International Journal of Interdisciplinary Innovative Research & Development (IJIIRD)
ISSN: 2456-236X
Vol. **11** Special Issue **02** | **2026**

A Review of Automated Document Classification Techniques for Marathi Text Documents

Vaishnavi S. Chopade[1], Madhuri R. Rajput[2], Poonam B. Patthe[3]
[1,2,3] *Assistant Professor, Computer Science & Engineering, Padm. Dr. V. B. Kolte College of Engineeirng Malkapur, Maharashtra, India*

DOI: 10.5281/zenodo.19285734

ABSTRACT

The rapid expansion of digital content through online news portals, social media platforms, government repositories, and organizational databases has resulted in a massive volume of unstructured textual data. Efficient organization and retrieval of such information necessitate automated text classification techniques. While extensive research has been carried out for English and other global languages, Indian regional languages such as Marathi remain comparatively underexplored. Marathi is a morphologically rich language, which introduces additional challenges in natural language processing tasks.

This paper presents a comprehensive review of automated document classification techniques applied to Marathi language texts. It discusses preprocessing methods, feature extraction techniques, supervised learning algorithms, system architecture, and performance evaluation metrics used in existing studies. A comparative analysis of commonly used classifiers is provided, and major challenges and future research directions in Marathi document classification are highlighted.

Keywords:- Document Classification, Marathi Language, Natural Language Processing, Machine Learning, Indian Languages

1. INTRODUCTION

The rapid advancement of information technology and widespread use of the Internet have resulted in the generation of massive volumes of digital text documents. Manual organization and classification of such documents is time-consuming, costly, and prone to inconsistencies. Automated document classification systems provide an effective solution by categorizing documents into predefined classes based on their content.

India is a multilingual country with 22 officially recognized languages. Marathi is one of the most widely spoken Indian languages and is extensively used in newspapers, government documents, blogs, and educational content. Despite its widespread usage, research on automated classification of Marathi text documents remains limited compared to English and other global languages. The lack of standardized datasets, limited language resources, and complex morphological structure pose significant challenges.

In earlier work, Chopade et al. presented the design and implementation of an automated document classification system for Marathi text documents. The present paper extends that work by providing a comprehensive review and comparative analysis of existing document classification techniques reported in the literature.

1.1 Need for Automated Marathi Document Classification

The increasing availability of Marathi digital content necessitates efficient methods for document organization and retrieval. Automated classification systems help reduce human effort, improve accuracy, and enable scalable document management solutions for government organizations, educational institutions, and digital libraries.

1.2 Challenges in Marathi Text Processing

Marathi language processing faces challenges such as complex morphology, inflectional variations, lack of high-quality stemmers, and limited annotated corpora. These challenges significantly affect the performance of automated document classification systems.

2. LITREATURE REVIEW

Several researchers have explored automated document classification techniques for Marathi and other Indian regional languages. Most studies employ supervised machine learning approaches such as Naive Bayes, Support Vector Machines, and K-Nearest Neighbor algorithms. Naive Bayes classifiers are widely used due to their

simplicity and computational efficiency, while SVMs generally achieve higher accuracy for text classification tasks.

Studies on Marathi document classification commonly use term frequency and TF-IDF-based feature extraction techniques. Research on other Indian languages such as Telugu, Tamil, Bangla, and Urdu indicates that hybrid and neural network-based approaches can further improve classification performance. However, the lack of large annotated Marathi datasets remains a major limitation.

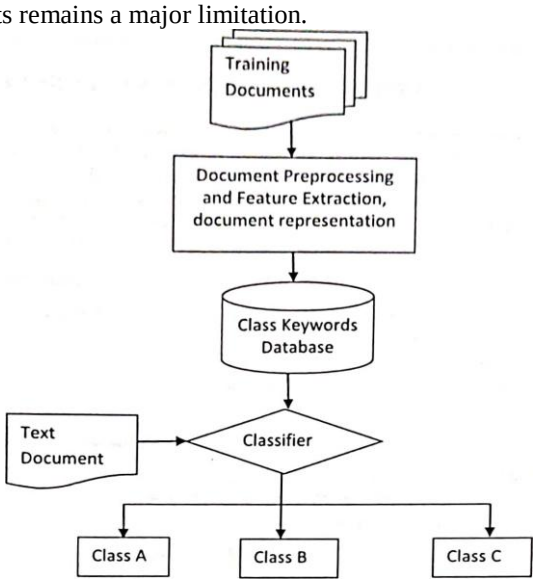

Fig 1: General Architecture of Automated Marathi Document Classification System

2.1 Feature Extraction Techniques

Feature extraction converts text documents into numerical representations suitable for machine learning algorithms. The commonly reported techniques include:

