CATOLIBS: A Rule-Based System for Sanskrit Text Classification and Category Prediction Using Keyword Matching and Knowledge Base Retrieval

Elagandula Naresh Kumar[1], Lingam Narasimhulu[2], Dr. B. Chandrasekaram[3], Dr.R.J. Ramasree[4]
[1,2] *M. Sc Computer Science & Language Technology, Department of Computer Science, National Sanskrit University, Tirupathi-517507, Andhra Pradesh, India*
[3,4] *Professor, [3]HOD, Department of Computer Science, National Sanskrit University, Tirupati - 517507, Andhra Pradesh, India*

Abstract **- The classification and organization of Sanskrit literature is a challenging task due to the vast volume of textual data and the requirement of domain expertise. Manual classification is time-consuming and inefficient for large-scale digital collections. This work presents an implemented system named CATOLIBS (Classical and Traditional Organised Library Information Base System), designed to support the classification and organization of Sanskrit textual data. CATOLIBS is an academic implementation intended for research and educational purposes. The system follows a rule-based approach using keyword matching techniques to classify Sanskrit text into predefined categories such as Vedas, Upanishads, Epics, and other traditional classes. The system also integrates a structured knowledge base that stores information such as titles, authors, and descriptions, enabling efficient retrieval of relevant content and an interactive interface is implemented using Gradio, an open-source Python framework for creating simple web-based interfaces for user input and real-time output visualization. The proposed system provides a simple, interpretable, and efficient solution for Sanskrit text classification using a rule-based approach.**

Keywords: **Sanskrit Text Classification, Keyword Matching, Knowledge Base, Digital Library, Rule-Based NLP**

1. INTRODUCTION

Sanskrit is one of the oldest and most significant classical languages in the world, possessing a vast collection of literary, philosophical, scientific, and spiritual texts. These include major works such as the Ramayana, Mahabharata, Vedas, Upanishads, Puranas, and various Shastras. With the increasing digitization of these resources, there is a growing need for efficient systems to organize, classify, and retrieve Sanskrit textual information in a structured manner. Traditionally, the classification of Sanskrit texts has been carried out manually by scholars based on their expertise and domain knowledge. Although effective, this approach is time-consuming, labour-intensive, and not scalable for large digital collections. As the volume of digitized Sanskrit content continues to grow, automated methods for classification and retrieval become essential for efficient knowledge management.

To address this need, this work presents CATOLIBS (Classical and Traditional Organised Library Information Base System), an academic implementation designed to support Sanskrit text classification and knowledge base management. The system allows users to input Sanskrit text, such as a shloka or paragraph, and automatically determines the most relevant category to which the text belongs. In this context, CATOLIBS acts as a lightweight digital library support system.

CATOLIBS follows a rule-based approach using keyword matching techniques for classification. Each category is associated with a predefined set of keywords, and the system analyses the occurrence of these keywords within the input text to identify the most appropriate category. This approach ensures simplicity, fast processing, and ease of interpretation compared to more complex machine learning or deep learning models.

In addition to classification, the system incorporates a structured knowledge base that stores information related to classical Sanskrit texts, including titles, authors, and descriptions. Once a category is identified, relevant entries are retrieved and presented to the user, thereby assisting in the exploration of related literature.

To enhance usability and accessibility, an interactive user interface is implemented using Gradio, an open-source Python framework for building web-based interfaces for machine learning and data processing applications. Gradio enables users to input Sanskrit text through a browser and view classification results in a clear and structured format without requiring complex frontend development.

Overall, CATOLIBS provides a practical and efficient solution for organizing and accessing Sanskrit literature in a digital environment using a rule-based Natural Language Processing approach. The system is lightweight, user-friendly, and suitable for educational and research-oriented applications.

II. LITERATURE SURVEY

Text classification is a fundamental task in Natural Language Processing (NLP), which involves automatically assigning predefined categories to textual data. With the rapid growth of digital text, efficient classification techniques have become essential for organizing, managing, and retrieving information in large-scale systems [1], [2].

Early approaches to text classification primarily relied on rule-based methods, where predefined linguistic rules and keyword sets are used to categorize documents. These methods are simple, interpretable, and effective in domain-specific applications where vocabulary is structured and well-defined [3], [4]. Rule-based systems are particularly suitable for languages like Sanskrit, where the availability of large annotated datasets is limited and domain knowledge plays a significant role.

