- Sanskrit OCR: Kataria & Jethva (2019), Madake et al. (2023)
- Devanāgari ancient script: Narang et al. (2019, 2021, 2022)
- Tamil palm leaf manuscripts: Subramani &Murugavalli (2019); Maheswari et al. (2024)
- Malayalam character recognition: Sudarsan & Sankar (2022, 2024)
- Region-specific datasets: Nair & Rani (2023), Kesiman et al. (2018)

Innovations include noise reduction, character segmentation, and dataset normalization (Krishnan et al., 2025).

3. Ancient Text NLP: Summarization, Keyword Extraction & Knowledge Mining

Highlights

- Sanskrit summarization using BERT-based extractive models (Bhatnagar et al., 2023) and emerging abstractive methods (Sinha, 2025).
- Keyword extraction enhanced via contextual embeddings (Khan et al., 2022; Patil & Ramteke, 2023).
- Domain-specific systems extract knowledge from Ayurveda and Itihasa texts (Bagchi et al.; Jain et al., 2025; Shankar et al.)

4. Machine Translation, Named Entity Recognition & Linguistic Modeling

Key Advancements

- Sanskrit↔English translation via NMT and rule-based hybrids (Koul & Manvi, 2021; Sethi et al., 2022; Sitender& Bawa, 2022).
- Word embeddings and morphological parsing for semantic preservation (Sandhan et al., 2021; Krishnan et al., 2025).
- NER dataset creation for Sanskrit heritage texts (Sujoy et al., 2023).
- Stress prediction and sentiment analysis using deep learning (Kumar et al., 2023; Kumari & Malik, 2024).

5. Generative AI, LLMs & Digital Humanities Applications

Recent publications emphasize:

- LLM-assisted summarization and data curation (Glickman & Zhang, 2024)
- Generative AI adoption in libraries (Formanek, 2025)
- Dataset politics and inclusivity (Scheuerman et al., 2021)

The identified risk is in the form of "Bias in LLM outputs if datasets underrepresent regional languages."

Sl.No.	Research Gap	Evidence	Opportunities
1	Lack of large annotated Indic manuscript datasets	Multiple works highlight data scarcity	National digitization + crowdsourced annotation
2	Script variation, ligatures & damaged documents	OCR studies repeatedly mention segmentation difficulty	Multimodal models + restoration pipelines
3	Limited semantic	Knowledge mining still	Knowledge graphs + cultural

	understanding for Sanskrit	shallow	ontology development
4	Benchmarking inconsistencies across studies	Varied metrics and corpora	Unified evaluation frameworks for Indic OCR/NLP
5	Metadata workflows not integrated into libraries	Large language models and generative artificial intelligence (GPT) still at intermediate stage	LLM-driven metadata automation in repositories

Transitions and developments

The transitions in processing the unstructured text in manuscripts is given as following:

1. Digitization, Preservation, and Image Analysis of Manuscripts

It focuses on safeguarding the physical and informational integrity of ancient palm-leaf, Sanskrit, Tamil, and Malayalam manuscripts. More emphasis on degradation and non-uniform writing surfaces require specialized computer-vision workflows. Dataset creation and image benchmarking demonstrate the foundational need for high-quality annotated data before applying advanced recognition models.

2. OCR and Handwritten Text Recognition for Indic Scripts

Deep learning has largely replaced traditional feature-engineered OCR strategies for palm-leaf and Devanāgari manuscripts. Approaches include Convolutional Neural Network (CNN) , Long Short-Term Memory (LSTM) network, Capsule Network (CapsNet) , and Hybrid Discrete Wavelet Transform-Convolutional Neural Network (**DWT-CNN**) architectures. Performance is strongly tied to the diversity and volume of training samples, which remain limited due to the rare nature of ancient content.

3. Knowledge Extraction and Summarization

The goal is to facilitate the conversion of unstructured Sanskrit prose, medical heritage, and epic literature into well-defined, searchable knowledge units. Some of the techniques utilized in summarization are clustering-driven extraction, BERT-based semantic compression, and rule-based information retrieval. These techniques have the potential to considerably lessen the cognitive load and make available to the general public a vast amount of historical knowledge hidden in classical texts. However, process is hindered by semantic ambiguity and the lack of domain-specific gold standards for evaluation.

4. Metadata Extraction from Scholarly Documents

The implementation of metadata extraction has changed from being rule-based heuristics to scalable neural architectures. CERMINE extracts PDFs citations- titles, authors, and structural metadata, whereas multi-lingual frameworks aim at worldwide interoperability. Standardization and the presence of

benchmarks are the main factors that determine the progress in reproducibility and discoverability in academic repositories

5. NLP for Sanskrit Language Processing

Language processing focus on morphological parsing, POS-tagging, Named Entity Recognition, deep embedding evaluation, and computational vocabulary modeling. The morphology of Sanskrit is very complicated and the phenomena of sandhi and rich morphology make the language very difficult for conventional NLP techniques, especially when there are few annotated corpora.

6. Machine Translation and Bilingual Systems

Among the methods used by the Grammarians is the utilization of rule-based grammar mechanisms together with neural machine translation to connect Sanskrit with English and Hindi. The employment of hybrid systems is instrumental in solving the problem of sparse data; thus, at the same time, they are aimed at retaining the semantic nuance. Machine translation is very important, in particular, for non-expert users and the cultural tourism sector. To a large extent, accuracy is still limited by the small number of parallel corpora, but there is quite a bit of recent progress making transliteration and translation solutions more user-friendly

7. Keyword Extraction and Indexing

By incorporating contextual embeddings and word matching in ancient scripts, keyword detection has immensely reduced the time of manuscript study. These instruments enable thematic keyword exploration and indexing to be carried out efficiently.

8. AI for Big Data and Unstructured Text Analytics

AI acts as an enabler in revealing the deep-seated patterns and in the conversion of the unstructured textual heritage of the past into the structured computational objects. It brings together the processes of the ancient document digitization, modern Knowledge-Graph and Data-Mining practices

9. Digital Humanities and Cultural Knowledge Systems

The emphasisis on the important cultural aspects of making Sanskrit available digitally to everyone. The tools that provide instant retrieval, epic narrative mapping, and tourism assistance, among other things, help to continue the societal value that comes from conservation. These projects bring the humanistic side of the story to the forefront, which is the driving force behind the technical innovations.

10. Resource Creation and Standardization

The most crucial knowledge infrastructure is formation of large, labeled datasets for handwriting, word segmentation, and computational lexicons. These resources demonstrate a community shift toward open datasets and benchmarking, essential prerequisites for reproducible progress.

References:

