AI-Driven Digitization of Marathi Indian Knowledge Systems: A Conceptual Framework for Sustainable and Holistic Development

Ujwala Mahajan
Centre for Internal Quality Assurance (CIQA),
Yashwantrao Chavan Maharashtra Open University,
Nashik, Maharashtra, India

Mohit Mahajan
Modern College of Engineering, Pune, Maharashtra,
India

ABSTRACT
Indian Knowledge Systems (IKS), which integrate philosophy, science, ecology, and holistic well-being, comprise an extensive collection of traditional wisdom developed over centuries in the Indian subcontinent. A significant portion of this knowledge is preserved in unstructured textual formats, such as manuscripts, traditional literature, and oral documentation, and is found in regional languages like Marathi. Much of this knowledge is still unavailable in the digital age due to a lack of systematic digitization and computational processing. Computational frameworks that can organize, extract, and interpret original knowledge are therefore becoming increasingly necessary.

This paper presents a conceptual AI-driven framework for digitizing and organizing Indian Knowledge Systems found in Marathi. The framework includes text preprocessing, morphological analysis, Named Entity Recognition (NER), Multiword Expression (MWE) detection, semantic concept extraction, and thematic clustering. These NLP techniques enable automatic identification and arrangement of knowledge about sustainability, ethics, ecology, and community-based development.

The suggested model places a strong emphasis on combining modern AI technologies with knowledge structures that originate in culture. The framework offers a scalable method for creating digital archives of traditional knowledge by permitting semantic interpretation and knowledge representation of Marathi texts. Initiatives for sustainable development, education, research, and cultural preservation can all benefit from such repositories. By offering a conceptual architecture for maintaining and restoring traditional knowledge systems through computational techniques, the study adds to the growing integration of artificial intelligence and original knowledge digitization.

General Terms
Natural language Processing, Indian knowledge System et. al.

Keywords
Indian Knowledge Systems, Artificial Intelligence, Marathi NLP, Multiword Expression Named Entity Recognition, Knowledge Digitization

1. INTRODUCTION
Indian Knowledge Systems (IKS) constitute a broad and diverse intellectual heritage developed in the Indian subcontinent over thousands of years. Philosophy, medicine, ecology, astronomy,

mathematics, agriculture, ethics, and governance are among the knowledge domains examined by these systems. Indian Knowledge recognizes the interdependence of human life, nature, and society, in contrast to contemporary disciplinary knowledge structures. This complete viewpoint encourages ethical behavior, sustainable living, and environmental harmony.

Sanskrit, Marathi, Tamil, and other classical and regional languages have historically preserved a large amount of Indian

knowledge. These languages are still used in a large number of manuscripts, traditional texts, commentaries, and oral traditions.

Nevertheless, a large portion of this information is still unorganized and unavailable in digital information systems. The accessibility, analysis, and distribution of such knowledge are restricted by the lack of systematic digitization and computational processing.

Recent advances in Natural Language Processing (NLP) and artificial intelligence have made it possible to digitize and analyze vast amounts of textual data. By carrying out operations like tokenization, morphological analysis, named entity recognition, semantic extraction, and topic modeling, NLP techniques allow machines to process human language. The interpretation and preservation of traditional knowledge can be greatly aided by these technologies.

Even though NLP research has advanced very far, many Indian languages still don't have enough computational resources. Marathi, a major language in India, continues to encounter difficulties stemming from insufficient annotated corpora, linguistic resources, and specialized models. To fill this gap, recent research projects have tried to create datasets and transformer- based language models for Marathi NLP tasks. The MahaCorpus and MahaBERT language models, for example, offer large Marathi corpora and pretrained models that can be used for tasks like sentiment analysis and named entity recognition.

Similarly, the development of the L3Cube-MahaNER dataset has contributed to improving named entity recognition systems for Marathi, demonstrating the growing potential of NLP research for regional languages.

Most current research, on the other hand, is more focused on general NLP tasks like text classification or entity recognition

than on digitizing and organizing traditional systems of knowledge. There is still a need for conceptual frameworks that combine AI methods with the special features of traditional knowledge literature.

This paper proposes a conceptual AI-driven framework for digitizing Indian Knowledge Systems documented in Marathi language resources. The framework combines many NLP parts so that traditional knowledge texts can be semantically extracted, organized, and classified by theme. The proposed framework seeks to facilitate sustainable development and full understanding preservation by combining current computational methods with culturally embedded knowledge frameworks.

