

Enhanced Question Classification for Marathi: A Comparative Study of Rule-Based and Machine Learning–Based Approaches

Aarti P. Raut ¹, , C. Namrata Mahender ¹

1. Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Chhatrapati Sambhajinagar, IND

Received: June 12, 2026 | Review began: July 20, 2026 | Review ended: July 24, 2026 | Published: August 04, 2026

© Copyright 2026

This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Accurate question classification is a crucial component in developing effective Marathi Question Answering Systems, as it enables improved information retrieval and automated responses. However, Marathi, being a morphologically rich and low-resource language, poses significant challenges for automatic classification. This paper presents a comparative study of rule-based and machine learning approaches for classifying Marathi questions into WH and Non-WH categories. The study evaluates a rule-based classifier based on linguistic patterns and WH keywords, alongside machine learning models using Term Frequency-Inverse Document Frequency-based word- and character-level features. By comparing their performance under the same experimental conditions, the study identifies the most effective approach for Marathi question classification. Multiple classifiers, including Naive Bayes, Decision Tree, Support Vector Machine (SVM), and K-Nearest Neighbors, are employed for classification. Experimental results show that the Decision Tree classifier achieves the highest accuracy of 99.66%, followed closely by SVM with 99.41%, demonstrating the effectiveness of pattern-based learning in Marathi. The rule-based approach also performs strongly, achieving an accuracy of 95.64%, highlighting the importance of lexical features such as WH keywords. Further evaluation using precision and recall indicates that Decision Tree and SVM provide balanced and reliable performance with minimal misclassification. While rule-based methods are effective for structured patterns, they lack generalization for complex sentence variations, which are better handled by machine learning models. Future research will focus on expanding datasets, implementing deep learning architectures, and exploring hybrid models that combine linguistic rules with advanced AI techniques to further enhance Marathi question classification.

Categories: Natural Language Processing (NLP), Machine Learning (ML)

Keywords: question answering (qa), rule based, machine learning, evaluation methodologies, natural language processing

Introduction

Natural Language Processing (NLP) has become a cornerstone of intelligent systems such as chatbots, search engines, virtual assistants, and automated support tools. Within NLP, question classification is a foundational task that determines how questions are categorized based on their linguistic structure and intent [1]. Accurate classification enables Question Answering (QA) systems to choose suitable retrieval and response strategies, leading to more precise and context-aware answers.

While extensive research has been conducted on high-resource languages such as English, French, and Chinese, research on low-resource languages like Marathi remains scarce. Marathi, one of the major Indo-Aryan languages spoken by more than 83 million people in India, lacks sufficient annotated datasets and NLP resources, which restricts the performance of

How to cite this article:

Raut A P, Mahender C (August 04, 2026) Enhanced Question Classification for Marathi: A Comparative Study of Rule-Based and Machine Learning–Based Approaches. Cureus J Comput Sci 3 : es44389-026-00204-w. DOI <https://doi.org/10.7759/s44389-026-00204-w>

QA systems [1]. Consequently, most existing systems fail to accurately interpret Marathi questions, particularly in educational and conversational domains.

This study presents a comparative evaluation of rule-based and machine learning approaches for Marathi question classification. The primary objective is to investigate the effectiveness of these approaches in classifying Marathi questions into WH-type and Non-WH-type categories and to identify the most suitable method for improving Marathi QA systems. The rule-based approach utilizes predefined linguistic rules based on Marathi interrogative words, whereas the machine learning approach employs Term Frequency-Inverse Document Frequency (TF-IDF)-based feature extraction with supervised classifiers. By evaluating both approaches under the same experimental setting, this study provides baseline results and highlights their respective strengths and limitations for Marathi question classification.

Challenges in Marathi question classification

Developing accurate Marathi question classification models raises several unique challenges:

Morphological Complexities

Marathi is a morphologically rich language, which means that words can appear in several different forms depending on their grammatical context, such as gender, number, tense, person, and case. These variations arise because Marathi uses suffixes and inflections to express grammatical relationships within a sentence. As a result, a single root word may generate many different surface forms. This characteristic creates difficulties for traditional systems, since they rely on predefined patterns and rules and may not effectively handle the large number of possible word variations and complex sentence structures.

For example:

"काय आहे?" (What is it?) vs. "काय करतोस?" (What are you doing?).

In both sentences, the word "काय" functions as an interrogative pronoun meaning "what." Although the question word remains the same, the verbs in the sentences differ in their grammatical roles.

"आहे" is a copular verb (a form of the verb "to be"), which is used to express existence or state.

"करतोस" is an action verb derived from the root verb "करणे" (to do) and represents the present tense, second-person singular form.

A rule-based system must therefore recognize that while the interrogative word "काय" remains constant, the verbs "आहे" and "करतोस" carry different grammatical meanings and functions within the sentence. Handling such variations requires complex linguistic rules, which makes rule-based approaches less effective for morphologically rich languages like Marathi.

Lack of Annotated Datasets

A major challenge in Marathi question classification is the lack of large annotated datasets. Unlike English, which has widely used datasets such as TREC and the Stanford Question Answering Dataset, Marathi has very limited publicly available labeled data. This scarcity makes it difficult to effectively train and evaluate machine learning models, thereby limiting the development and performance of Marathi QA systems [2].