1. Adnan, K., & Akbar, R. (2019). An analytical study of information extraction from unstructured and multidimensional big data. Journal of Big Data, 6(91), 1-38.
2. Adnan, K., & Akbar, R. (2019). Limitations of information extraction methods and techniques for heterogeneous unstructured big data. International Journal of Engineering Business Management, 11, 1847979019890771.
3. Agarwal, M., Indu, S., Jayanthi, N. (2024). An Approach to the Classification of Ancient Indian Scripts Using the CNN Model. In: Mehta, G., Wickramasinghe, N., Kakkar, D. (eds) Innovations in VLSI, Signal Processing and Computational Technologies. WREC 2023. Lecture Notes in Electrical Engineering, vol 1095. Springer, Singapore.
4. Ahmed, M. W., & Afzal, M. T. (2020). FLAG-PDFe: Features Oriented Metadata Extraction Framework for Scientific Publications," in IEEE Access, vol. 8, pp. 99458-99469, 2020, doi: 10.1109/ACCESS.2020.2997907.
5. Al-Barhamtoshy, H., & Eassa, F. (2014). A data analytic framework for unstructured text. Life Science Journal, 11(10), 339-350
6. Alghamdi, H., Dawwas, W., Almutairi, T. H., & Rahman, A. (2022). Extracting ToC and Metadata from PDF Books: A Rule-Based Approach. ICIC Express Letters, Part B: Applications, 13(2), 133-143.
7. Amur, Z. H., Hooi, Y. K., Soomro, G. M., Bhanbhro, H., Karyem, S., & Sohu, N. (2023). Unlocking the Potential of Keyword Extraction: The Need for Access to High-Quality Datasets. Applied Sciences, 13(12), 7228. https://doi.org/10.3390/app13127228.
8. Anoop, C. S., & Ramakrishnan, A. G. (2019, July). Automatic speech recognition for Sanskrit. In 2019 2nd international conference on intelligent computing, instrumentation and control technologies (ICICICT), Kannur, India, 2019, pp. 1146-1151, doi: 10.1109/ICICICT46008.2019.8993283.
9. Azimjonov, J., & Alikhanov, J. (2018). Rule based metadata extraction framework from academic articles. *ArXiv, abs/1807.09009*.
10. Bagchi, P., Jain, V., & Kharat, A. (2024). NLP-based knowledge extraction from Charak Samhita for navigating ancient wisdom: a Django framework approach. Proceedings of RSU International Research Conference (RSUCON-2024), 555-566.
11. Barve, S., Desai, S., Sardinha, R. (2016). Query-Based Extractive Text Summarization for Sanskrit. In: Das, S., Pal, T., Kar, S., Satapathy, S., Mandal, J. (eds) Proceedings of the 4th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA) 2015. Advances in Intelligent Systems and Computing, vol 404. Springer, New Delhi. https://doi.org/10.1007/978-81-322-2695-6_47.
12. Bhatnagar, K., Lonka, S., Kunal, J., & G, M. R. M. (2023, April 4). San-BERT: Extractive Summarization for Sanskrit Documents using BERT and it's variants. ArXiv.org. https://doi.org/10.48550/arXiv.2304.01894
13. Botelle, R., Bhavsar, V., Kadra-Scalzo, G., Mascio, A., Williams, M. V., Roberts, A., Velupillai, S., & Stewart, R. (2022). Can natural language processing models extract and classify instances of interpersonal violence in mental healthcare electronic records: an applied evaluative study. BMJ Open, 12(2), e052911. https://doi.org/10.1136/bmjopen-2021-052911.
14. Chadha, S., Mittal, S., & Singhal, V. (2019). An Insight of Script Text Extraction Performance using Machine Learning Techniques. International Journal of Innovative Technology and Exploring Engineering, 9(1), 2581–2588. https://doi.org/10.35940/ijitee.a5224.119119

15. Chand, A., Agarwal, P., & Sharma, S. (2023), Real-Time Retrieving Vedic Sanskrit Text into Multi-Lingual Text and Audio for Cultural Tourism Motivation," 2023 International Conference for Advancement in Technology (ICONAT), Goa, India, 2023, pp. 1–6, doi: 10.1109/ICONAT57137.2023.10080862.
16. Cheong, H., Li, W., Cheung, A., Nogueira, A., & Iorio, F. (2017). Automated Extraction of Function Knowledge From Text. Journal of Mechanical Design, 139(11), 111407. https://doi.org/10.1115/1.4037817.
17. Choi, W., Yoon, H.-M., Hyun, M.-H., Lee, H.-J., Seol, J.-W., Lee, K. D., Yoon, Y. J., & Kong, H. (2023). Building an annotated corpus for automatic metadata extraction from multilingual journal article references. PLoS ONE, 18(1), e0280637–e0280637. https://doi.org/10.1371/journal.pone.0280637.
18. Deepthi, C. V. S., & Seenu, A. (2022, December). A Systematic Review on OCRs for Indic Documents & Manuscripts. In 2022 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI) (Vol. 1, pp. 1-4). IEEE.
19. Dhingra, V., & Joshi, M. M. (2022). Rule based approach for compound segmentation and paraphrase generation in Sanskrit. International Journal of Information Technology, 14(6), 3183-3191.
20. Dhruva, G., Kore, V., Vijitha, M., Rao, S., & Preethi, P. (2023, December). Comprehensive dataset building of isolated handwritten Sanskrit characters. In International Conference on Applied Soft Computing and Communication Networks (pp. 489-503). Singapore: Springer Nature Singapore.
21. Felix, C., Pandey, A. V., & Bertini, E. (2016). Texttile: An interactive visualization tool for seamless exploratory analysis of structured data and unstructured text. IEEE transactions on visualization and computer graphics, 23(1), 161-170.
22. Formanek, M. (2025). Exploring the potential of large language models and generative artificial intelligence (GPT): Applications in Library and Information Science. Journal of Librarianship and Information Science, 57(2), 568-590.
23. Fulda, J., Brehmer, M., &Munzner, T. (2015). TimeLineCurator: Interactive authoring of visual timelines from unstructured text. IEEE transactions on visualization and computer graphics, 22(1), 300-309.
24. Garg, A., Tiwari, L., Juj, T., Indu, S., & Jayanthi, N. (2021). Language and Era Prediction of Digitized Indian Manuscripts Using Convolutional Neural Networks. In Sentimental Analysis and Deep Learning: Proceedings of ICSADL 2021 (pp. 703-718). Singapore: Springer Singapore.
25. Glickman, M., & Zhang, Y. (2024). AI and generative AI for research discovery and summarization. arXiv preprint arXiv:2401.06795.
26. Gudadhe, S. R., Bardekar, A. A., & Ranit, A. B. (2024, October). A novel approach using machine learning and NLP for revolutionizing Pali manuscript conservation. In 2024 4th International Conference on Sustainable Expert Systems (ICSES) (pp. 844-848). IEEE.
27. Gupta, V. K., & Shah, H. R. (2025, February). Summarization of Sanskrit Text: Approaches and Techniques. In 2025 International Conference on Computational, Communication and Information Technology (ICCCIT) (pp. 643-648). IEEE.
28. Guruprasad, P., & KS Rao, G. (2021). Recognition of Handwritten Nandinagari Palm Leaf Manuscript Text. In Computational Intelligence Methods for Super-Resolution in Image Processing Applications (pp. 177-190). Cham: Springer International Publishing.
29. Hameed, P., Koikara, R., & Sharma, C. (2015, September). Segmentation of Ancient and Historical Gilgit Manuscripts. In Proceedings of the Second International Conference on

Computer and Communication Technologies: IC3T 2015, Volume 1 (pp. 443-447). New Delhi: Springer India.

30. Jain, V., Bagchi, P., Kharat, A., & Shivani, V. (2025). Extracting Invaluable Insights from Sushruta Samhita Using Natural Language Processing. International Journal of Public Mental Health and Neurosciences, 12(2), 10-14.
31. Jaiswal, P., & Singh, A. P. (2022). Conservation of knowledge heritage and national mission for manuscripts in Uttar Pradesh and Kerala.
32. Jindal, A., & Ghosh, R. (2024). A semi-self-supervised learning model to recognize handwritten characters in ancient documents in Indian scripts. Neural Computing and Applications, 36(20), 11791-11808.
33. Kabra, D., Gohel, R., Prajapati, S., & Gupta, M. K. (2025). Statistical Analysis of Hindi and Sanskrit Languages. Authorea Preprints.
34. Kakimoto, H., Hayashi, T., Wang, Y., Kawai, Y., & Sumiya, K. (2018). Query keyword extraction from video caption data based on spatio-temporal features. In Proceedings of the International MultiConference of Engineers and Computer Scientists (Vol. 1, pp. 405-408).
35. Kataria, B., & Jethva, H. B. (2018). Review of Advances in Digital Recognition of Indian Language Manuscripts. International Journal of Scientific Research in Science, Engineering and Technology, 4(1), 1302-1318.
36. Kataria, B., & Jethva, H. B. (2019). CNN-bidirectional LSTM based optical character recognition of Sanskrit manuscripts: a comprehensive systematic literature review. Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.(IJSRCSEIT), 5(2), 2456-3307.
37. Kataria, D. B., & Jethva, H. B. (2021). Optical Character Recognition of Sanskrit Manuscripts Using Convolution Neural Networks. Webology (ISSN: 1735-188X) Volume, 18.
38. Kesiman, M. W. A., Valy, D., Burie, J. C., Paulus, E., Suryani, M., Hadi, S., ... & Ogier, J. M. (2018). Benchmarking of document image analysis tasks for palm leaf manuscripts from southeast asia. Journal of Imaging, 4(2), 43.
39. Khan, M. Q., Shahid, A., Uddin, M. I., Roman, M., Alharbi, A., Alosaimi, W., ... & Alshahrani, S. M. (2022). Impact analysis of keyword extraction using contextual word embedding. PeerJ Computer Science, 8, e967.
40. Kim, S., Choi, H., Kim, N., Chung, E., & Lee, J. Y. (2018). Comparative analysis of manuscript management systems for scholarly publishing. Science Editing, 5(2), 124-134.
41. King, G., Lam, P., & Roberts, M. E. (2017). Computer-assisted keyword and document set discovery from unstructured text. American Journal of Political Science, 61(4), 971-988.
42. Kore, V., Dhruva, G., Rao, S., Vijitha, M., & Preethi, P. (2025). A systematic framework for Sanskrit character recognition using deep learning. ELCVIA Electronic Letters on Computer Vision and Image Analysis, 24(1), 81-103.
43. Koul, N., & Manvi, S. S. (2021). A proposed model for neural machine translation of Sanskrit into English. International Journal of Information Technology, 13(1), 375-381.
44. Krishnan, S., Kulkarni, A., & Huet, G. (2025). Normalized dataset for Sanskrit word segmentation and morphological parsing. Language Resources and Evaluation, 59(2), 1279-1330.
45. Kulkarni, I., Tikkal, S., Chaware, S., Kharate, P., & Pandit, A. (2022, February). Proposed Design to Recognize Ancient Sanskrit Manuscripts with Translation Using Machine Learning. In Proceedings of the International Conference on Innovative Computing & Communication (ICICC).