2. INDIAN KNOWLEDGE SYSTEMS AND SUSTAINABLE DEVELOPMENT

Indian Knowledge Systems represent a comprehensive approach to understanding human life and its relationship with nature and society. These systems have evolved through centuries of philosophical reflection, empirical observation, and practical experience.

The fundamental principle underlying Indian Knowledge Systems is the idea of interconnectedness between individuals, communities, and the natural environment. Concepts such as dharma (ethical conduct), loka-sangraha (social welfare), and ecological balance emphasize sustainable and responsible living.

Traditional Indian knowledge includes multiple domains, such as:

- Ayurveda and traditional medicine
- Agricultural knowledge and ecological practices
- Yoga and holistic health systems
- Environmental ethics and biodiversity preservation
- Community-based governance systems

These knowledge traditions provide valuable insights into sustainable development and environmental conservation.

In the context of modern technological development, there is increasing recognition of the importance of integrating indigenous knowledge with contemporary scientific research. Digitizing Indian Knowledge Systems can support interdisciplinary research and contribute to sustainable innovation.

Recent initiatives in India have begun using artificial intelligence technologies to digitize and analyze historical manuscripts and traditional texts. For example, several projects aim to transform scanned manuscripts into searchable digital knowledge repositories through AI-based transcription and analysis systems.

Such initiatives demonstrate the potential of AI technologies for preserving cultural heritage and enabling new forms of knowledge discovery.

3. LITERATURE REVIEW

The literature relevant to this study can be broadly categorized into three major domains: digitization of indigenous knowledge systems, artificial intelligence techniques for knowledge extraction, and Natural Language Processing (NLP) research for Indian and low-resource languages. These domains collectively provide the conceptual and technological foundation for developing computational frameworks for digitizing Indian Knowledge Systems.

3.1 Digitization of Indigenous Knowledge

The digitization of cultural and indigenous knowledge has become an important research area in the fields of digital humanities, information science, and knowledge management. Traditional knowledge systems, often preserved in manuscripts, oral traditions, and regional literature, represent valuable intellectual heritage that requires systematic preservation and accessibility.

Several global initiatives have focused on digitizing cultural heritage through the creation of digital archives and knowledge repositories. Projects such as digital manuscript libraries and cultural heritage digitization programs aim to convert physical documents into digital formats for long-term preservation and wider dissemination [26]. Similarly, UNESCO has emphasized the importance of digitizing indigenous knowledge to ensure its protection and transmission across generations [27].

In India, various initiatives have been undertaken to digitize traditional knowledge resources, including projects related to Ayurvedic texts, classical literature, and archival manuscripts. These initiatives primarily focus on scanning, cataloguing, and storing documents in digital repositories [3].

However, many digitization efforts remain limited to image-based archiving and document storage without enabling deeper semantic understanding of the content. As a result, the digitized materials often lack structured metadata or knowledge representation mechanisms that allow for advanced search, analysis, and knowledge discovery [10]. This limitation highlights the need for computational frameworks that go beyond simple digitization and incorporate artificial intelligence techniques for knowledge extraction and semantic interpretation.

3.2 Artificial Intelligence in Knowledge Extraction

Artificial intelligence has significantly transformed the way knowledge is extracted and organized from large volumes of unstructured textual data. AI-driven approaches, particularly those based on Natural Language Processing, enable automated processing of textual information and facilitate the extraction of meaningful insights.

Key NLP techniques used in knowledge extraction include Named Entity Recognition (NER), relation extraction, topic modeling, and knowledge graph construction. These techniques allow systems to identify entities, detect relationships between concepts, and organize information into structured knowledge representations [13].

Machine learning and deep learning models have further enhanced the capabilities of knowledge extraction systems. Neural architectures, including transformer-based models, have demonstrated strong performance in tasks such as entity recognition, semantic classification, and information retrieval [8].

Knowledge graphs and ontologies have also emerged as powerful tools for representing structured knowledge derived from textual sources. Ontology-based systems enable semantic representation of domain knowledge and support advanced reasoning and query mechanisms [10].

Applications of AI-based knowledge extraction have been successfully implemented in various domains, including biomedical research, legal document analysis, scientific literature mining, and digital libraries [1]. These applications demonstrate the potential of AI techniques to transform unstructured textual information into structured knowledge resources.