Syntactic Variations

Another challenge in Marathi question classification is syntactic variation, where the structure of sentences can change even when the meaning of the question is similar. Marathi generally follows a Subject-Object-Verb sentence order, but in practice, the word order can vary depending on emphasis, context, or style. This flexibility makes it difficult for rule-based systems, which often rely on fixed sentence patterns [3].

For example:

"अर्जुन कुठे आहे?" (Where is Arjun?) vs. "अर्जुन कुठे गेला?" (Where did Arjun go?)

How to cite this article:

Raut A P, Mahender C (August 04, 2026) Enhanced Question Classification for Marathi: A Comparative Study of Rule-Based and Machine Learning-Based Approaches. Cureus J Comput Sci 3 : es44389-026-00204-w. DOI <https://doi.org/10.7759/s44389-026-00204-w>

In both questions, "अर्जुन" is the subject and "कुठे" means "where." However, the verbs differ: "आहे" indicates a state of being, while "गेली" indicates an action in the past tense. Although both questions ask about Arjun's location, their syntactic structures and verb forms change.

WH-Type and Non-WH-Type Distinctions

Questions in Marathi can be broadly divided into two main types:

i) WH-type Questions: These are descriptive questions that begin with interrogative WH-words such as कोण (who), काय (What), कुठे (Where), केव्हा (When), का (Why), and कसे (How). These questions typically seek specific information.

For example:

- कोण आला? - Who came?
- काय घडले? - What happened?
- कुठे जात आहेस? - Where are you going?
- केव्हा सुरू होईल? - When will it start?
- का उशीर झाला? - Why were you late?
- कसे करायचे? - How should it be done?

ii) Non-WH-type Questions: These questions do not start with WH-words and generally expect Yes/No answers, factual responses, or selection from given options.

For example:

- तू शाळेत गेला का? - Did you go to school? (Yes/No question)
- हा भारताचा राष्ट्रध्वज आहे का? - Is this the national flag of India?

Traditional rule-based approaches often struggle to accurately distinguish between these question types because of variations in sentence structure and word forms. Therefore, machine learning models are required to automatically learn patterns and improve the accuracy of Marathi question classification [4].

Need for automatic Marathi question classification

With the growing use of Marathi QA systems, automated question classification is essential for improving response accuracy. In virtual assistants and chatbots, accurate classification helps distinguish between WH-type and Non-WH-type questions, ensuring relevant answers. AI-based educational systems [5] require classification to determine whether a question seeks a definition, explanation, or factual confirmation. Similarly, Marathi search engines rely on efficient classification to differentiate between descriptive and factual queries, which improves search results. In automated customer support, QA systems must classify questions to decide whether a query requires a detailed response or a simple Yes/No answer. Developing an effective Marathi question classification model will therefore enhance the accuracy and usability of Marathi QA systems and improve their performance in real-world applications.

Additionally, the National Education Policy 2020 emphasizes the importance of promoting Indian regional languages in education and digital learning platforms. The policy encourages the development of technological tools and educational resources in native languages to make learning more accessible. In this context, automatic Marathi question classification can support AI-driven learning systems, educational chatbots, and digital learning platforms, enabling students to interact and access knowledge in their native language more effectively. This further highlights the need for robust Marathi NLP solutions to support the goals of the National Education Policy and strengthen language-based educational technologies.

How to cite this article:

Raut A P, Mahender C (August 04, 2026) Enhanced Question Classification for Marathi: A Comparative Study of Rule-Based and Machine Learning-Based Approaches. Cureus J Comput Sci 3 : es44389-026-00204-w. DOI <https://doi.org/10.7759/s44389-026-00204-w>

Materials And Methods

Extensive research has been conducted on question classification in widely spoken languages, such as English and Bengali, but Marathi remains largely unexplored in this area. The complexity of Marathi's inflectional morphology, flexible word order, and agglutinative structure makes automatic classification particularly challenging. Most existing approaches for automatic question classification rely on machine learning models, often trained on large datasets, but the lack of annotated Marathi datasets poses a major challenge. Since direct research on Marathi question classification is limited, it is essential to review studies from other languages to identify key challenges and methodologies that can be adapted.

How to cite this article:

Raut A P, Mahender C (August 04, 2026) Enhanced Question Classification for Marathi: A Comparative Study of Rule-Based and Machine Learning-Based Approaches. Cureus J Comput Sci 3 : es44389-026-00204-w. DOI <https://doi.org/10.7759/s44389-026-00204-w>

Study	Language	Approach	Accuracy/Finding	Limitation
Al Faraby et al. [6]	English	Random Forest embeddings	Macro F1-score rose from 0.57 to 0.70 (Random Forest)	Less accurate for specific question types and not effective beyond textbooks
Hamza et al. [7]	Arabic	Word embedding with machine learning	Achieved 90% accuracy in Arabic question classification, outperforming existing methods	Limited generalization to domain-specific questions and dependent on word representation quality
Mohammed and Omar [8]	English	TFPOS-IDF and word2vec	Achieved 89.7% F1-score, demonstrating effectiveness in multi-domain exam question classification	Performance relies on dataset size and domain variations, limited generalization to new subjects
Yilmaz and Toklu [9]	Turkish	LSTM and CNN	Models achieved 94% accuracy on the question dataset, with word2Vec embeddings	Limited datasets, language-specific challenge
Mohasseb et al. [10]	English	Grammar-based ML	Grammar-based question classification using J48 achieved 90.1% accuracy, but it was grammar-dependent and domain-specific	Relies on grammatical structure, struggles with informal or ungrammatical questions, and lacks cross-domain adaptability
Van-Tu and Anh-Cuong [11]	English	Feature selection algorithm and SVM	Achieved 95.2% (coarse) and 91.6% (fine) accuracy	Limited to English, dependent on feature selection, and ineffective for complex question structures
Banerjee et al. [12]	Bengali	Naive Bayes and Decision Tree classifiers	Combining classifiers improves performance by 4.02% for coarse-grained classes, with stacking reaches 87.79% accuracy in fine-grained classification	Limited to Bengali, lacks generalization to other languages, and has high computational overhead from multiple models
Kamble and Baskar [13]	Marathi	ML translation-based	73.5% (coarse) and 47.5% (fine-grained)	Lower accuracy and limited dataset size