46. Kumar, P., Pathania, K., & Raman, B. (2023). Zero-shot learning based cross-lingual sentiment analysis for sanskrit text with insufficient labeled data. Applied Intelligence, 53(9), 10096-10113.
47. Kumar, R., Tewari, P., Thakur, R. K., & Kumar, R. (2024). ENGLISH TO SANSKRIT TRANSLATION USING NMT. Available at SSRN 4938136.
48. Kumari, S., & Malik, A. (2024). Making Machines Talk In Sanskrit: A systematic exploration Of Text-To-Speech Synthesis For Sanskrit Language. Journal of Computational Analysis & Applications, 33(8).
49. Kumari, S., & Malik, A. (2024). Predicting Stress in Sanskrit Texts: A Deep Learning Approach to Sentiment Analysis. International Journal of Multiphysics, 18(3).
50. Kuźma, M., &Mościcka, A. (2020). Evaluation of metadata describing topographic maps in a National Library. Heritage Science, 8(1).
51. Leipzig, J., Nüst, D., Hoyt, C. T., Ram, K., & Greenberg, J. (2021). The role of metadata in reproducible computational research. Patterns, 2(9).
52. Leung, J. K., Griva, I., & Kennedy, W. G. (2020). Making use of affective features from media content metadata for better movie recommendation making. arXiv preprint arXiv:2007.00636.
53. Löffler, F., Wesp, V., König-Ries, B., & Klan, F. (2021). Dataset search in biodiversity research: Do metadata in data repositories reflect scholarly information needs?. PloS one, 16(3), e0246099.
54. Lomte, M. V. M., & Doye, D. D. (2022). Handwritten Vedic Sanskrit text recognition using deep learning. Journal of Algebraic Statistics, 13(3), 2190-2198.
55. Lydia, E. L., Kannan, S., SumanRajest, S., & Satyanarayana, S. (2020). Correlative study and analysis for hidden patterns in text analytics unstructured data using supervised and unsupervised learning techniques. International Journal of Cloud Computing, 9(2-3), 150-162.
56. Madake, J., Yedle, Y., Shahabade, V., &Bhatlawande, S. (2023, June). Sanskrit OCR System. In International Conference on Advanced Communication and Intelligent Systems (pp. 188-200). Cham: Springer Nature Switzerland.
57. Mahadevkar, S. V., Patil, S., Kotecha, K., Soong, L. W., & Choudhury, T. (2024). Exploring AI-driven approaches for unstructured document analysis and future horizons. Journal of Big Data, 11(1), 92.
58. Maheshwari, A., Ajmera, R., &Dharamdasani, D. K. (2023). Handwritten Vedic Sanskrit Text Recognition Using Deep Learning and Convolutional Neural Networks. Auricle Global Society of Education and Research.
59. Maheswari, S. U., Maheswari, P. U., & Aakaash, G. S. (2024). An intelligent character segmentation system coupled with deep learning based recognition for the digitization of ancient Tamil palm leaf manuscripts. Heritage Science, 12(1), 342.
60. Malashin, I., Masich, I., Tynchenko, V., Gantimurov, A., Nelyub, V., &Borodulin, A. (2024). Image text extraction and natural language processing of unstructured data from medical reports. Machine Learning and Knowledge Extraction, 6(2), 1361-1377.
61. Meng, B., Hou, L., Yang, E., & Li, J. (2018, October). Metadata extraction for scientific papers. In China National Conference on Chinese Computational Linguistics (pp. 111-122). Cham: Springer International Publishing.
62. Menon, A., Choi, J., & Tabakovic, H. (2018, July). What you say your strategy is and why it matters: natural language processing of unstructured text. In Academy of management proceedings (Vol. 1, p. 18319). Briarcliff Manor, NY 10510: Academy of Management.

63. Mirończuk, M. M. (2020). Information extraction system for transforming unstructured text data in fire reports into structured forms: a Polish case study. Fire technology, 56(2), 545-581.
64. Mohammed, H., Märgner, V., & Ciotti, G. (2021). Learning-free pattern detection for manuscript research: An efficient approach toward making manuscript images searchable. International Journal on Document Analysis and Recognition (IJDAR), 24(3), 167-179.
65. Moudgil, A., Singh, S., & Gautam, V. (2021). An overview of recent trends in OCR systems for manuscripts. Cyber Intelligence and Information Retrieval: Proceedings of CIIR 2021, 525-533.
66. Moudgil, A., Singh, S., Gautam, V., Rani, S., & Shah, S. H. (2023). Handwritten devanāgari manuscript characters recognition using capsnet. International Journal of Cognitive Computing in Engineering, 4, 47-54.
67. MS, R., Mallikarjuna, C., & VS, A. (2020). NLP-Driven Knowledge Extraction and Thematic Classification of Translated Ancient Indian Medical Texts. Rajeevan, MS, Mini devi, B., Anoop, VS, & Mallikarjuna, C.(2025). NLP-driven Knowledge Extraction and Thematic Classification of Translated Ancient Indian Medical Text. Reimagining LIS Education: Collaborative Integration of Indian Knowledge System with NEP, 1, 351.
68. Muehlberger, G., Seaward, L., Terras, M., Ares Oliveira, S., Bosch, V., Bryan, M., ... &Zagoris, K. (2019). Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. Journal of documentation, 75(5), 954-976.
69. Müngen, A. A., & Kaya, M. (2018). Extracting abstract and keywords from context for academic articles. Social Network Analysis and Mining, 8(1), 45.
70. Nair, B. B., & Rani, N. S. (2023). HMPLMD: Handwritten Malayalam palm leaf manuscript dataset. Data in Brief, 47, 108960.
71. Nam, S. T., Shin, S. Y., & Jin, C. Y. (2021). Text Mining and Visualization of Unstructured Data Using Big Data Analytical Tool R. 한국정보통신학회논문지, 25(9), 1199-1205.
72. Narang, S. R., Jindal, M. K., Ahuja, S., & Kumar, M. (2020). On the recognition of Devanāgari ancient handwritten characters using SIFT and Gabor features. Soft Computing, 24(22), 17279-17289.
73. Narang, S. R., Kumar, M., & Jindal, M. K. (2021). DeepNetDevanagari: a deep learning model for Devanāgari ancient character recognition. Multimedia Tools and Applications, 80(13), 20671-20686.
74. Narang, S. R., Kumar, M., & Jindal, M. K. (2022, December). Optimization of Character Classes in Devanāgari Ancient Manuscripts and Dataset Generation. In International Conference on Frontiers in Computing and Systems (pp. 59-69). Singapore: Springer Nature Singapore.
75. Narang, S., Jindal, M. K., & Kumar, M. (2019). Devanāgari ancient documents recognition using statistical feature extraction techniques. Sādhanā, 44(6), 141.
76. Nigam, A., & Chandra, S. (2022, June). Digital Accessibility and Information Mining of Dharmaśāstric Knowledge Traditions. In Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference (pp. 42-47).
77. Nigam, A., & Chandra, S. (2022). Digital World of Dharmaśāstric Knowledge Tradition: An Instant Information Retrieval System for Manusmṛiti. GIS: Science Journal, 9(8), 241-249.
78. Patil, N. P., & Ramteke, R. J. (2023). A novel optimized deep learning framework to spot keywords and query matching process in Devanāgari scripts. Multimedia Tools and Applications, 82(19), 30177-30199.