Despite these advances, the application of AI-driven knowledge extraction in the context of traditional knowledge systems and indigenous literature remains relatively unexplored, particularly for regional languages.

3.3 Natural Language Processing for Indian Languages

Natural Language Processing research for Indian languages has experienced significant growth in recent years, driven by advances in deep learning and the development of multilingual language models. However, many Indian languages are still considered low-resource languages due to the limited availability of annotated datasets and linguistic resources [12].

The emergence of transformer-based architectures has improved performance in various NLP tasks such as machine translation, sentiment analysis, and named entity recognition. Models such as BERT and its multilingual variants have been widely adopted for processing multiple languages, including Indian languages [8].

For Marathi language processing, recent research has contributed to the development of large-scale corpora and pretrained language models. The MahaCorpus dataset and MahaBERT model introduced by Joshi et al. (2022) provide one of the largest Marathi text corpora and pretrained transformer models for downstream NLP tasks [11].

Named Entity Recognition has also been explored for Marathi using both traditional machine learning approaches and deep learning techniques. The L3Cube-MahaNER dataset developed by Patil et al. (2022) provides annotated data for training and evaluating NER models for Marathi text [20].

Recent studies have further investigated hybrid approaches combining deep learning architectures with linguistic features to improve performance in low-resource languages [6]. These approaches highlight the importance of leveraging both linguistic knowledge and machine learning techniques when processing morphologically rich languages such as Marathi.

Despite these developments, relatively little research has focused on applying NLP techniques to the digitization and semantic structuring of Indian Knowledge Systems. Most existing studies address general language processing tasks rather than domain-specific knowledge extraction from traditional literature. This gap highlights the need for integrated AI-driven frameworks capable of processing indigenous knowledge texts and organizing them into structured digital knowledge repositories.

4. RESEARCH GAP

The majority of current research concentrates on general tasks like sentiment analysis, text classification, or machine translation, regardless of the fact that a number of studies have investigated NLP techniques for Indian languages. The issue of extracting structured knowledge from original knowledge systems or traditional literature is rarely studied.

Furthermore, document preservation is frequently given precedence over semantic knowledge extraction in current digitization initiatives. Traditional knowledge texts are, therefore, still mostly unstructured and challenging to computationally analyze.

The absence of frameworks created especially for regional languages like Marathi is another important gap. The majority of the literature on Indian Knowledge Systems is written in regional languages, so it is crucial to create computational methods that are adapted to their linguistic features.

In order to efficiently digitize and organize Marathi knowledge resources, an integrated AI-based framework that integrates linguistic processing, semantic analysis, and knowledge organization is required.

5. PROPOSED CONCEPTUAL AI FRAMEWORK

In order to digitize Indian Knowledge Systems recorded in Marathi language resources, this study suggests a conceptual Artificial Intelligence (AI) framework. The process converts unstructured textual knowledge into structured digital knowledge repositories using a multi-layered processing architecture. To extract and arrange significant knowledge components from conventional texts, the framework combines methods from Natural Language Processing (NLP), linguistic analysis, and knowledge representation.

Starting with the gathering of knowledge sources and concluding with the structured representation and application of knowledge, the suggested methodology involves five consecutive steps. Figure 1 presents the framework's overall workflow.

5.1 Indian Knowledge Data Sources

Finding and gathering textual materials that contain Indian Knowledge Systems is the first step in the methodology. The main input for the suggested framework is these resources.

Manuscripts, printed books, archival documents, and cultural literature written in regional languages like Marathi are just a few of the ways that Indian knowledge systems are conserved. Philosophy, ethics, environmental practices, traditional medicine, agriculture, and social governance are all topics that are frequently covered in these texts.

The data sources for the proposed framework include:

- Traditional manuscripts and archival records
- Books and scholarly publications related to Indian Knowledge Systems
- Cultural and philosophical literature in Marathi
- Historical documents and digitized archives

Since much of this knowledge exists in unstructured textual formats, computational techniques are required to convert the content into machine-readable and analyzable forms.

5.2 Preprocessing and Digitization

The methodology's second step focuses on transforming unstructured textual materials into digital text that is ready for computational analysis. Preprocessing and digitization are part of this step.

Optical Character Recognition (OCR) is used to transform text

images into a machine-readable digital format when the knowledge sources are physical documents or scanned images. When it comes to converting printed materials and historical manuscripts into searchable text, OCR technology is crucial.