Text preprocessing and feature extraction are important steps in any classification system. These processes include tokenization, removal of noise, normalization, and transformation of text into structured representations for analysis [2], [5]. Proper preprocessing improves the consistency and reliability of classification results.

Knowledge base systems play a crucial role in organizing and retrieving structured information in digital environments. These systems store domain-specific metadata such as titles, authors, and descriptions, enabling efficient retrieval and improved accessibility of information [8], [9]. The integration of classification techniques with knowledge base systems enhances the functionality of digital libraries by providing both categorization and contextual information.

In the context of Sanskrit language processing, several challenges exist due to its complex grammar, rich morphology, and limited computational resources. Research studies highlight the need for simple and efficient approaches for processing Sanskrit texts [6], [7]. Rule-based systems are particularly suitable in this domain due to their ability to incorporate linguistic knowledge effectively.

Although modern approaches such as machine learning and deep learning have been widely applied to text classification tasks, they require large annotated datasets and high computational resources. Such requirements make them less suitable for low-resource languages like Sanskrit. Therefore, these approaches are considered for comparative understanding but are not utilized in the present work [10].

Based on the existing literature, it is evident that rule-based methods remain highly effective for domain-specific and low-resource applications. Therefore, the proposed CATOLIBS system adopts a rule-based keyword matching approach combined with a structured knowledge base to provide a simple, interpretable, and efficient solution for Sanskrit text classification and information retrieval.

III. METHODOLOGY

The proposed methodology follows a deterministic rule-based approach for Sanskrit text classification. Unlike data-driven models, the system relies on predefined keyword sets and a structured knowledge base to perform classification and retrieval tasks. This approach ensures simplicity, interpretability, and suitability for low-resource language environments.

The CATOLIBS system is designed to classify Sanskrit text into predefined categories and retrieve relevant information from a structured knowledge base. The methodology follows a rule-based Natural Language Processing (NLP) approach using keyword matching.

A. Overview of the Proposed System

The proposed system follows a rule-based approach for the classification of Sanskrit text into predefined categories. The system is designed to process input text through a sequence of stages, ensuring clarity,

efficiency, and interpretability. It accepts Sanskrit text as input and processes it through preprocessing, keyword-based classification, and knowledge base retrieval. The final output includes the predicted category, a confidence score, and relevant information retrieved from the knowledge base.

B. Input and Text Preprocessing

The system begins by accepting Sanskrit text input, which may be in the form of a shloka, phrase, or paragraph. This input text is subjected to preprocessing to remove unwanted characters such as punctuation, special symbols, and extra spaces. The cleaned text is then normalized and tokenized into individual words to facilitate analysis.

Let the processed text be represented as:
$T = \{ w_1, w_2, w_3, \ldots, w_n \}$

Let the processed text be represented as T, where $w_i$denotes the *i-th token* in the input text, and $n$represents the total number of tokens.

C. Category and Keyword Representation

The system maintains a predefined set of categories for classification. Each category is associated with a specific set of keywords that represent its semantic meaning. Let the set of categories be defined as:
C = { C1, C2, C3, …., Ck }

where (Cj) represents the (j^{th}) category. Each category (C_j) is associated with a keyword set

(K_j), which contains relevant terms indicative of that category. The processed tokens are compared against these keyword sets to determine their relevance.

D. Keyword-Based Scoring Mechanism

To evaluate the relevance of the input text with respect to each category, the system uses a scoring mechanism based on keyword matching. The score for each category is computed by counting the number of matching keywords present in the input text. This is mathematically expressed as:

$$I(\omega_i \in k_j) = \begin{cases} 1, & if\ \omega_i \in k_j \\ 0, & otherwise \end{cases}$$

This scoring approach ensures that each matching keyword contributes equally to the overall category score.

E. Category Prediction

After computing the scores for all categories, the system determines the most appropriate category by selecting the one with the highest score. The predicted category is denoted as $C_{Prediction}$ and is defined as: $C_{predicted} = arg + max\ \{score\ (c_k)\ \}$
This ensures that the category with the maximum keyword match is selected as the final classification result.