How to cite this article:

Raut A P, Mahender C (August 04, 2026) Enhanced Question Classification for Marathi: A Comparative Study of Rule-Based and Machine Learning-Based Approaches. Cureus J Comput Sci 3 : es44389-026-00204-w. DOI <https://doi.org/10.7759/s44389-026-00204-w>

			accuracy	
Govilkar and Bakal [14]	Marathi	Ontology-based reasoning	Ontology-based QA enhances semantic understanding, improving accuracy in retrieving precise answers in Marathi	Performance depends on ontology completeness, domain coverage, and handling complex queries remains a challenge
Devi and Dua [15]	Hindi	KNN and NN	Smith-Waterman excelled in similarity functions; KNN achieved 97% and 95.6% accuracy, while NN scored 93.3% and 95% on two datasets	Limited to Hindi objective questions, lacks generalization to other languages or formats, and performance varies by dataset

TABLE 1: Review on Languages for Question Classifications

CNN, Convolutional Neural Network; KNN, K-Nearest Neighbors; LSTM, Long Short-Term Memory; ML, Machine Learning; NN, Neural Network; QA, Question Answering; SVM, Support Vector Machine; TFPOS-IDF, Term Frequency–Inverse Document Frequency based on Part-Of-Speech

After reviewing multilingual studies, several important insights can be identified that are relevant to Marathi question classification (Table 7). These studies provide valuable approaches for handling different question types, feature extraction methods, and classification algorithms that can be adapted for Marathi. At the same time, they highlight existing research gaps, such as the need to manage complex question structures, address morphological variations, and develop suitable annotated datasets for Marathi. Analyzing these multilingual approaches, therefore, helps identify suitable techniques and design improved models tailored to Marathi question classification.

Proposed methodology

The Marathi language currently lacks an automatic question classification system. Most existing approaches focus on widely used languages such as English and Hindi, leaving Marathi underexplored in this domain. Given the increasing importance of Marathi-based educational resources, there is a strong need for a dedicated question classification system. To address this gap, this study presents a comparative framework that evaluates rule-based and machine learning approaches for classifying Marathi questions into WH-type and Non-WH-type categories. The rule-based approach utilizes predefined linguistic rules based on Marathi interrogative words, whereas the machine learning approach employs TF-IDF-based feature extraction with supervised classifiers. Both approaches are evaluated under the same experimental conditions to compare their effectiveness and identify the most suitable method for Marathi question classification. The overall workflow of the proposed comparative framework is illustrated in Figure 7 and consists of the following steps:

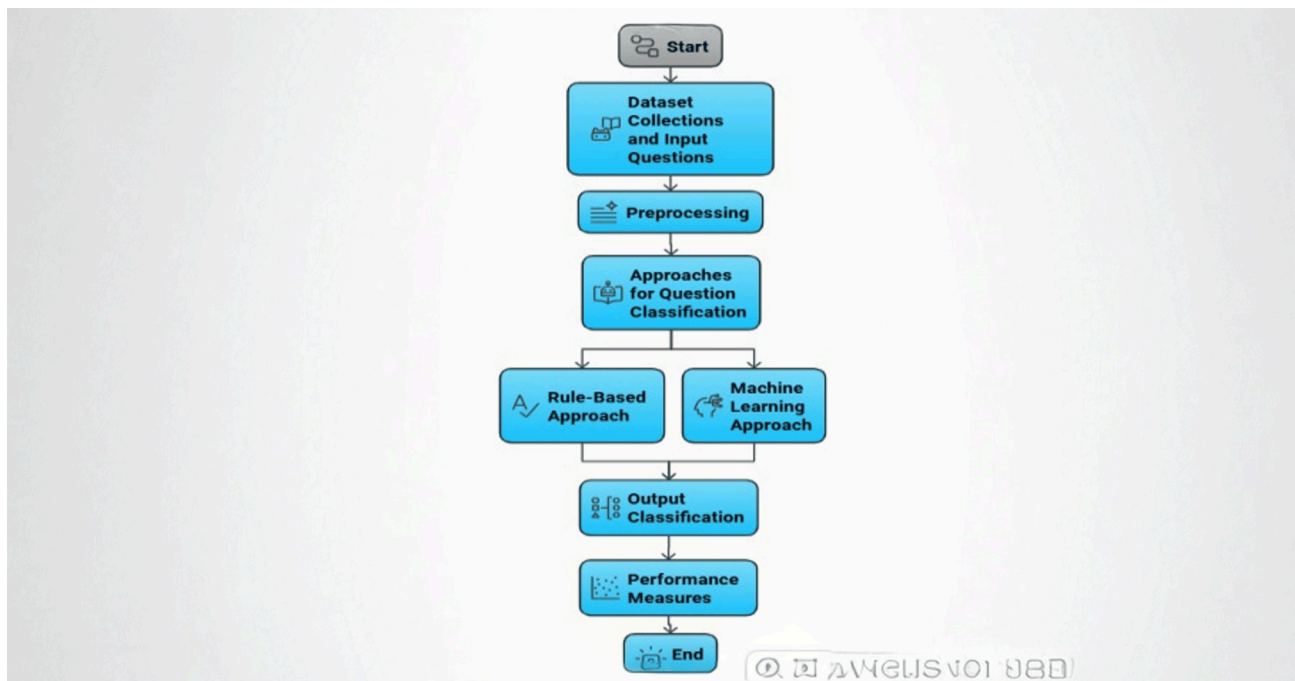


FIGURE 1: Comparative Framework for Marathi Question Classification Using Rule-Based and Machine Learning Approaches

1) Start: The process begins by initiating the Marathi question classification system, which is designed for automatic question identification and categorization.

2) Dataset Collection and Input Questions: The dataset used in this study consists of 1,193 Marathi questions collected from Balbharati Marathi textbooks prescribed for Standards II, III, and IV. These questions were manually extracted and annotated into two categories: WH-type and Non-WH-type, based on their linguistic structure and expected response type. The dataset was selected because Balbharati textbooks are standardized educational resources widely used in Maharashtra and provide grammatically correct and curriculum-aligned question patterns suitable for evaluating Marathi question classification methods.

Although the dataset adequately represents the target educational domain, it is limited to primary-level textbooks and, therefore, does not cover questions from secondary education, higher education, competitive examinations, newspapers, conversational text, or other real-world domains. Consequently, the experimental results should be interpreted within the scope of primary-level educational question classification. Future work will expand the corpus by incorporating more diverse Marathi question sources to improve the robustness and generalizability of the proposed approaches.

Example questions from the dataset include:

- बुधवार पेठ कोठे आहे?
- सावित्रीबाई फुले यांना लोकांनी समर्थन दिले होते का?

These questions are provided as input to both the rule-based and machine learning classifiers for comparative evaluation.

3) Preprocessing: Preprocessing is crucial for preparing the input question for further analysis in a Marathi QA system. Since Marathi is a morphologically rich language, preprocessing helps normalize and structure the text. This step ensures that the input is clean and structured so that the system can accurately interpret and generate appropriate responses.

How to cite this article:

Raut A P, Mahender C (August 04, 2026) Enhanced Question Classification for Marathi: A Comparative Study of Rule-Based and Machine Learning-Based Approaches. Cureus J Comput Sci 3 : es44389-026-00204-w. DOI <https://doi.org/10.7759/s44389-026-00204-w>

The key preprocessing techniques include:

- **Removing Stopwords:** Stopwords are commonly used words that do not add significant meaning to a question, such as "आणि" (and), "हा" (this), etc., in Marathi. Removing these words helps the system focus on the core words that define the intent and meaning of the question.
- **Tokenization:** Tokenization involves breaking down the input question into individual words or tokens. This helps the system analyze each component of the question separately. For example, the question "भारताची राजधानी कोणती आहे?" (What is the capital of India?) would be tokenized into ["भारताची," "राजधानी," "कोणती," "आहे"].
- **Stemming or Lemmatization:** These techniques are used to reduce words to their root forms, improving the system's understanding of their base meanings. Stemming involves removing suffixes (e.g., "चालत आहे" → "चाल" for "walking"). Lemmatization converts words to their dictionary form (e.g., "मोठ्या" → "मोठा" for "big").

By applying these preprocessing techniques, the QA system ensures that the input question is clean, structured, and ready for further analysis. This step is essential for accurate question interpretation and helps improve the overall performance of the system in generating relevant answers.

i) **Feature Extraction and Implementation Details:** After preprocessing, the normalized Marathi questions were converted into numerical feature representations for machine learning classification. The preprocessing stage was performed using the Stanza NLP toolkit, which provides linguistic processing support for the Marathi language. Stanza was used to perform tokenization and lemmatization, while Marathi stop words were removed to retain only meaningful lexical terms. This preprocessing pipeline reduced morphological variations and ensured consistent text representation before feature extraction. The preprocessed questions were subsequently transformed into numerical vectors using the TF-IDF technique. TF-IDF assigns higher weights to informative terms while reducing the influence of frequently occurring words, thereby generating discriminative feature vectors suitable for text classification. These TF-IDF feature vectors served as input to the machine learning classifiers. All machine learning experiments were implemented in Python using the Scikit-learn library. The generated TF-IDF feature vectors were provided as input to the Naive Bayes, Decision Tree, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) classifiers. To ensure a fair comparison, identical preprocessing, feature extraction, and evaluation procedures were applied to all classifiers throughout the experiments.

4) **Approaches for Question Classification:** Question classification is a critical step in a QA system, as it helps categorize questions into predefined types. This categorization enables the system to understand the user's intent and retrieve relevant answers effectively. However, in the context of the Marathi language, question classification poses challenges due to its rich morphology, complex linguistic structures, and limited resources. Based on the preprocessed input, automatic question classification can be performed using the following approaches:

- i) **Rule-Based Approach:** The rule-based approach in a Marathi QA system works by applying predefined linguistic rules to analyze and generate responses. This method relies on identifying question words (प्रश्नवाचक शब्द) such as "कोण" (Who), "काय" (What), "कोठे" (Where), "केव्हा" (When), "का" (Why), and "कसा" (How) to determine the type of question.
- **Applying Rules for Question Classification:** In this step, specific rules are applied to process the identified question type, ensuring that the system accurately interprets the question and selects the most appropriate method for generating an answer (Table 2).