Preprocessing procedures are used to enhance data quality and guarantee consistency after the text has been digitized. These preprocessing operations include:

- **Text cleaning:** removal of unwanted characters, formatting symbols, and noise introduced during scanning.
- **Normalization:** standardization of spelling, punctuation, and character encoding.
- **Sentence segmentation:** dividing the text into individual sentences for further analysis.

These preprocessing steps ensure that the text is properly structured and suitable for subsequent linguistic and semantic processing.

5.3 Linguistic Processing

The third stage of the framework performs linguistic analysis of the processed text. Linguistic processing is particularly important for the Marathi language due to its rich morphological structure and complex grammatical patterns**.**

This stage includes several fundamental NLP operations:

Tokenization

Tokenization involves breaking the text into smaller units, such as words or tokens. These tokens form the basic units used for further linguistic analysis.

Part-of-Speech Tagging

Part-of-speech (POS) tagging assigns grammatical categories to each token, such as nouns (NN), verbs (VB), adjectives (ADJ), and adverbs (ADV). This process helps identify syntactic relationships between words within a sentence.

Multiword Expression Tagging

Multiword expressions (MWEs) are combinations of words that function as a single semantic unit. In Marathi texts related to Indian Knowledge Systems, many important concepts appear in the form of multiword expressions. Detecting these expressions is essential for capturing domain-specific knowledge accurately.

Morphological Analysis

Morphological analysis examines the internal structure of words to identify roots, prefixes, and suffixes. Since Marathi is a morphologically rich language, this step helps normalize variations of words and improves the accuracy of semantic interpretation.

Together, these linguistic processing steps prepare the text for deeper semantic analysis.

5.4 Natural Language Processing Modules

After linguistic processing, the framework applies advanced NLP modules to extract knowledge elements from the text. These modules perform semantic analysis and identify meaningful patterns within the knowledge sources.

Named Entity Recognition (NER)

Named Entity Recognition identifies and classifies entities present in the text. In the context of Indian Knowledge Systems, entities may include:

- Names of scholars or historical figures
- Cultural practices or traditions
- Geographic locations
- Important philosophical concepts

NER helps convert unstructured text into structured information by identifying relevant entities.

Multiword Expression Detection (MWE)

Multiword Expression Detection focuses on identifying compound phrases that represent meaningful knowledge units. Many important ideas in Indian Knowledge Systems are expressed through phrases rather than individual words. Detecting MWEs helps capture these domain-specific expressions accurately.

Semantic Concept Extraction

Semantic concept extraction identifies key knowledge concepts embedded within the text. These concepts may relate to sustainability, ethical principles, traditional practices, or ecological knowledge.

Natural language processing techniques such as semantic similarity analysis and contextual representation models can be used to extract these conceptual relationships.

Thematic Clustering

Thematic clustering groups related knowledge concepts into thematic categories. This process helps organize extracted knowledge into meaningful topics such as:

- Environmental sustainability
- Ethical philosophy
- Traditional medicine
- Community-based governance

Thematic clustering improves the interpretability of the extracted knowledge and supports knowledge discovery.

5.5 Knowledge Representation

The final stage of the framework focuses on organizing the extracted knowledge into structured digital formats**.** The knowledge elements identified during NLP processing are stored in structured representations such as:

- **Knowledge graphs**, which capture relationships between concepts and entities
- **Semantic databases**, which enable structured querying and retrieval of information
- **Digital repositories**, which provide centralized storage for indigenous knowledge resources

These structured representations enable efficient knowledge management and facilitate interdisciplinary research on Indian

Knowledge Systems.The resulting digital knowledge repositories can support various applications, including cultural preservation, educational initiatives, and sustainable development research.

5.6 Applications and Sustainable Insights

The structured knowledge generated by the framework can be used to support several practical applications. These include the development of digital knowledge repositories, AI-driven knowledge discovery systems, and educational resources related to Indian Knowledge Systems.

Furthermore, the extracted knowledge can provide insights into traditional practices related to ecological sustainability, ethical living, and community-centered development. By integrating traditional knowledge with modern computational technologies, the framework contributes to the preservation and regeneration of original knowledge traditions in the digital era.