79. Pradeep, A., & Mamidi, R. (2025). Sandaŕśana: A Survey on Sanskrit Computational Linguistics and Digital Infrastructure for Sanskrit. ACM Computing Surveys, 57(10), 1-38.
80. Puri, S., & Singh, S. P. (2019). An efficient Devanāgari character classification in printed and handwritten documents using SVM. Procedia Computer Science, 152, 111-121.
81. Qayyum, F., & Afzal, M. T. (2019). Identification of important citations by exploiting research articles' metadata and cue-terms from content. Scientometrics, 118(1), 21-43.
82. Rahman, A. U., Musleh, D., Nabil, M., Alubaidan, H., Gollapalli, M., Krishnasamy, G., ... & Mahmud, M. (2022). Assessment of Information Extraction Techniques, Models and Systems. Mathematical Modelling of Engineering Problems, 9(3).
83. Razack, H. I. A., Mathew, S. T., Saad, F. F. A., & Alqahtani, S. A. (2021). Artificial intelligence-assisted tools for redefining the communication landscape of the scholarly world. Science editing, 8(2), 134-144.
84. Saeed, M. Y., Awais, M., Talib, R., & Younas, M. (2020). Unstructured text documents summarization with multi-stage clustering. IEEE Access, 8, 212838-212854.
85. Safder, I., Hassan, S. U., Visvizi, A., Noraset, T., Nawaz, R., &Tuarob, S. (2020). Deep learning-based extraction of algorithmic metadata in full-text scholarly documents. Information processing & management, 57(6), 102269.
86. Saini, J. R., & Bafna, P. B. (2020). Measuring the Similarity between the Sanskrit Documents using the Context of the Corpus. International Journal of Advanced Computer Science and Applications, 11(5).
87. Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalan, K. (2017). Using text mining techniques for extracting information from research articles. In Intelligent natural language processing: Trends and Applications (pp. 373-397). Cham: Springer International Publishing.
88. Samantaray, S., Mohapatra, S. K., & Mohapatra, S. (2025, April). A Systematic Literature Review of Recognizing Handwritten Vedic Text on Palm Leaves Using Machine Learning. In International Conference on Green Artificial Intelligence and Industrial Applications (pp. 146-160). Cham: Springer Nature Switzerland.
89. Sandhan, J., Adideva, O., Komal, D., Behera, L., & Goyal, P. (2021). Evaluating neural word embeddings for Sanskrit. arXiv preprint arXiv:2104.00270.
90. Sanyal, K., Goswami, P. K., & Pathak, N. (2024). Evaluating the Amarkosha to Generate Computational Model for Sanskrit Vocabulary and Sanskrit Word Bank. Journal of Computational Analysis & Applications, 33(7), 1138-1144
91. Schembera, B. (2021). Like a rainbow in the dark: metadata annotation for HPC applications in the age of dark data. Journal of Supercomputing, 77(8).
92. Scheuerman, M. K., Hanna, A., & Denton, R. (2021). Do datasets have politics? Disciplinary values in computer vision dataset development. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), 1-37.
93. Sethi, N., Dev, A., & Bansal, P. (2022, December). A bilingual machine transliteration system for Sanskrit-English using rule-based approach. In 2022 4th International Conference on Artificial Intelligence and Speech Technology (AIST), Delhi, India, 2022, pp. 1-5, doi: 10.1109/AIST55798.2022.10064993
94. Shankar, B., Mishra, P., Sagnika, S., & Pattanaik, A. (2019). Engaging with an Indian Epic: A Digital Approach. International Journal of Computer Applications, 975, 8887.
95. Sharada, B., Sushma, S. N., &Bharathlal. (2018). Keyword Spotting in Historical Devanāgari Manuscripts by Word Matching. In Data Analytics and Learning: Proceedings of DAL 2018 (pp. 65-76). Singapore: Springer Singapore.

96. Shelke, S. V., Chandwadkar, D. M., Ugale, S. P., &Chothe, R. V. (2025). Discrete wavelet transforms and convolutional neural network-based handwritten Sanskrit character recognition. Indonesian Journal of Electrical Engineering and Computer Science, 38(2), 1367-1375.
97. Singh, B., & Ahuja, N. J. (2019). Mining the treasure of palm leaf manuscripts through information retrieval techniques. Digital Library Perspectives, 35(3-4), 146-156.
98. Singh, S. (2018). Scope of Handwriting Recognition in Indic Scripts. International Journal of Research and Analytical Reviews, 6 (1), 305-307
99. Sinha, S. (2025). Abstractive Text Summarization for Contemporary Sanskrit Prose: Issues and Challenges. arXiv preprint arXiv:2501.01933.
100. Sinha, S., & Jha, G. N. (2020, May). Abstractive text summarization for Sanskrit prose: a study of methods and approaches. In Proceedings of the WILDRE5–5th Workshop on Indian Language Data: Resources and Evaluation (pp. 60-65).
101. Sitender, & Bawa, S. (2022). Sanskrit to universal networking language EnConverter system based on deep learning and context-free grammar. Multimedia Systems, 28(6), 2105-2121.
102. Sitender, Bawa, S., Kumar, M., & Sangeeta. (2023). A comprehensive survey on machine translation for English, Hindi and Sanskrit languages. Journal of Ambient Intelligence and Humanized Computing, 14(4), 3441-3474.
103. Skluzacek, T. J., Wong, R., Li, Z., Chard, R., Chard, K., & Foster, I. (2021, June). A serverless framework for distributed bulk metadata extraction. In Proceedings of the 30th International Symposium on High-Performance Parallel and Distributed Computing (pp. 7-18).
104. Sleimi, A., Sannier, N., Sabetzadeh, M., Briand, L., Ceci, M., & Dann, J. (2021). An automated framework for the extraction of semantic legal metadata from legal texts. Empirical Software Engineering, 26(3), 43.
105. Song, S., Huang, H., & Ruan, T. (2019). Abstractive text summarization using LSTM-CNN based deep learning. Multimedia Tools and Applications, 78(1), 857-875.
106. Srivastava, P., Chauhan, K., Aggarwal, D., Shukla, A., Dhar, J., & Jain, V.P. (2018). Deep Learning Based Unsupervised POS Tagging for Sanskrit. *Proceedings of the 2018* International Conference on Algorithms, Computing and Artificial Intelligence. (pp. 1-6).
107. Subramani, K., &Murugavalli, S. (2019). Recognizing ancient characters from Tamil palm leaf manuscripts using convolution-based deep learning. International Journal of Recent Technology and Engineering, 8(3), 6873-6880.
108. Sudarsan, D., & Sankar, D. (2022). A novel complete denoising solution for old Malayalam palm leaf manuscripts. Pattern Recognition and Image Analysis, 32(1), 187-204.
109. Sudarsan, D., & Sankar, D. (2024). An ensemble neural network model for Malayalam character recognition from palm leaf manuscripts. ACM Transactions on Asian and Low-Resource Language Information Processing.https://doi.org/10.1145/3686311
110. Sujoy, S., Krishna, A., & Goyal, P. (2023, January). Pre-annotation-based approach for development of a Sanskrit named entity recognition dataset. In Proceedings of the Computational Sanskrit & Digital Humanities: Selected papers presented at the 18th World Sanskrit Conference (pp. 59-70).
111. Tapaswi, N. (2024). An efficient part-of-speech tagger rule-based approach of Sanskrit language analysis. International Journal of Information Technology, 16(2), 901-908.https://doi.org/10.1007/s41870-023-01668-y
112. Tapaswi, N. (2025). Shabda sculptor: carving morphological excellence in Sanskrit spellcheck. International Journal of Information Technology, 17(1), 591-597.