How to cite this article:

Question Type	Rule Category	Processing Steps	Example (Marathi)	System Interpretation and Response
WH-Type Questions	Language-based Rules	- Identify main subject (e.g., topic or entity). - Understand action/context (e.g., what, who, where). - Retrieve relevant information from the knowledge base. - Construct factual or descriptive response.	छत्रपती शिवाजी महाराज कोण होते? (Who was Chhatrapati Shivaji Maharaj?)	Recognized as a WH-type query seeking factual information. The system recognizes "कोण" as a question about a person and retrieves the correct answer: "छत्रपती शिवाजी महाराज हे मराठा साम्राज्याचे संस्थापक होते."
Non-WH-Type Questions	Logical Rules	- Detect Yes/No or numerical question format. - Check conditions or keyword presence. - Verify factual correctness from stored data. - Respond with confirmation or denial.	भारताची राजधानी दिल्ली आहे का? (Is Delhi the capital of India?)	Identified as a Non-WH-type (Yes/No) question. The system validates the fact from its database and responds "हो" (Yes).

TABLE 2: Rules Applied for Marathi Question Classification

This approach ensures structured and accurate responses for well-defined questions. However, it faces challenges with complex or ambiguous queries, where machine learning methods can provide better accuracy.

ii) Machine Learning Approach: The machine learning approach in a Marathi QA system involves training models on large datasets to recognize patterns in questions and generate relevant answers (Table 3). Unlike the rule-based approach, which relies on predefined rules, machine learning models learn from data and improve their accuracy over time.

How to cite this article:

Raut A P, Mahender C (August 04, 2026) Enhanced Question Classification for Marathi: A Comparative Study of Rule-Based and Machine Learning-Based Approaches. Cureus J Comput Sci 3 : es44389-026-00204-w. DOI <https://doi.org/10.7759/s44389-026-00204-w>

Algorithm	Description	Example
Naive Bayes	Naive Bayes is a probabilistic classifier based on Bayes' Theorem. It assumes independence between words in a question, making it fast and efficient for text classification tasks. It performs well even with small datasets and is suitable for categorizing WH and Non-WH questions using word probability distributions.	WH-Type: "भारताची राजधानी काय आहे?" → contains "काय" → classified as WH-type. Non-WH-Type: "तु आज शाळेत गेला का?" → contains "का" → classified as Non-WH-type.
SVM	SVM is a supervised learning model that identifies an optimal decision boundary (hyperplane) to separate different classes. It is effective for high-dimensional text data and provides high classification accuracy. It is particularly useful for distinguishing WH-type and Non-WH-type questions based on learned patterns.	WH-Type: "मुंबई कुठे आहे?" → contains "कुठे" → classified as WH-type. Non-WH-Type: "पृथ्वी सूर्याभोवती फिरते का?" → classified as Non-WH-type. SVM learns from labeled examples to classify the question structure.
Decision Tree	A Decision Tree is a hierarchical model that classifies data using feature-based splitting. It is easy to interpret and useful for rule-based and hybrid classification. The model typically checks the presence of WH-words to determine the class.	WH-Type: "का तू शाळेत गेला नाही?" → contains "का" (Why) → classified as WH-type. Non-WH-Type: "ही व्यक्ती प्रसिद्ध आहे का?" → Yes/No pattern → classified as Non-WH-type. The model follows Decision Tree splitting rules starting at the root node: Does the question contain a WH-word? If yes, it is classified as a WH-type question. If no, it is classified as a Non-WH-type question.
KNN	KNN is a distance-based algorithm that classifies a question based on similarity to previously labeled examples. It is simple and effective for small datasets and groups similar WH-type and Non-WH-type questions using feature similarity.	WH-Type: "समुद्राचा रंग कोणता आहे?" → contains "कोणता" → classified as WH-type. Non-WH-Type: "समुद्राचा रंग निळा का असतो?" → classified as Non-WH-type.

TABLE 3: Machine Learning Algorithm for Marathi Question Classification

SVM, Support Vector Machine

iii) Output Classification: After applying both rule-based and machine learning classification approaches, the system produces the predicted question class. The output indicates whether the input question belongs to WH-type or the Non-WH-type category. This classification helps in retrieving appropriate answers efficiently.

iv) Performance Measures:

How to cite this article:

Raut A P, Mahender C (August 04, 2026) Enhanced Question Classification for Marathi: A Comparative Study of Rule-Based and Machine Learning-Based Approaches. Cureus J Comput Sci 3 : es44389-026-00204-w. DOI <https://doi.org/10.7759/s44389-026-00204-w>

- **Accuracy:** The system performance is evaluated using accuracy. Accuracy is a fundamental metric used to assess the overall performance of the system. It measures the proportion of correctly classified questions out of the total number of questions processed. It provides a general indication of how well the system understands and classifies Marathi queries. The accuracy of the system is calculated using the formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

where, TP (True Positives) represents correctly identified WH questions, TN (True Negatives) represents correctly identified Non-WH questions, FP (False Positives) represents Non-WH questions incorrectly classified as WH, and FN (False Negatives) represents WH questions incorrectly classified as Non-WH.