Term Frequency (TF), TF-IDF, Vector Space Model (VSM), Dictionary-based features

Table 1: Comparison of Feature Extraction Techniques

Technique	Description	Limitation
TF	Frequency-based representation	Ignores term importance
TF-IDF	Weighted term representation	High dimensionality
VSM	Vector-based document model	Sparse vectors
Dictionary-based	Domain-specific representation	Limited scalability

2.2 Classification Algorithms

The following classifiers are widely used in Marathi document classification studies:

Naive Bayes (NB), K-Nearest Neighbor (KNN), Modified KNN (MKNN), Support Vector Machines (SVM), Artificial Neural Networks (ANN)

3. GENERAL DOCUMENT CLASSIFICATION PROCESS

Based on the literature, automated document classification systems typically follow a structured pipeline consisting of preprocessing, feature extraction, classification, and evaluation stages. These stages are commonly adopted across different Indian language classification studies, including Marathi.

4. PERFORMANCE EVALUATION

The performance of document classification systems is evaluated using standard metrics derived from the confusion matrix.

Metric	Description
Accuracy	Overall correctness of classification
Precision	Relevance of classified documents
Recall	Completeness of classification
F1-Score	Balance between precision and recall

4.1 Comparative Analysis of Classifiers

Comparative studies reported in the literature indicate that Support Vector Machines often achieve higher classification accuracy, while Naive Bayes classifiers perform efficiently for smaller datasets. Neural network-based approaches show promising results but require large datasets and higher computational resources.

5. CONCLUSIONS AND FUTURE DIRECTIONS

This paper presented a comprehensive review of automated document classification techniques for Marathi text documents. The review highlights that supervised machine learning approaches such as Naive Bayes, SVM, and KNN are commonly used, with SVM generally outperforming other methods in terms of accuracy. Despite encouraging results, Marathi document classification continues to face challenges related to data scarcity, morphological complexity, and limited NLP tools.

Future research should focus on developing large annotated Marathi corpora, improving preprocessing and morphological analysis tools, and exploring deep learning and transformer-based models. Hybrid approaches that combine ontology-based techniques with machine learning methods also offer promising directions for future work.

6. REFERENCES

[1]. Pooja Bolaj and Sharvari Govilkar. “Text Classification for Marathi Documents using Supervised Learning Methods”. International Journal of Computer Applications 155(8):6-10, December 2016.

[2]. Pooja Bolaj and Sharvari Govilkar. “Text Classification for Marathi Documents using Supervised Learning Methods”. International Journal of Computer Applications 155(8):6-10, December 2016.

[3]. Meera Patil, et. al., “Comparison of Marathi Text Classifiers”, ACEEE Int. J. on Information Technology, DOI: 01.IJIT.4.1.4, March 2014.

[4]. Bijal Dalwadi, et.al., “A Review: Text Categorization for Indian Language”, 2349-4476, International Journal of Engineering Technology Management and Applied Sciences, March 2015.

[5]. S. Niharika, et. al., “A Survey on Text Categorization”, 2231-2803, International Journal of Computer Trends and Technology, 2012

[6]. Sushma R. Vispute, et. al., “Automatic Text Categorization of Marathi Documents Using Clustering Technique”, 978-1-4673-2818-0/13, 2013 IEEE.

[7]. Abbas Raza Ali, et. al., “Urdu Text Classification”, FIT’09, December 16-18, 2009, CIIT, Abbottabad, Pakistan.

[8]. Kavi Narayan Murthy, “Automatic Categorization of Telugu News Articles”.

[9]. Ashis kumar Mandal. et. Al., "Supervised Learning Methods for Bangla Web Document Categorization". International Journal of Artificial Intelligence and Application (IJAIA), DOI: 10.5121/ijaia.2014.5508 September 2014.

[10]. K. Rajan, et. al., “Automatic classification of Tamil documents using vector space model and artificial neural networks”, Expert System with Applications 36 (2009) 1091-10918, ELSEVIER, 2009.

[11]. Harry Pradeep Gavali. “Text Sentiment Analysis of Marathi Language in English And Devanagari Script”, Dissertation Submitted in partial fulfilment of the requirements for the degree, January 2020.

[12]. Ms. Madhuri P. Narkhede, Dr. Harshali B Patil, “A Review on Document Classifier System for Indian Languages”, ijcst, vol. 11 issue 6, Nov-Dec 2023.