International Journal of Interdisciplinary Innovative Research & Development (IJIIRD)

ISSN: 2456-236X

Vol. 10 Issue 01 | 2025

An Automated Document Classifier System for Classifying Marathi Text Documents

Vaishnavi S. Chopade[1], Ankush S. Narkhede[2], Madhuri R. Rajput[3], Poonam B. Patthe[4]

[1,2,3,4] *Assistant Professor, Computer Science & Engineering, Padm. Dr. V. B. Kolte College of Engineeirng Malkapur, Maharashtra, India*

DOI: 10.5281/zenodo.15751442

ABSTRACT

With the rapid growth of digital content, a vast amount of textual data is being generated through sources such as news articles, social media interactions, user comments, corporate reports, product reviews, medical records, and tweets. This multilingual data presents significant challenges in terms of organization and knowledge extraction. Manual processing of such large-scale data is both time-consuming and inefficient, thereby emphasizing the need for automated solutions. Text classification systems, particularly those powered by machine learning, play a vital role in efficiently handling and analyzing these documents This paper provides an overview of different document classification systems designed for the Marathi language, emphasizing techniques based on machine learning. It also examines the classification methodology, classification techniques, and performance evaluation measures employed in these approaches.

Keyword: - Document Classifier Systems, Marathi Language

1. INTRODUCTION

The tremendous growth in computers and internet users causes the generation of massive amount of digital documents. India, being a highly diverse nation where diversity is found everywhere, from religions to the cultures and languages the people speaks. India has 22 official languages, and massive amounts of critical textual data are available for these languages. Searching this vast amount of data manually is a time-consuming task. An automated document classifier solves this problem. Automated document classification is useful to improve the performance of information retrieval systems. Efficient document classification is helpful for various Platforms, such as E-commerce, news agencies, content curators, blogs, directories, and user likes.

As Marathi is the official language of Maharashtra, it is predominantly spoken by the people of the state and is therefore used in many important documents. As a result, there is a growing necessity for automated methods to handle Marathi text documents efficiently. While a variety of text classification systems have been developed for many languages, the area of Marathi document classification still requires further research and development. This paper provides an extensive literature review of document classification systems implemented in different languages.

Text classification is the process of assigning tags or categories to the text according to its content. It’s one of the fundamental tasks in Natural Language Processing (NLP) with wide ranging applications including sentiment analysis, topic identification, spam filtering, and intent recognition. Following figure depicts the document classification problem. The text Classification process can include different levels of scope: sub-sentence level, sentence level, paragraph level, and document level.

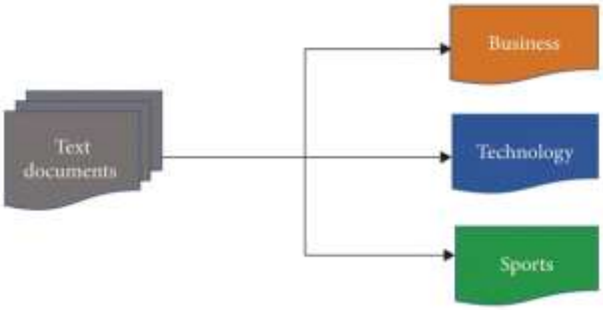

Fig.1: Document classification

2. REVIEW OF LITREATURE

Document classification is one or the important research problem in the field of natural language processing. In recent years, a vast amount of text documents in various Indian languages have become readily available on the internet. For better management and retrieval of such documents, automatic classification can be helpful. The literature reveals numerous document classification systems developed for various languages. This project reviews the documents classifiers systems developed for various Indian languages. The following section presents a comprehensive survey of research related to text classification.

Meera Patil, et.al. [2] proposed an efficient system for classifying Marathi text documents Naive Bayes (NB), Centroid, K-Nearest Neighbor (KNN) and Modified K-Nearest Neighbor (MKNN) classifiers. Then the comparison is done among these four classifiers in terms of accuracy and classification time efficiency. The results indicate that Naïve Bayes (NB) outperforms the other methods in terms of both accuracy and classification time for Marathi document classification, while K-Nearest Neighbor (KNN) demonstrates the lowest accuracy among the four techniques.