- **Precision:** Precision measures the correctness of positive predictions. It indicates how many of the questions classified as WH are actually WH questions. The following formula is used to calculate precision:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

- **Recall:** Recall measures the ability of the system to identify all relevant instances. It indicates how many actual WH questions are correctly identified by the system.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

These metrics together provide a detailed evaluation of the system's effectiveness. While accuracy gives an overall performance measure, precision and recall provide deeper insight into classification quality, especially in handling class imbalance and misclassification scenarios.

v) **End:** Finally, the system outputs the predicted question type and completes the classification process. The identified question category helps the Marathi QA system retrieve appropriate answers efficiently.

Results And Discussion

Result and analysis

To ensure reproducibility, all experiments were conducted using identical preprocessing and feature extraction procedures for each classifier. The Marathi question dataset contained 1,193 manually annotated questions, categorized into WH-type and Non-WH-type classes. Before training, the dataset was preprocessed through tokenization, stop-word removal, and lemmatization, followed by TF-IDF feature extraction. The rule-based approach and all machine learning classifiers were evaluated under the same experimental settings to ensure a fair comparison. The machine learning models were implemented using the Stanza library in Python. TF-IDF vectors generated from the preprocessed questions were used as input features for Naive Bayes, Decision Tree, SVM, and KNN classifiers. Performance was measured using accuracy, precision, and recall, as described in Table 4.

How to cite this article:

Approaches	Accuracy (%)	Precision (%)	Recall (%)
Rule-Based	95.64	100	94.92
Naive Bayes	92.11	88.02	100
Decision Tree	99.66	100	99.61
SVM	99.41	99.52	99.80
KNN	90.36	92.15	97.08

TABLE 4: Accuracy, Precision, and Recall of Rule-Based and ML-Based Approaches

KNN, K-Nearest Neighbors; ML, Machine Learning; SVM, Support Vector Machine

The results demonstrate the strengths and limitations of each method in handling WH and Non-WH question types. The rule-based approach achieves an accuracy of 95.64%, with a precision of 100% and a recall of 94.92%, confirming that Marathi question classification relies heavily on lexical patterns. The perfect precision indicates that whenever the system predicts a question as WH, it is always correct (i.e., no FP). However, the slightly lower recall shows that some WH questions are missed, highlighting the limitation of fixed linguistic rules in handling variations and complex sentence structures.

Naive Bayes achieves an accuracy of 92.11%, with a precision of 88.02% and a recall of 100%. The perfect recall indicates that the model successfully identifies all WH questions (no FN), but the lower precision suggests that it incorrectly classifies some Non-WH questions as WH (higher FP). This behavior is attributed to the probabilistic nature of Naive Bayes and its assumption of feature independence, which may not hold for natural language data.

The results clearly indicate that the Decision Tree classifier outperforms all other models. This is primarily due to its ability to effectively capture rule-based patterns present in Marathi questions, such as the occurrence of WH words like "काय," "कोण," and "कुठे." Since the dataset is highly pattern-driven, the Decision Tree learns decision boundaries efficiently, resulting in near-perfect classification performance. It achieves the highest accuracy of 99.66%, along with perfect precision (100%) and very high recall (99.61%). These results demonstrate its effectiveness in modeling underlying linguistic patterns with minimal classification errors, making it the most suitable model for this dataset.

SVM also performs exceptionally well, achieving an accuracy of 99.41%, precision of 99.52%, and recall of 99.80%. These balanced and high values indicate that SVM effectively handles high-dimensional TF-IDF features and generalizes well across different question patterns. The slight difference compared to the Decision Tree suggests that, although SVM is robust, it may not capture explicit rule-based patterns as directly as tree-based models.

KNN achieves an accuracy of 90.36%, with a precision of 92.15% and a recall of 97.08%. The relatively lower accuracy indicates that KNN struggles with high-dimensional sparse data, such as TF-IDF representations. Although the recall is high, indicating that most WH questions are detected, the lower precision suggests the presence of FP. Figure 2 presents a comparison of performance measures, including accuracy, precision, and recall, for both rule-based and machine-learning approaches. The chart clearly illustrates that the Decision Tree and SVM outperform the other models across all evaluation metrics.

How to cite this article:

Raut A P, Mahender C (August 04, 2026) Enhanced Question Classification for Marathi: A Comparative Study of Rule-Based and Machine Learning-Based Approaches. Cureus J Comput Sci 3 : es44389-026-00204-w. DOI <https://doi.org/10.7759/s44389-026-00204-w>