113. Therrell, G. (2019). More product, more process: metadata in digital image collections. Digital Library Perspectives, 35(1), 2-14.
114. Thottempudi, S. G. (2021). A visual narrative of ramayana using extractive summarization topic modeling and named entity recognition. In CEUR Workshop Proc. (Vol. 2823, pp. 3-10).
115. Tkaczyk, D. (2017). New methods for metadata extraction from scientific literature. arXiv preprint arXiv:1710.10201.
116. Tkaczyk, D., Szostek, P., Fedoryszak, M., Dendek, P. J., &Bolikowski, Ł. (2015). CERMINE: automatic extraction of structured metadata from scientific literature. International Journal on Document Analysis and Recognition (IJDAR), 18(4), 317-335.
117. Tomar, A., Choudhary, M., &Yerpude, A. (2015). Ancient Indian scripts image pre-processing and dimensionality reduction for feature extraction and classification: a survey. International Journal of Computer Trends and Technology (IJCTT), 21(2), 101-124.
118. Tuarob, S., Bhatia, S., Mitra, P., & Giles, C. L. (2016). AlgorithmSeer: A system for extracting and searching for algorithms in scholarly big data. IEEE Transactions on Big Data, 2(1), 3-17.
119. Valaboju, B., Dwivedi, S., Chincholikar, K., Gopalan, K., & Vidwans, V. (2025, May). A Semi-Automatic Text Recognition Tool for Pre-Colonial Handwritten Manuscripts in Devanāgari Script. In International Conference on Human-Computer Interaction (pp. 152-160). Cham: Springer Nature Switzerland.
120. Vijayarajan, V., Dinakaran, M., Tejaswin, P., & Lohani, M. (2016). A generic framework for ontology-based information retrieval and image retrieval in web data. Human-centric Computing and Information Sciences, 6(1), 18.
121. Vijitha, M., Vrinda, K., Dhruva, G., Sahana, R., & Preethi, P. (2024, March). Segmentation of handwritten Sanskrit words using Image-Processing techniques. In International Conference on Innovations in Cybersecurity and Data Science Proceedings of ICICDS (pp. 13-27). Singapore: Springer Nature Singapore.
122. Wu, J., Killian, J., Yang, H., Williams, K., Choudhury, S. R., Tuarob, S., ... & Giles, C. L. (2015, October). Pdfmef: A multi-entity knowledge extraction framework for scholarly documents and semantic search. In Proceedings of the 8th International Conference on Knowledge Capture (pp. 1-8).
123. Yang, W., Fu, R., Amin, M. B., & Kang, B. (2025). The impact of modern ai in metadata management. Human-Centric Intelligent Systems, 5(3), 323-350.
124. Yeghiazaryan, A., Khechoyan, K., Nalbandyan, G., & Muradyan, S. (2022). Tokengrid: Toward More Efficient Data Extraction From Unstructured Documents. IEEE Access, 10, 39261-39268.
125. Zaman, G., Mahdin, H., Hussain, K., & Rahman, A. (2020). Information extraction from semi and unstructured data sources: A systematic literature review. ICIC Express Letters, 14(6), 593-603.
126. Zhang, Y., Tuo, M., Yin, Q., Qi, L., Wang, X., & Liu, T. (2020). Keywords extraction with deep neural network model. Neurocomputing, 383, 113-121.

Appendix 1: Thematic Classification

1. **Digitization, Preservation and Datasets**

Sl. No.	Titles/ Studies
1.	"Amur, Z. H., Hooi, Y. K., Soomro, G. M., Bhanbhro, H., Karyem, S., & Sohu, N. (2023). Unlocking the potential of keyword extraction: the need for access to high-quality datasets. Applied Sciences, 13(12), 7228."
2.	"Deepthi, C. V. S., & Seenu, A. (2022, December). A Systematic Review on OCRs for Indic Documents & Manuscripts. In 2022 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI) (Vol. 1, pp. 1-4). IEEE."
3.	"Dhruva, G., Kore, V., Vijitha, M., Rao, S., & Preethi, P. (2023, December). Comprehensive dataset building of isolated handwritten Sanskrit characters. In International Conference on Applied Soft Computing and Communication Networks (pp. 489-503). Singapore: Springer Nature Singapore."
4.	"Garg, A., Tiwari, L., Juj, T., Indu, S., & Jayanthi, N. (2021). Language and Era Prediction of Digitized Indian Manuscripts Using Convolutional Neural Networks. In Sentimental Analysis and Deep Learning: Proceedings of ICSADL 2021 (pp. 703-718). Singapore: Springer Singapore."
5.	"Gudadhe, S. R., Bardekar, A. A., & Ranit, A. B. (2024, October). A novel approach using machine learning and NLP for revolutionizing Pali manuscript conservation. In 2024 4th International Conference on Sustainable Expert Systems (ICSES) (pp. 844-848). IEEE."
6.	"Guruprasad, P., & KS Rao, G. (2021). Recognition of Handwritten Nandinagari Palm Leaf Manuscript Text. In Computational Intelligence Methods for Super-Resolution in Image Processing Applications (pp. 177-190). Cham: Springer International Publishing."
7.	"Hameed, P., Koikara, R., & Sharma, C. (2015, September). Segmentation of Ancient and Historical Gilgit Manuscripts. In Proceedings of the Second International Conference on Computer and Communication Technologies: IC3T 2015, Volume 1 (pp. 443-447). New Delhi: Springer India."
8.	"Jaiswal, P., & Singh, A. P. (2022). Conservation of knowledge heritage and national mission for manuscripts in Uttar Pradesh and Kerala."
9.	"Kataria, B., & Jethva, H. B. (2018). Review of Advances in Digital Recognition of Indian Language Manuscripts. International Journal of Scientific Research in Science, Engineering and Technology, 4(1), 1302-1318."
10.	"Kataria, B., & Jethva, H. B. (2019). CNN-bidirectional LSTM based optical character recognition of Sanskrit manuscripts: a comprehensive systematic literature review. Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.(IJSRCSEIT), 5(2), 2456-3307."
11.	"Kataria, D. B., & Jethva, H. B. (2021). Optical Character Recognition of Sanskrit Manuscripts Using Convolution Neural Networks. Webology (ISSN: 1735-188X) Volume, 18."
12.	"Kesiman, M. W. A., Valy, D., Burie, J. C., Paulus, E., Suryani, M., Hadi, S., ... & Ogier, J. M. (2018). Benchmarking of document image analysis tasks for palm leaf manuscripts from southeast asia. Journal of Imaging, 4(2), 43."
13.	"Kim, S., Choi, H., Kim, N., Chung, E., & Lee, J. Y. (2018). Comparative analysis of manuscript management systems for scholarly publishing. Science Editing, 5(2), 124-134."
14.	"Krishnan, S., Kulkarni, A., & Huet, G. (2025). Normalized dataset for Sanskrit word segmentation and morphological parsing. Language Resources and Evaluation, 59(2), 1279-1330."