Sushma R. Vispute, et al. [5] Developed a smart system for classifying Marathi documents using the LINGO algorithm, which relies on the Vector Space Model (VSM). It also highlights the provision of personalized Marathi language documents to end users, determined by analyzing their browsing history. The comparative analysis reveals that the Vector Space Model (VSM) outperforms other models like the Boolean model and the probabilistic model.

Abbas Raza Ali, et. al. [6] categorized Urdu text documents using statistical methods like Naive Bayes (NB) and Support Vector Machine (SVM). created an intelligent system forThe preprocessing steps include tokenization, normalization, diacritics elimination, stop words elimination and affixes based stemming. The experimental results show that NB classifier is very efficient but has accuracy less than SVM classifier

Kavi Narayana Murthy [7] proposed automatic text classification for Telugu news articles using Naive Bayes (NB) classifier. The four primary categories identified are Politics, Sports, Business, and Cinema. The performance of Naive Bayes (NB) is evaluated using precision, recall, and F-measure. The approach proposed by the author does not incorporate stop word removal, stemming, or morphological analysis.

Ashis Kumar Mandal and Rikta Sen [8] Explained the classification of web documents in the Bangla language using four effective supervised learning algorithms, namely Decision Tree (DT), K-Nearest Neighbor (KNN), Naive Bayes (NB), and Support Vector Machines (SVM). A news corpus collected from various Bangla websites is utilized to assess the performance of these techniques.

The experimental results show that SVM gives better result in terms of accuracy compared to NB. K. Rajan et al [9] presented classification of morphologically rich Dravidian language Tamil documents using Vector Space Model and Artificial Neural Networks. The result analysis shows that artificial neural network model achieves 93.33% on Tamil text documents.

Nadimapalli V Ganapathi have implemented the K-Nearest Neighbour (K-NN) algorithm, which is known to be one of the top performing classifiers applied for the English text. The results show that K-NN is applicable to Telugu text.

ArunaDevi K. and Saveetha R proposed an efficient methodology for retrieving C-feature which is used to classify Tamil text documents. Classification of text documents in Tamil language becomes easy when feature extraction is used. As feature consists of a pair of term to classify document to a predefined category.

From the literature survey it is observed that classification of Marathi documents is unexplored. Therefore, the authors presented efficient Marathi text classifiers which use these supervised learning methods. The input Marathi text documents are preprocessed to remove unwanted data, then feature vector is computed which is supplied to supervised learning methods for classification.

2.1 NEED

In the modern era, the Internet has become a major source of information. The evolution of the Internet has led to the collection of large number of digital documents. Most of the digital data is available in English language and hence majority of data mining and natural language processing research work is focused on English language.

Today, millions of digital documents are available in Indian regional languages. While manual document classification can be extremely detailed and accurate, it has two significant drawbacks making it impractical it also takes a long time and is subjective.

The time required to categorize text increases proportionally with the volume of text to be processed. Consider the amount of content on a corporate intranet, all of a governmental institution's regulations and laws, all of the news items in a newspaper's archive, or even all of the info on the internet that could be useful for a company's business-impossible it's for humans to manage this volume of data in a reasonable amount of time.

This is where the automatic document classification software comes in handy. It allows enterprises and organizations in every sector to organize content and make it available at any time easily. It is scalable, faster and objective.

3. DOCUMENT CLASSIFICATION PROCESS

The document classification process for Marathi texts utilizes supervised learning methods and ontology-based approaches Marathi text documents are provided as input to the system, which then classifies them into appropriate categories based on predefined class labels. The classes considered for implementation includes Festival, Sport, Tourism, Literature, Movies etc.