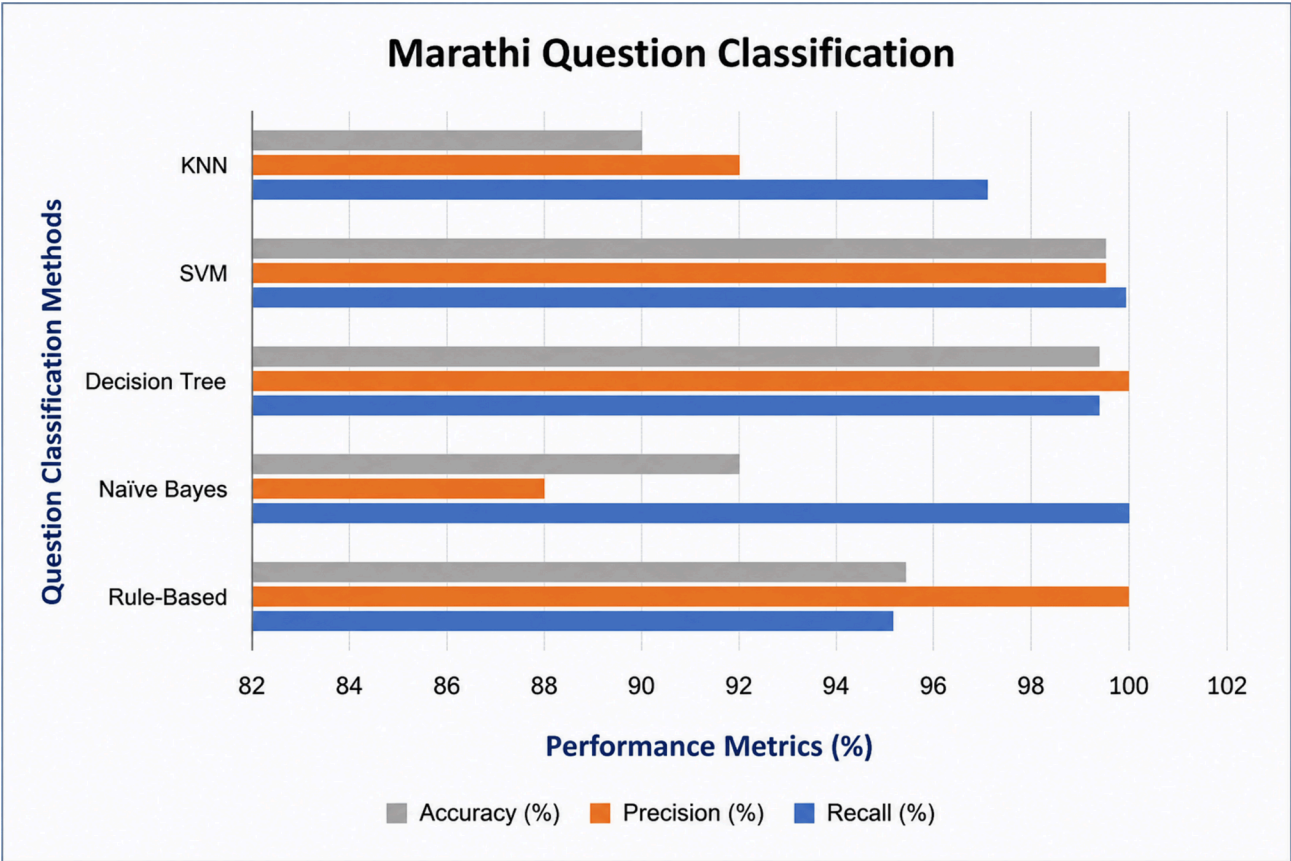


FIGURE 2: Comparison of Performance Measures of Rule-Based and ML-Based Approaches

KNN, K-Nearest Neighbors; ML, Machine Learning; SVM, Support Vector Machine

Overall, the experimental results indicate that the Decision Tree and SVM classifiers achieved the highest performance on the collected Marathi question dataset, with the Decision Tree achieving the best classification accuracy (99.66%). The strong performance of the rule-based approach (95.64%) further confirms that Marathi question classification is largely driven by interrogative lexical patterns. Although the Decision Tree classifier achieved the highest accuracy, tree-based models may be susceptible to overfitting, particularly when trained on relatively small and domain-specific datasets. Since the present study uses 1,193 questions collected from Balbharati textbooks for Standards II-IV, the reported performance should be interpreted within this specific educational domain rather than generalized to all Marathi question types. Error analysis indicates that most misclassifications occurred in questions containing uncommon sentence structures, implicit interrogative expressions, or linguistic variations that were not sufficiently represented in the dataset. These observations suggest that increasing dataset diversity would further improve model robustness and generalization. Overall, the comparative evaluation demonstrates that machine learning approaches outperform manually designed linguistic rules on the current dataset. However, additional validation using larger and more diverse Marathi corpora, along with modern transformer-based language models, is necessary before drawing broader conclusions regarding Marathi QA systems.

Conclusions

This study presented a comparative evaluation of rule-based and machine learning approaches for automatic Marathi question classification. The proposed approaches classified Marathi questions into WH-type and Non-WH-type categories using linguistic rules and TF-IDF-based feature representations with supervised machine learning classifiers. Experimental

How to cite this article:

evaluation on a dataset of 1,193 Marathi questions demonstrated that the Decision Tree classifier achieved the highest classification accuracy (99.66%), followed by SVM (99.41%), while the rule-based approach also achieved competitive performance with an accuracy of 95.64%. The experimental results indicate that machine learning approaches are more effective than manually designed linguistic rules for the collected Balbharati dataset. However, the findings are limited to the current dataset, which represents primary-level educational questions, and therefore should not be generalized to all Marathi question classification tasks without further validation. The present work provides baseline experimental results for Marathi question classification and establishes a foundation for future research in low-resource Marathi NLP. Future work will focus on expanding the dataset by incorporating questions from secondary and higher-secondary textbooks, competitive examinations, and other real-world domains. In addition, transformer-based language models such as IndicBERT, MuRIL, mBERT, and XLM-R will be investigated and compared with traditional machine learning approaches. Further research will also include comprehensive experimental validation using larger datasets, detailed error analysis, statistical significance testing, and the extension of the task to fine-grained multi-class Marathi question classification.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Aarti P. Raut

Acquisition, analysis, or interpretation of data: Aarti P. Raut, C. Namrata Mahender

Drafting of the manuscript: Aarti P. Raut

Supervision: Aarti P. Raut, C. Namrata Mahender

Critical review of the manuscript for important intellectual content: C. Namrata Mahender

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

The datasets (and/or code) supporting this study are available from the corresponding author upon reasonable request.