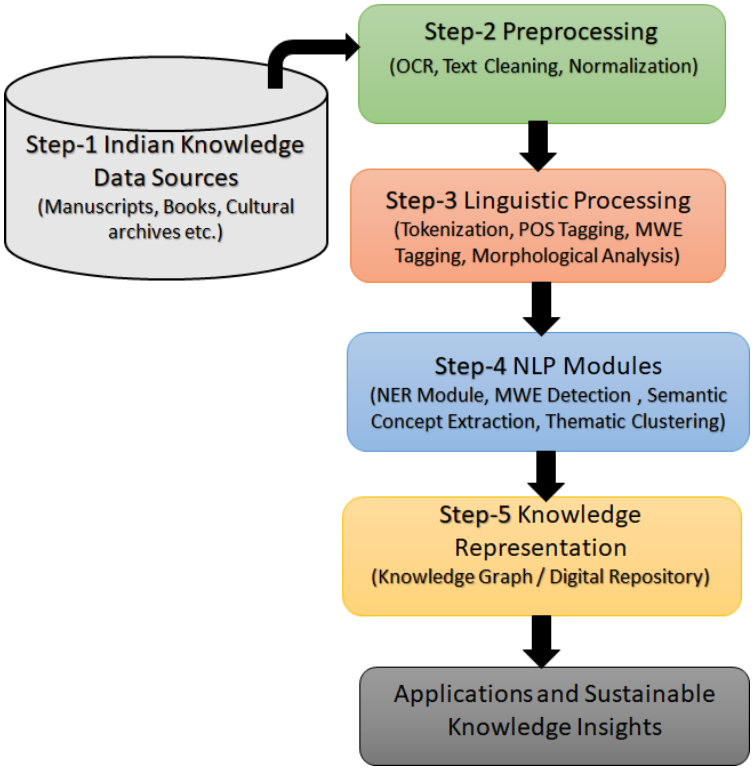

Figure 1: Conceptual AI Framework for Digitizing Indian Knowledge Systems in Marathi

The overall workflow of the proposed AI-based framework for digitizing Indian Knowledge Systems in Marathi is illustrated in Figure 1. The framework demonstrates the sequential processing stages from knowledge source collection to knowledge representation and application.

6. APPLICATIONS AND IMPLICATIONS

The proposed AI-driven framework for digitizing Indian Knowledge Systems (IKS) in Marathi presents several important implications for research, technology development, cultural preservation, and knowledge dissemination. By integrating Natural Language Processing techniques with indigenous knowledge sources, the framework provides a structured approach for transforming unstructured traditional literature into organized digital knowledge resources.

Academic Research

The development of structured digital repositories of Indian Knowledge Systems can significantly support interdisciplinary academic research. Researchers in fields such as philosophy, linguistics, environmental studies, history, and cultural studies can access organized knowledge resources for systematic analysis. The framework also facilitates computational research in digital humanities and indigenous knowledge studies by enabling machine-assisted exploration of traditional texts.

Cultural Preservation

A large portion of Indian Knowledge Systems exists in manuscripts, regional literature, and oral traditions that are vulnerable to loss over time. The proposed framework supports long-term cultural preservation by converting such resources into structured digital formats. Digitized knowledge repositories can help preserve linguistic heritage and ensure that valuable indigenous knowledge remains accessible for future generations.

Technological Innovation

The integration of artificial intelligence and Natural Language Processing techniques with indigenous knowledge resources opens new opportunities for technological innovation. The proposed framework contributes to the development of AI applications tailored for low-resource languages such as Marathi. In addition, structured knowledge extraction from

traditional texts can inspire new research directions in sustainability, traditional medicine, environmental practices, and community-based development models.

Educational Applications

Structured knowledge repositories generated through the proposed framework can support educational initiatives related to Indian Knowledge Systems. Such repositories may assist in the development of curriculum materials, digital learning platforms, and research-based teaching resources. Educational institutions can utilize these resources to promote awareness of indigenous knowledge traditions and their relevance to contemporary societal challenges.

Sustainable Knowledge Integration

Traditional Indian knowledge often emphasizes ecological balance, ethical conduct, and sustainable living practices. By extracting and organizing these knowledge elements through AI-driven methods, the framework can contribute to identifying sustainability-oriented insights embedded in traditional literature. These insights may support modern discussions on sustainable development, environmental ethics, and holistic well-being.

Overall, the proposed framework demonstrates how artificial intelligence can serve as an enabling technology for preserving, organizing, and interpreting traditional knowledge systems in the digital era.