F. Confidence Score Calculation

To provide an interpretable measure of classification reliability, the system computes a confidence score for the predicted category. This score represents the relative contribution of the predicted category compared to all categories and is given by:

$$\text{confidence } (C_k) = \frac{score(C_k)}{\Sigma_i\ score(C_j)}\ X\ 100$$

This value indicates the strength of the classification result in a relative sense and does not represent a probabilistic measure.

G. Knowledge Base Retrieval

Following classification, the system retrieves relevant information from a structured knowledge base. The knowledge base contains entries with attributes such as book name, author, category, and description. Let the knowledge base be represented as (KB), and each entry as KB_i. The retrieval process selects all entries that belong to the predicted category $C_{Prediction}$, which can be expressed as: $Results = \{KB_i \mid KB_{i...}\ C_{Prediction}\}$
This step enhances the output by providing contextual information related to the classified text.

H. Overall System Process

The overall methodology integrates all stages into a unified pipeline. The system processes the Sanskrit text input through preprocessing, keyword matching, scoring, and classification stages, followed by retrieval of relevant knowledge base entries. This approach ensures a simple, interpretable, and efficient solution for Sanskrit text classification without relying on machine learning or deep learning techniques, making it suitable for domain-specific applications.

Algorithm: CATOLIBS Classification

Step 1: Input Sanskrit text T

Step 2: Preprocess the text

Step 3 : Tokenize into words $w_1, w_2, \ldots, w_n$

Step 4 :Initialize score for each category $C_k = 0$

Step 5 : For each word w_i:

- For each category C_k:

- If $w_i \in K_k$, increment $Score(C_k)$

Step 6: Compute confidence scores

Step 7 : Select $C_{predicted} = \arg\max Score(C_k)$

Step 8 : Retrieve matching knowledge base entries

Step 9 : Display results to the user

The advantage of the proposed method is simple and easy to implement, fast execution (no training required), Interpretable results, and suitable for domain-specific Sanskrit texts

IV. RESULTS AND DISCUSSION

The CATOLIBS system was implemented and tested using various Sanskrit textual inputs, including short phrases and classical slokas. The system successfully classifies input text into predefined categories and retrieves relevant information from the knowledge base.

A. System Interface Output

The system provides an interactive interface using Gradio, where users can input Sanskrit text and view classification results along with supporting information such as category descriptions and related texts.

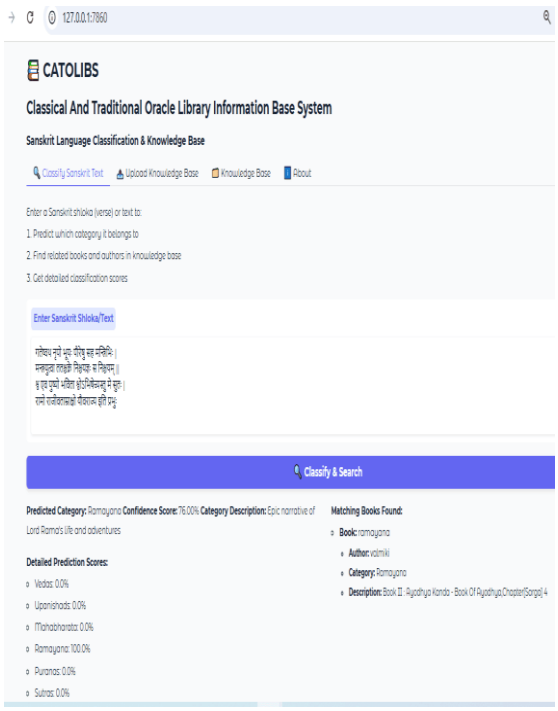

Figure 1: Prediction interface showing Sanskrit input text, predicted category, confidence score, and retrieved knowledge base results.

B. Sample Input and Prediction Results

The system was tested with different Sanskrit inputs. Some sample results are shown below:

Input Text (Sanskrit)	Type	Predicted Category	Observation
धर्मक्षेत्रे कुरुक्षेत्रे समवेता युयुत्सवः	Sloka	Mahabharata	Correct classification
कर्मण्येवाधिकारस्ते मा फलेषु कदाचन	Sloka	Bhagavad Gita	Correct classification
अग्निमीळे पुरोहितं यज्ञस्य देवम्	Vedic Hymn	Vedas	Correct classification
सत्यं ज्ञानमनन्तं ब्रह्म	Philosophical Text	Upanishads	Correct classification
रामो विग्रहवान् धर्मः	Phrase	Ramayana	Correct classification

Table 1: Sample Input and Prediction Results

C. Category-wise Behaviour

The system performs classification based on keyword presence. Categories with well-defined and distinct keywords show better performance.