15.	"Kulkarni, I., Tikkal, S., Chaware, S., Kharate, P., & Pandit, A. (2022, February). Proposed Design to Recognize Ancient Sanskrit Manuscripts with Translation Using Machine Learning. In Proceedings of the International Conference on Innovative Computing & Communication (ICICC)."
16.	"Löffler, F., Wesp, V., König-Ries, B., & Klan, F. (2021). Dataset search in biodiversity research: Do metadata in data repositories reflect scholarly information needs?. PloS one, 16(3), e0246099."
17.	"Maheswari, S. U., Maheswari, P. U., & Aakaash, G. S. (2024). An intelligent character segmentation system coupled with deep learning based recognition for the digitization of ancient Tamil palm leaf manuscripts. Heritage Science, 12(1), 342."
18.	"Mohammed, H., Märgner, V., & Ciotti, G. (2021). Learning-free pattern detection for manuscript research: An efficient approach toward making manuscript images searchable. International Journal on Document Analysis and Recognition (IJDAR), 24(3), 167-179."
19.	"Moudgil, A., Singh, S., & Gautam, V. (2021). An overview of recent trends in OCR systems for manuscripts. Cyber Intelligence and Information Retrieval: Proceedings of CIIR 2021, 525-533."
20.	"Moudgil, A., Singh, S., Gautam, V., Rani, S., & Shah, S. H. (2023). Handwritten devanāgari manuscript characters recognition using capsnet. International Journal of Cognitive Computing in Engineering, 4, 47-54".
21.	"Nair, B. B., & Rani, N. S. (2023). HMPLMD: Handwritten Malayalam palm leaf manuscript dataset. Data in Brief, 47, 108960."
22.	"Narang, S. R., Kumar, M., & Jindal, M. K. (2022, December). Optimization of Character Classes in Devanāgari Ancient Manuscripts and Dataset Generation. In International Conference on Frontiers in Computing and Systems (pp. 59-69). Singapore: Springer Nature Singapore."
23.	"Samantaray, S., Mohapatra, S. K., & Mohapatra, S. (2025, April). A Systematic Literature Review of Recognizing Handwritten Vedic Text on Palm Leaves Using Machine Learning. In International Conference on Green Artificial Intelligence and Industrial Applications (pp. 146-160). Cham: Springer Nature Switzerland."
24.	"Scheuerman, M. K., Hanna, A., & Denton, R. (2021). Do datasets have politics? Disciplinary values in computer vision dataset development. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), 1-37."
25.	"Sharada, B., Sushma, S. N., &Bharathlal. (2018). Keyword Spotting in Historical Devanāgari Manuscripts by Word Matching. In Data Analytics and Learning: Proceedings of DAL 2018 (pp. 65-76). Singapore: Springer Singapore."
26.	"Singh, B., & Ahuja, N. J. (2019). Mining the treasure of palm leaf manuscripts through information retrieval techniques. Digital Library Perspectives, 35(3-4), 146-156."
27.	"Subramani, K., &Murugavalli, S. (2019). Recognizing ancient characters from tamil palm leaf manuscripts using convolution based deep learning. International Journal of Recent Technology and Engineering, 8(3), 6873-6880."
28.	"Sudarsan, D., & Sankar, D. (2022). A novel complete denoising solution for old Malayalam palm leaf manuscripts. Pattern Recognition and Image Analysis, 32(1), 187-204."
29.	"Sudarsan, D., & Sankar, D. (2024). An ensemble neural network model for Malayalam character recognition from palm leaf manuscripts. ACM Transactions on Asian and Low-Resource Language Information Processing."
30.	"Sujoy, S., Krishna, A., & Goyal, P. (2023, January). Pre-annotation based approach for development of a Sanskrit named entity recognition dataset. In Proceedings of the

	Computational Sanskrit & Digital Humanities: Selected papers presented at the 18th World Sanskrit Conference (pp. 59-70).”
31.	“Valaboju, B., Dwivedi, S., Chincholikar, K., Gopalan, K., & Vidwans, V. (2025, May). A Semi-Automatic Text Recognition Tool for Pre-Colonial Handwritten Manuscripts in Devanāgari Script. In International Conference on Human-Computer Interaction (pp. 152-160). Cham: Springer Nature Switzerland.”

2. **OCR and Script Recognition for Indic Scripts**

Sl.No.	Titles/ Studies
1.	“Agarwal, M., Indu, S., & Jayanthi, N. (2023, April). An Approach to the Classification of Ancient Indian Scripts Using the CNN Model. In International Conference on Women Researchers in Electronics and Computing (pp. 367-377). Singapore: Springer Nature Singapore.”
2.	“Anoop, C. S., & Ramakrishnan, A. G. (2019, July). Automatic speech recognition for Sanskrit. In 2019 2nd international conference on intelligent computing, instrumentation and control technologies (ICICICT) (Vol. 1, pp. 1146-1151). IEEE.”
3.	“Chadha, S., Mittal, S., & Singhal, V. (2019). An insight of script text extraction performance using machine learning techniques. International Journal of Innovative Technology and Exploring Engineering, 9(1), 2581-2588.”
4.	“Dhingra, V., & Joshi, M. M. (2022). Rule based approach for compound segmentation and paraphrase generation in Sanskrit. International Journal of Information Technology, 14(6), 3183-3191.”
5.	“Jindal, A., & Ghosh, R. (2024). A semi-self-supervised learning model to recognize handwritten characters in ancient documents in Indian scripts. Neural Computing and Applications, 36(20), 11791-11808.”
6.	“Kore, V., Dhruva, G., Rao, S., Vijitha, M., & Preethi, P. (2025). A systematic framework for Sanskrit character recognition using deep learning. ELCVIA Electronic Letters on Computer Vision and Image Analysis, 24(1), 81-103.”
7.	“Lomte, M. V. M., & Doye, D. D. (2022). Handwritten Vedic Sanskrit text recognition using deep learning. Journal of Algebraic Statistics, 13(3), 2190-2198.”
8.	“Madake, J., Yedle, Y., Shahabade, V., &Bhatlawande, S. (2023, June). Sanskrit OCR System. In International Conference on Advanced Communication and Intelligent Systems (pp. 188-200). Cham: Springer Nature Switzerland.”
9.	“Maheshwari, A., Ajmera, R., &Dharamdasani, D. K. (2023). Handwritten Vedic Sanskrit Text Recognition Using Deep Learning and Convolutional Neural Networks. Auricle Global Society of Education and Research.”
10.	“Muehlberger, G., Seaward, L., Terras, M., Ares Oliveira, S., Bosch, V., Bryan, M., ... &Zagoris, K. (2019). Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. Journal of documentation, 75(5), 954-976.”
11.	“Narang, S. R., Jindal, M. K., Ahuja, S., & Kumar, M. (2020). On the recognition of Devanāgari ancient handwritten characters using SIFT and Gabor features. Soft Computing, 24(22), 17279-17289.”
12.	“Narang, S. R., Kumar, M., & Jindal, M. K. (2021). DeepNetDevanagari: a deep learning model for Devanāgari ancient character recognition. Multimedia Tools and Applications, 80(13), 20671-20686.”
13.	“Narang, S., Jindal, M. K., & Kumar, M. (2019). Devanāgari ancient documents recognition using statistical feature extraction techniques. Sādhanā, 44(6), 141.”
14.	“Patil, N. P., & Ramteke, R. J. (2023). A novel optimized deep learning framework to spot

	keywords and query matching process in Devanāgari scripts. Multimedia Tools and Applications, 82(19), 30177-30199."
15.	"Puri, S., & Singh, S. P. (2019). An efficient Devanāgari character classification in printed and handwritten documents using SVM. Procedia Computer Science, 152, 111-121."
16.	"Shelke, S. V., Chandwadkar, D. M., Ugale, S. P., &Chothe, R. V. (2025). Discrete wavelet transform and convolutional neural network based handwritten Sanskrit character recognition. Indonesian Journal of Electrical Engineering and Computer Science, 38(2), 1367-1375."
17.	"Song, S., Huang, H., & Ruan, T. (2019). Abstractive text summarization using LSTM-CNN based deep learning. Multimedia Tools and Applications, 78(1), 857-875."
18.	"Thottempudi, S. G. (2021). A visual narrative of ramayana using extractive summarization topic modeling and named entity recognition. In CEUR Workshop Proc. (Vol. 2823, pp. 3-10)."
19.	"Tkaczyk, D., Szostek, P., Fedoryszak, M., Dendek, P. J., &Bolikowski, Ł. (2015). CERMINE: automatic extraction of structured metadata from scientific literature. International Journal on Document Analysis and Recognition (IJDAR), 18(4), 317-335."
20.	"Tomar, A., Choudhary, M., &Yerpude, A. (2015). Ancient Indian scripts image pre-processing and dimensionality reduction for feature extraction and classification: a survey. International Journal of Computer Trends and Technology (IJCTT), 21(2), 101-124."
21.	"Vijitha, M., Vrinda, K., Dhruva, G., Sahana, R., & Preethi, P. (2024, March). Segmentation of handwritten Sanskrit words using Image-Processing techniques. In International Conference on Innovations in Cybersecurity and Data Science Proceedings of ICICDS (pp. 13-27). Singapore: Springer Nature Singapore."