Acknowledgements

The authors extend their gratitude to the Chhatrapati Shahu Maharaj Research Training and Human Development Institute (SARTHI), Pune, for awarding a fellowship. Additionally, they acknowledge the university scheme SEED Money for the Minor Research Project for University Faculty, sanctioned under Ref. No. Plan&Stat/RDC/2022-23/776-79. They also express their appreciation to the Computational and Psycholinguistics Research Lab Facility for its support and to the Department of Computer Science and Information Technology at Dr. Babasaheb Ambedkar Marathwada University, Chhatrapati Sambhajanagar, Maharashtra, India.

How to cite this article:

Raut A P, Mahender C (August 04, 2026) Enhanced Question Classification for Marathi: A Comparative Study of Rule-Based and Machine Learning-Based Approaches. *Cureus J Comput Sci* 3 : es44389-026-00204-w. DOI <https://doi.org/10.7759/s44389-026-00204-w>

References

1. Kakwani D, Kunchukuttan A, Golla S, Gokul NC, Bhattacharyya A, Khapra MM, Kumar P: [IndicNLP Suite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages](#). Findings of the Association for Computational Linguistics: EMNLP 2020. 2020, 4948-4961. [10.18653/v1/2020.findings-emnlp.445](#)
2. Essam M, Deif MA, Elgohary R: [Deciphering Arabic question: a dedicated survey on Arabic question analysis methods, challenges, limitations and future pathways](#). Artificial Intelligence Review. 2024, 57:251. [10.1007/s10462-024-10880-6](#)
3. Shaikh S, Daudpotta SM, Imran AS: [Bloom's learning outcomes' automatic classification using LSTM and pretrained word embeddings](#). IEEE Access. 2021, 9:117887-117909. [10.1109/access.2021.3106443](#)
4. Sharma H, Mathur R, Chintala T, Dhanalakshmi S, Senthil R: [An effective deep learning pipeline for improved question classification into Bloom's taxonomy's domains](#). Education and Information Technologies. 2023, 28:5105-5145. [10.1007/s10639-022-11356-2](#)
5. Zhang J, Wong C, Giacaman N, Luxton-Reilly A: [Automated classification of computing education questions using Bloom's taxonomy](#). ACE '21: Proceedings of the 23rd Australasian Computing Education Conference. 2021, 58-65. [10.1145/3441636.3442305](#)
6. Al Faraby S, Romadhony A, Adiwijaya: [Analysis of LLMs for educational question classification and generation](#). Computers and Education: Artificial Intelligence. 2024, 7:100298. [10.1016/j.caeai.2024.100298](#)
7. Hamza A, En-Nahnahi N, Alaoui Zidani K, El Alaoui Ouati S: [An Arabic question classification method based on new taxonomy and continuous distributed representation of words](#). Journal of King Saud University Computer and Information Sciences. 2021, 33:218-224. [10.1016/j.jksuci.2019.01.001](#)
8. Mohammed M, Omar N: [Question classification based on Bloom's taxonomy cognitive domain using modified TF-IDF and word2vec](#). PLoS ONE. 2020, 15:e0230442. [10.1371/journal.pone.0230442](#)
9. Yilmaz S, Toklu S: [A deep learning analysis on question classification task using Word2vec representations](#). Neural Computing and Applications. 2020, 32:2909-2928. [10.1007/s00521-020-04725-w](#)
10. Mohasseb A, Bader-El-Den M, Cocea M: [Question categorization and classification using grammar based approach](#). Information Processing & Management. 2018, 54:1228-1243. [10.1016/j.ipm.2018.05.001](#)
11. Van-Tu N, Anh-Cuong L: [Improving question classification by feature extraction and selection](#). Indian Journal of Science and Technology. 2016, 9:1-8. [10.17485/ijst/2016/v9i17/93160](#)
12. Banerjee S, Naskar SK, Rosso P, Bandyopadhyay S: [Classifier combination approach for question classification for Bengali question answering system](#). Sādhanā. 2019, 44:247. [10.1007/s12046-019-1224-8](#)
13. Kamble S, Baskar S: [Learning to classify Marathi questions and identify answer type using machine learning technique](#). Advances in Machine Learning and Data Science. Advances in Intelligent Systems and Computing. Reddy Edla D, Lingras P, Venkatanaresbhabu K (ed): Springer, Singapore; 2018. 705:33-41. [10.1007/978-981-10-8569-7_4](#)
14. Govilkar SS, Bakal JW: [Question answering system using ontology in Marathi language](#). International Journal of Artificial Intelligence and Applications. 2017, 8:53-64. [10.5121/ijaia.2017.8405](#)
15. Devi R, Dua M: [Performance evaluation of different similarity functions and classification methods using web based Hindi language question answering system](#). Procedia Computer Science. 2016, 92:520-525. [10.1016/j.procs.2016.07.377](#)

How to cite this article:

Raut A P, Mahender C (August 04, 2026) Enhanced Question Classification for Marathi: A Comparative Study of Rule-Based and Machine Learning-Based Approaches. Cureus J Comput Sci 3 : es44389-026-00204-w. DOI <https://doi.org/10.7759/s44389-026-00204-w>