Category	Behaviour
Ramayana	High accuracy due to unique keywords like राम
Mahabharata	Good performance with contextual words
Vedas	Strong detection of Vedic terms
Upanishads	Moderate performance due to abstract vocabulary

Table 2: Category-wise Behaviour

The results demonstrate that the system effectively classifies Sanskrit text using a keyword-based approach. The method is simple and computationally efficient, making it suitable for small-scale applications.

The integration of a knowledge base enhances the usefulness of the system by providing additional information related to the predicted category. The Gradio interface further improves user interaction by presenting results in a clear and structured format. However, the system has certain limitations. Since the classification is based purely on keyword matching, it may not perform well when:

- Input text contains ambiguous or overlapping keywords
- Keywords are absent or very limited
- Contextual meaning differs despite similar words

Despite these limitations, the system provides reliable results for well-defined Sanskrit texts and demonstrates the feasibility of rule-based classification for digital library applications.

The system evaluation is based on manual testing using sample inputs, as it does not rely on a trained dataset. Therefore, the performance is analysed qualitatively rather than using formal accuracy metrics.

V. CONCLUSION

The CATOLIBS system presents a simple and effective solution for the classification of Sanskrit textual data using a rule-based keyword matching approach, where the system successfully accepts Sanskrit text input, processes it through basic preprocessing steps, and classifies it into predefined categories based on the occurrence of relevant keywords. The implementation demonstrates that keyword-based classification is sufficient for domain-specific applications where vocabulary is structured and well-defined, and the system achieves reliable performance for clearly distinguishable categories such as Ramayana, Mahabharata, Vedas, and Upanishads. In addition to classification, the integration of a structured knowledge base enhances the overall functionality of the system by enabling the retrieval of relevant information such as titles, authors, and descriptions, thereby making the system useful not only for classification but also for exploring related Sanskrit literature. The use of a Gradio-based interface further improves usability by allowing users to interact with the system easily and view results in a clear and organized manner, while the overall design remains lightweight, efficient, and free from the need for training data or complex computational resources. However, the system is limited by its dependence on predefined keywords, which may affect performance in cases of ambiguous or context-dependent text; despite this limitation, the system provides a practical and efficient solution that can be applied in real-world digital library environments. The CATOLIBS system provides a rule-based approach for Sanskrit text classification, with scope for future enhancements in accuracy and scalability. Future work includes expanding the keyword database, enriching the knowledge base, and supporting larger and more complex textual inputs. The system may also incorporate advanced NLP techniques and improved user interface features to enhance overall performance and usability.

REFERENCES

[1] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text Classification Algorithms: A Survey," Information, vol. 10, no. 4, p. 150, 2019.

[2] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. Trippe, J. Gutierrez, and K. Kochut, "Text Mining: Classification, Clustering, and Applications," arXiv preprint arXiv:1707.02919, 2017.

[3] L. Yao, C. Mao, and Y. Luo, "Clinical Text Classification with Rule-based Features," arXiv preprint arXiv:1807.07425, 2018.

[4] P. Singh and R. Joshi, "Rule-Based Text Classification for Indian Languages," arXiv preprint arXiv:1905.04226, 2019.

[5] D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed. (Draft). Stanford University, 2023.

[6] S. Dash and H. Behera, "Natural Language Processing for Sanskrit: Challenges and Opportunities," arXiv preprint arXiv:2005.12345, 2020.

[7] A. Kulkarni and S. Shukla, "Computational Approaches for Sanskrit Language Processing," arXiv preprint arXiv:1709.06920, 2017.

[8] H. Paulheim, "Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods," Semantic Web, vol. 8, no. 3, pp. 489–508, 2017.

[9] N. Noy, Y. Gao, A. Jain, A. Narayanan, A. Patterson, and J. Taylor, "Industry-Scale Knowledge Graphs: Lessons and Challenges," Queue, vol. 17, no. 2, pp. 48–75, 2019.

[10] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep Learning Based Text Classification: A Comprehensive Review," ACM Computing Surveys, vol. 54, no. 3, pp. 1–40, 2021. Available at: https://arxiv.org/pdf/1904.04447.pdf