7. CONCLUSION

This study proposed a conceptual Artificial Intelligence framework for the digitization and semantic organization of Indian Knowledge Systems documented in Marathi language resources. The framework integrates multiple Natural Language Processing components, including text preprocessing, morphological analysis, Named Entity Recognition, Multiword Expression detection, semantic concept extraction, and thematic clustering. Through this multi-layered architecture, the framework provides a systematic approach for transforming unstructured traditional texts into structured digital knowledge repositories.

The proposed model highlights the potential of combining modern computational technologies with original knowledge traditions. By enabling automated extraction and organization of knowledge elements, the framework supports the preservation and accessibility of culturally significant information embedded in regional-language literature. The development of structured digital repositories can facilitate interdisciplinary research, support educational initiatives, and contribute to the preservation of cultural heritage.

Furthermore, the framework demonstrates the relevance of Indian Knowledge Systems in contemporary discussions related to sustainability and holistic development. Traditional knowledge sources offer useful details regarding community-based behaviors, ethical living, and sustainability. Digitizing and analyzing such knowledge through AI-based systems can help bridge the gap between traditional knowledge and modern technological innovation.

Future research may focus on implementing the proposed framework using large-scale Marathi textual corpora and evaluating its effectiveness in extracting domain-specific knowledge from traditional literature. Further studies may also explore the development of domain-specific ontologies, knowledge graphs, and multilingual NLP models to enhance the semantic representation and accessibility of Indian Knowledge Systems in digital environments.

8. REFERENCES

[1] Aggarwal, C. C., & Zhai, C. (2012). *Mining text data*. Springer.

[2] Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics, 6*, 587–604.

[3] Bhattacharyya, P. (2015). IndoWordNet: A lexical resource for Indian languages. In *Proceedings of the Global WordNet Conference*.

[4] Bhattacharyya, P. (2017). *Natural language processing for Indian languages*. Springer.

[5] Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'Reilly Media.

[6] Chaudhari, S., Patil, S., & Joshi, A. (2023). Hybrid approaches for named entity recognition in low-resource languages.

[7] *Natural Language Engineering, 29*(2), 325–341.

[8] Chen, D., & Manning, C. D. (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[9] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

[10] Goldberg, Y. (2017). *Neural network methods for natural language processing*. Morgan & Claypool.

[11] Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies, 43*(5–6), 907–928.

[12] Joshi, A., Patil, S., Kale, P., & Bhattacharyya, P. (2022). L3Cube-MahaCorpus and MahaBERT: Marathi monolingual corpus and language model. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

[13] Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the Association for Computational Linguistics (ACL)*.

[14] Jurafsky, D., & Martin, J. H. (2020). *Speech and language processing* (3rd ed.). Stanford University.

[15] Kumar, A., Singh, R., & Sharma, P. (2021). Cross-lingual natural language processing models for low-resource languages.

[16] *Artificial Intelligence Review, 54*(6), 4125–4145.

[17] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*.

[18] Liu, Y., Ott, M., Goyal, N., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

[19] Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

[20] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[21] Navigli, R., & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation, and application of a wide-coverage multilingual semantic network. *Artificial Intelligence, 193*, 217–250.

[22] Patil, S., Joshi, A., Kale, P., & Bhattacharyya, P. (2022). L3Cube-MahaNER: Named entity recognition dataset and baseline models for Marathi. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

[23] Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[24] Ruder, S. (2019). Neural transfer learning for natural language processing. *PhD Thesis*, National University of Ireland.

[25] Sarkar, D. (2016). *Text analytics with Python: A practical, real-world approach to gaining actionable insights from your data*. Apress.

[26] Saurav, S., Singh, A., & Bhattacharyya, P. (2020). Word embeddings for Indian languages: A survey. *ACM Computing Surveys, 53*(5), 1–38.

[27] Singh, A., & Kumar, R. (2021). Artificial intelligence in cultural heritage preservation: Opportunities and challenges. *Journal of Cultural Heritage Management and Sustainable Development, 11*(3), 412–428.

[28] Smith, A. (2011). Preservation in the digital age. *Library Trends, 59*(3), 333–352.

[29] UNESCO. (2017). *Safeguarding intangible cultural heritage in the digital era*. UNESCO Publishing.

[30] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[31] Zhang, Y., Jin, R., & Zhou, Z. H. (2010). Understanding the bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics, 1*(1–4), 43–52.