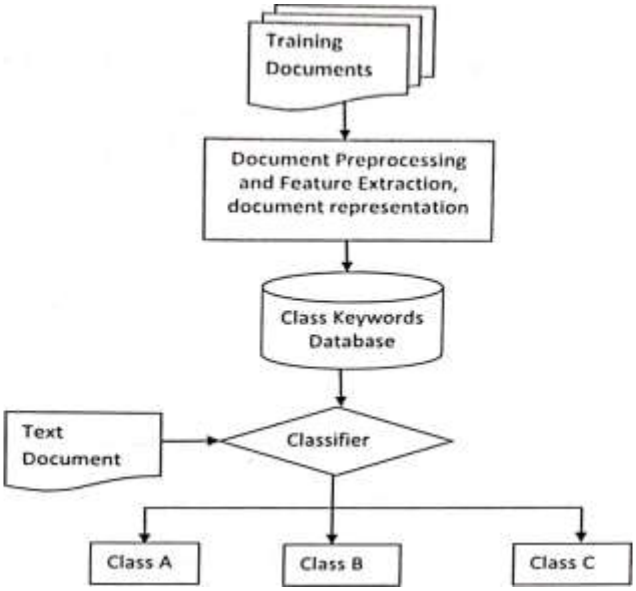

Fig 2: Architecture of proposed Document Classifier system

3.1 METHODOLOGY OF DOCUMENT CLASSIFICATION PROCESS

The system takes a set of Marathi language documents as input and processes them through several preprocessing steps including input validation, tokenization, removal of stop words, stemming, and morphological analysis. Once this preprocessing is complete, key features are identified and extracted from the cleaned tokens. To categorize the documents, the system utilizes a mix of supervised machine learning approaches and ontology-based techniques. The final output consists of Marathi documents classified under specific predefined categories.

The document classification system comprises the following phases:

1. **Preprocessing**

A) **Input Validation**: The first step in the preprocessing phase involves validating the input documents, which are a set of Marathi text documents. The input document may contain words or sentences in a different script or language. This step ensures that the input documents are written in valid Devanagari script. Any words or sentences that do not adhere to the Devanagari script are removed before proceeding with further processing.[1]

B) **Tokenization**: The process of splitting text input into individual tokens is known as tokenization. This task is achieved by identifying spaces between words.[1]

C) **Stop Word Removal:** Stop words are the most commonly occurring words that slow down the document processing, as they are considered irrelevant to the content's meaning. Hence the removal of stop words enhances the speed of searching by comparing with a corpus of stop words.[1]

D) **Stemming**: Stemming is important in the system, which uses a suffix list to remove suffixes from words and thus reduces the word to its stem. The result of stemming is stem of word that can be given as input to morphological analysis for further processing.[1]

2. Feature Extraction:

At this stage, the pre-processor constructs a feature vector for the input document using a Marathi dictionary. This vector contains the key features identified in the document, along with how frequently each one appears.

3. Supervised Learning Methods:

To classify documents written in Marathi, the system applies both supervised machine learning techniques and ontology-based classification approaches. The supervised techniques applied consist of Naïve Bayes (NB), Modified K-Nearest Neighbor (MKNN), and Support Vector Machine (SVM).

4. Output – Categorized Marathi Documents:

In the final step, the system outputs Marathi documents sorted into their respective categories. These categories include topics like Festival, Sports, History, Literature, Tourism, and more. For example, if two Marathi documents are provided as input, after preprocessing and applying the classification techniques, the system might categorize both under the festival label, such as Diwali.

3.2 DOCUMENT CLASSIFICATION TECHNIQUES

To classify documents written in the Marathi language, the system makes use of supervised learning methods along with ontology-based classification approaches. Various document classification techniques have been explored in existing literature, including Expectation Maximization (EM), Naive Bayes (NB) classifier, term frequency–inverse document frequency (tf-idf), instantaneously trained neural networks, latent semantic indexing, Support Vector Machines (SVM), artificial neural networks, K-Nearest Neighbor (KNN), decision tree algorithms like ID3 and C4.5, concept mining, rough set-based classifiers, soft set-based classifiers, multiple-instance learning, and natural language processing techniques.

4. CONCLUSIONS & FUTURE SCOPE

Automatic text classification plays important role in Information Retrieval systemIt plays a crucial role in organizing and retrieving information efficiently. However, limited research has been conducted on Indian regional languages like Marathi. To address this gap, the proposed system focuses on classifying Marathi language documents using supervised learning algorithms and classification techniques. It has been noted that algorithms such as SVM, KNN and its variants, as well as Naive Bayes, are commonly applied for document classification in Indian languages. Specifically for Marathi, substantial work in this area is still lacking.

In the future, it would be valuable to create an automated Marathi document classifier using various techniques, comparing the performance of these classifiers in the Marathi document classification process. This could be tested with a larger corpus and expanded to include additional domains.

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

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

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

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

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

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

[8]. 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.

[9]. 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.

[10]. 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.

[11]. 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.