3. **Metadata Extraction and Information Extraction**

Sl.No.	Titles/ Studies
1.	"Adnan, K., & Akbar, R. (2019). An analytical study of information extraction from unstructured and multidimensional big data. Journal of Big Data, 6(1), 1-38."
2.	"Adnan, K., & Akbar, R. (2019). Limitations of information extraction methods and techniques for heterogeneous unstructured big data. International Journal of Engineering Business Management, 11, 1847979019890771."
3.	"Ahmed, M. W., & Afzal, M. T. (2020). FLAG-PDFe: Features oriented metadata extraction framework for scientific publications. IEEE Access, 8, 99458-99469."
4.	"Al-Barhamtoshy, H., & Eassa, F. (2015). A data analytic framework for unstructured text. Life Science Journal."
5.	"Alghamdi, H., Dawwas, W., Almutairi, T. H., & Rahman, A. (2022). Extracting ToC and Metadata from PDF Books: A Rule-Based Approach. ICIC Express Letters, Part B: Applications, 13(2), 133-143."
6.	"Azimjonov, J., & Alikhanov, J. (2018). Rule based metadata extraction framework from academic articles. arXiv preprint arXiv:1807.09009."
7.	"Choi, W., Yoon, H. M., Hyun, M. H., Lee, H. J., Seol, J. W., Lee, K. D., ... & Kong, H. (2023). Building an annotated corpus for automatic metadata extraction from multi-lingual journal article references. PloS one, 18(1), e0280637."
8.	"Felix, C., Pandey, A. V., & Bertini, E. (2016). Texttile: An interactive visualization tool for seamless exploratory analysis of structured data and unstructured text. IEEE transactions on visualization and computer graphics, 23(1), 161-170."

9.	"Fulda, J., Brehmer, M., &Munzner, T. (2015). TimeLineCurator: Interactive authoring of visual timelines from unstructured text. IEEE transactions on visualization and computer graphics, 22(1), 300-309."
10.	"King, G., Lam, P., & Roberts, M. E. (2017). Computer-assisted keyword and document set discovery from unstructured text. American Journal of Political Science, 61(4), 971-988."
11.	"Kuźma, M., &Mościcka, A. (2020). Evaluation of metadata describing topographic maps in a National Library. Heritage Science, 8(1)."
12.	"Leipzig, J., Nüst, D., Hoyt, C. T., Ram, K., & Greenberg, J. (2021). The role of metadata in reproducible computational research. Patterns, 2(9)."
13.	"Leung, J. K., Griva, I., & Kennedy, W. G. (2020). Making use of affective features from media content metadata for better movie recommendation making. arXiv preprint arXiv:2007.00636."
14.	"Lydia, E. L., Kannan, S., SumanRajest, S., & Satyanarayana, S. (2020). Correlative study and analysis for hidden patterns in text analytics unstructured data using supervised and unsupervised learning techniques. International Journal of Cloud Computing, 9(2-3), 150-162."
15.	"Mahadevkar, S. V., Patil, S., Kotecha, K., Soong, L. W., & Choudhury, T. (2024). Exploring AI-driven approaches for unstructured document analysis and future horizons. Journal of Big Data, 11(1), 92."
16.	"Malashin, I., Masich, I., Tynchenko, V., Gantimurov, A., Nelyub, V., &Borodulin, A. (2024). Image text extraction and natural language processing of unstructured data from medical reports. Machine Learning and Knowledge Extraction, 6(2), 1361-1377."
17.	"Meng, B., Hou, L., Yang, E., & Li, J. (2018, October). Metadata extraction for scientific papers. In China National Conference on Chinese Computational Linguistics (pp. 111-122). Cham: Springer International Publishing."
18.	"Menon, A., Choi, J., & Tabakovic, H. (2018, July). What you say your strategy is and why it matters: natural language processing of unstructured text. In Academy of management proceedings (Vol. 1, p. 18319). Briarcliff Manor, NY 10510: Academy of Management."
19.	"Mirończuk, M. M. (2020). Information extraction system for transforming unstructured text data in fire reports into structured forms: a Polish case study. Fire technology, 56(2), 545-581."
20.	"Nam, S. T., Shin, S. Y., & Jin, C. Y. (2021). Text Mining and Visualization of Unstructured Data Using Big Data Analytical Tool R. 한국정보통신학회논문지, 25(9), 1199-1205."
21.	"Qayyum, F., & Afzal, M. T. (2019). Identification of important citations by exploiting research articles' metadata and cue-terms from content. Scientometrics, 118(1), 21-43."
22.	"Rahman, A. U., Musleh, D., Nabil, M., Alubaidan, H., Gollapalli, M., Krishnasamy, G., ... & Mahmud, M. (2022). Assessment of Information Extraction Techniques, Models and Systems. Mathematical Modelling of Engineering Problems, 9(3)."
23.	"Saeed, M. Y., Awais, M., Talib, R., & Younas, M. (2020). Unstructured text documents summarization with multi-stage clustering. IEEE Access, 8, 212838-212854."
24.	"Safder, I., Hassan, S. U., Visvizi, A., Noraset, T., Nawaz, R., &Tuarob, S. (2020). Deep learning-based extraction of algorithmic metadata in full-text scholarly documents. Information processing & management, 57(6), 102269."
25.	"Schembera, B. (2021). Like a rainbow in the dark: metadata annotation for HPC applications in the age of dark data. Journal of Supercomputing, 77(8)."

26.	"Skluzacek, T. J., Wong, R., Li, Z., Chard, R., Chard, K., & Foster, I. (2021, June). A serverless framework for distributed bulk metadata extraction. In Proceedings of the 30th International Symposium on High-Performance Parallel and Distributed Computing (pp. 7-18)."
27.	"Sleimi, A., Sannier, N., Sabetzadeh, M., Briand, L., Ceci, M., & Dann, J. (2021). An automated framework for the extraction of semantic legal metadata from legal texts. Empirical Software Engineering, 26(3), 43."
28.	"Therrell, G. (2019). More product, more process: metadata in digital image collections. Digital Library Perspectives, 35(1), 2-14."
29.	"Tkaczyk, D. (2017). New methods for metadata extraction from scientific literature. arXiv preprint arXiv:1710.10201."
30.	"Wu, J., Killian, J., Yang, H., Williams, K., Choudhury, S. R., Tuarob, S., ... & Giles, C. L. (2015, October). Pdfmef: A multi-entity knowledge extraction framework for scholarly documents and semantic search. In Proceedings of the 8th International Conference on Knowledge Capture (pp. 1-8)."
31.	"Yang, W., Fu, R., Amin, M. B., & Kang, B. (2025). The impact of modern ai in metadata management. Human-Centric Intelligent Systems, 5(3), 323-350."
32.	"Yeghiazaryan, A., Khechoyan, K., Nalbandyan, G., & Muradyan, S. (2022). Tokengrid: Toward More Efficient Data Extraction From Unstructured Documents. IEEE Access, 10, 39261-39268."
33.	"Zaman, G., Mahdin, H., Hussain, K., & Rahman, A. (2020). Information extraction from semi and unstructured data sources: A systematic literature review. ICIC Express Letters, 14(6), 593-603."

4. **Sanskrit NLP and Computational Linguistics**

Sl.No.	Titles/ Studies
1.	"Barve, S., Desai, S., & Sardinha, R. (2015, October). Query-based extractive text summarization for Sanskrit. In Proceedings of the 4th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA) 2015 (pp. 559-568). New Delhi: Springer India."
2.	"Bhatnagar, K., Lonka, S., & Kunal, J. (2023). San-BERT: Extractive Summarization for Sanskrit Documents using BERT and it's variants. arXiv preprint arXiv:2304.01894."
3.	"Chand, A., Agarwal, P., & Sharma, S. (2023, January). Real-Time Retrieving Vedic Sanskrit Text into Multi-Lingual Text and Audio for Cultural Tourism Motivation. In 2023 International Conference for Advancement in Technology (ICONAT) (pp. 1-6). IEEE."
4.	"Formanek, M. (2025). Exploring the potential of large language models and generative artificial intelligence (GPT): Applications in Library and Information Science. Journal of Librarianship and Information Science, 57(2), 568-590."
5.	"Glickman, M., & Zhang, Y. (2024). AI and generative AI for research discovery and summarization. arXiv preprint arXiv:2401.06795."
6.	"Gupta, V. K., & Shah, H. R. (2025, February). Summarization of Sanskrit Text: Approaches and Techniques. In 2025 International Conference on Computational, Communication and Information Technology (ICCCIT) (pp. 643-648). IEEE."
7.	"Kabra, D., Gohel, R., Prajapati, S., & Gupta, M. K. (2025). Statistical Analysis of Hindi and Sanskrit Languages. Authorea Preprints."
8.	"Koul, N., & Manvi, S. S. (2021). A proposed model for neural machine translation of Sanskrit into English. International Journal of Information Technology, 13(1), 375-381."

9.	"Kumar, P., Pathania, K., & Raman, B. (2023). Zero-shot learning based cross-lingual sentiment analysis for sanskrit text with insufficient labeled data. Applied Intelligence, 53(9), 10096-10113."
10.	"Kumar, R., Tewari, P., Thakur, R. K., & Kumar, R. (2024). ENGLISH TO SANSKRIT TRANSLATION USING NMT. Available at SSRN 4938136."
11.	"Kumari, S., & Malik, A. (2024). Making Machines Talk In Sanskrit: A systematic exploration Of Text-To-Speech Synthesis For Sanskrit Language. Journal of Computational Analysis & Applications, 33(8)."
12.	"Kumari, S., & Malik, A. (2024). Predicting Stress in Sanskrit Texts: A Deep Learning Approach to Sentiment Analysis. International Journal of Multiphysics, 18(3)."
13.	"Pradeep, A., & Mamidi, R. (2025). Sandarśana: A Survey on Sanskrit Computational Linguistics and Digital Infrastructure for Sanskrit. ACM Computing Surveys, 57(10), 1-38."
14.	"Saini, J. R., & Bafna, P. B. (2020). Measuring the Similarity between the Sanskrit Documents using the Context of the Corpus. International Journal of Advanced Computer Science and Applications, 11(5)."
15.	"Sandhan, J., Adideva, O., Komal, D., Behera, L., & Goyal, P. (2021). Evaluating neural word embeddings for Sanskrit. arXiv preprint arXiv:2104.00270."
16.	"Sanyal, K., Goswami, P. K., & Pathak, N. (2024). Evaluating the Amarkosha to Generate Computational Model for Sanskrit Vocabulary and Sanskrit Word Bank. Journal of Computational Analysis & Applications, 33(7)."
17.	"Sethi, N., Dev, A., & Bansal, P. (2022, December). A bilingual machine transliteration system for Sanskrit-English using rule-based approach. In 2022 4th International Conference on Artificial Intelligence and Speech Technology (AIST) (pp. 1-5). IEEE."
18.	"Sinha, S. (2025). Abstractive Text Summarization for Contemporary Sanskrit Prose: Issues and Challenges. arXiv preprint arXiv:2501.01933."
19.	"Sinha, S., & Jha, G. N. (2020, May). Abstractive text summarization for Sanskrit prose: a study of methods and approaches. In Proceedings of the WILDRE5–5th Workshop on Indian Language Data: Resources and Evaluation (pp. 60-65)."
20.	"Sitender, & Bawa, S. (2022). Sanskrit to universal networking language EnConverter system based on deep learning and context-free grammar. Multimedia Systems, 28(6), 2105-2121."
21.	"Sitender, Bawa, S., Kumar, M., & Sangeeta. (2023). A comprehensive survey on machine translation for English, Hindi and Sanskrit languages. Journal of Ambient Intelligence and Humanized Computing, 14(4), 3441-3474."
22.	"Srivastava, P., Chauhan, K., Aggarwal, D., Shukla, A., Dhar, J., & Jain, V. P. (2018, December). Deep learning based unsupervised POS tagging for Sanskrit. In Proceedings of the 2018 international conference on algorithms, computing and artificial intelligence (pp. 1-6)."
23.	"Tapaswi, N. (2024). An efficient part-of-speech tagger rule-based approach of Sanskrit language analysis. International Journal of Information Technology, 16(2), 901-908."
24.	"Tapaswi, N. (2025). Shabda sculptor: carving morphological excellence in Sanskrit spellcheck. International Journal of Information Technology, 17(1), 591-597."
25.	"Vijayarajan, V., Dinakaran, M., Tejaswin, P., & Lohani, M. (2016). A generic framework for ontology-based information retrieval and image retrieval in web data. Human-centric Computing and Information Sciences, 6(1), 18."

5. **Knowledge Retrieval, Information Systems and Applications**

Sl.No.	Titles/ Studies
1.	"Nigam, A., & Chandra, S. (2022). Digital World of Dharmaśāstric Knowledge Tradition: An Instant Information Retrieval System for Manusmṛiti. GIS: Science Journal, 9(8), 241-249."
2.	"Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalan, K. (2017). Using text mining techniques for extracting information from research articles. In Intelligent natural language processing: Trends and Applications (pp. 373-397). Cham: Springer International Publishing."
3.	"Tuarob, S., Bhatia, S., Mitra, P., & Giles, C. L. (2016). AlgorithmSeer: A system for extracting and searching for algorithms in scholarly big data. IEEE Transactions on Big Data, 2(1), 3-17."

6. **Other Studies in metadata and keyword extraction**

Sl.No.	Titles/ Studies
1.	"Botelle, R., Bhavsar, V., Kadra-Scalzo, G., Mascio, A., Williams, M. V., Roberts, A., ... & Stewart, R. (2022). Can natural language processing models extract and classify instances of interpersonal violence in mental healthcare electronic records: an applied evaluative study. BMJ open, 12(2), e052911."
2.	"Cheong, H., Li, W., Cheung, A., Nogueira, A., & Iorio, F. (2017). Automated extraction of function knowledge from text. Journal of Mechanical Design, 139(11), 111407."
3.	"Jain, V., Bagchi, P., Kharat, A., & Shivani, V. (2025). Extracting Invaluable Insights from Sushruta Samhita Using Natural Language Processing. International Journal of Public Mental Health and Neurosciences, 12(2), 10-14."
4.	"Kakimoto, H., Hayashi, T., Wang, Y., Kawai, Y., & Sumiya, K. (2018). Query keyword extraction from video caption data based on spatio-temporal features. In Proceedings of the International MultiConference of Engineers and Computer Scientists (Vol. 1, pp. 405-408)."
5.	"Khan, M. Q., Shahid, A., Uddin, M. I., Roman, M., Alharbi, A., Alosaimi, W., ... & Alshahrani, S. M. (2022). Impact analysis of keyword extraction using contextual word embedding. PeerJ Computer Science, 8, e967."
6.	"MS, R., Mallikarjuna, C., & VS, A. (2020). NLP-Driven Knowledge Extraction and Thematic Classification of Translated Ancient Indian Medical Texts. Rajeevan, MS, Mini devi, B., Anoop, VS, & Mallikarjuna, C.(2025). NLP-driven Knowledge Extraction and Thematic Classification of Translated Ancient Indian Medical Text. Reimagining LIS Education: Collaborative Integration of Indian Knowledge System with NEP, 1, 351."
7.	"Müngen, A. A., & Kaya, M. (2018). Extracting abstract and keywords from context for academic articles. Social Network Analysis and Mining, 8(1), 45."
8.	"Nigam, A., & Chandra, S. (2022, June). Digital Accessibility and Information Mining of Dharmaśāstric Knowledge Traditions. In Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference (pp. 42-47)."
9.	"Razack, H. I. A., Mathew, S. T., Saad, F. F. A., & Alqahtani, S. A. (2021). Artificial intelligence-assisted tools for redefining the communication landscape of the scholarly world. Science editing, 8(2), 134-144."
10.	"Zhang, Y., Tuo, M., Yin, Q., Qi, L., Wang, X., & Liu, T. (2020). Keywords extraction with deep neural network model. Neurocomputing, 383, 113-121."