

Original Article

Marathi Hate Speech Detection Using Naive Bayes and SVM Techniques

Preeti V. Sarode¹, Harshali B. Patil²

¹Department of Computer Science and Information Technology K. M. Agrawal College
Kalyan, Maharashtra, India

²Department of Computer Science Dr. Annasaheb G. D. Bendale Mahila Mahavidyalaya
Jalgaon, Maharashtra, India

Manuscript ID:
IJERSD-2026-020308

ISSN: 3067-2325

Volume 2

Issue 3

Pp. 38-43

June 2026

Submitted: 12 May 2026

Revised: 19 May 2026

Accepted: 12 June 2026

Published: 30 June 2026

Correspondence Address:
Preeti V. Sarode

Department of Computer Science
and Information Technology K. M.
Agrawal College, Kalyan,
Maharashtra, India.
Email:
preetivsarodephd@gmail.com

Quick Response Code:



Web: <https://rlgjaar.com>

DOI: 10.5281/zenodo.23057817

DOI Link:
<https://doi.org/10.5281/zenodo.23057817>



Attribution-NonCommercial-
ShareAlike 4.0 International



Abstract

The incredible growth of social media has measured to growth in user-generated content in regional languages such as Marathi, which has also triggered in the spread of hate speech and offensive language. It is hard to detect such content due to multifaceted language and less resources. This paper proposes a machine learning-based framework using hybrid TF-IDF features and evaluates multiple classifiers. The L3Cube-MahaHate dataset was used for research. Text preprocessing and feature extraction using TF-IDF were applied to change text content into numerical illustrations. The results illustration that SVM achieves better with an accuracy of 88.45%, while Naive Bayes reaches to 85.01%. The study highlights the importance of feature engineering and model selection in low-resource language processing.

Keywords: Hate Speech Detection, Machine Learning, TF-IDF, Naive Bayes, SVM

Introduction

The exponential progress of social media platforms has transmuted online communication, leading to an extensive flow in user-generated content across regional languages such as Marathi. These platforms help people communicate easily; though, they also spread hate speech and offensive language, which can undesirably affect society and individuals. Therefore, automatic detection of such hurtful content is important. Ranasinghe and Zampieri, (2021) explored hate speech detection largely in high-resource languages such as English. For low-resource languages like Marathi, development remains limited due to challenges such as scarcity of annotated datasets, linguistic diversity, and the presence of informal and code-mixed text, noisy text, and complex language patterns [1].

According to Deshmukh (2024), the L3Cube-MahaHate dataset performed a dynamic role in addressing the insufficiency of resources by contributing a large-scale, manually annotated corpus for Marathi hate speech detection [2].

In this study, Naïve Bayes and Support Vector Machine are applied as classification algorithms for the detection of Marathi hate speech based on TF-IDF feature representation. This study emphasizes on building an automated recognition system of hate speech using machine learning methods.

Related Work

This section carries an inclusive overview of existing literature regarding detection in hate speech, highlighting machine learning-based methods and feature extraction techniques. It observes previous studies associated to Marathi and other low-resource languages, focusing on datasets, methodological advancements and classification models.

Mandl et al. (2020) demonstrated that hate speech detection research has perceptively developed across multiple languages [3].

Gretel (2020) extended multimodal offensive language analysis to English, Arabic, Danish, Greek, and Turkish, systematically comparing BERT against classical baselines including SVM, CNN, and BiLSTM [4]. Coltekin (2020) applied SVM classifiers to annotate Turkish tweets across binary offensive detection, targeted/non-targeted classification, and three-way target categorization [5]. Similarly, Moy et al. (2021) employed Naïve Bayes, SVM, Random Forest and Logistic Regression with n-gram features for detection of Indonesian hate speech, while Khan et al.

Creative Commons

This is an open access journal, and articles are distributed under the terms of the [Creative Commons Attribution NonCommercial-ShareAlike 4.0 International](https://creativecommons.org/licenses/by-nc-sa/4.0/), The Creative Commons Attribution license allows re-distribution and re-use of a licensed work on the condition that the creator is appropriately credited

How to cite this article:

Sarode, P. V., & Patil, H. B. (2026). Marathi Hate Speech Detection Using Naive Bayes and SVM Techniques. *International Journal of Engineering Research for Sustainable Development*, 2(3), 38–43. <https://doi.org/10.5281/zenodo.23057817>

(2020) developed the HS-RU-20 Roman Urdu corpus to evaluate these alongside CNN across multi-level classifications [6][7]. Dravidian language studies progressed with Chakravarthi et al.

(2021) leveraging Logistic Regression and Random Forest with TF-IDF for Tamil datasets [11].

Vasantharajan & Thayasivam (2021) assessing BERT models across Tamil, Kannada, and Malayalam code-mixed offensive content [12]. Velankar et al. (2021) explored Marathi and Hindi offensive and hate speech detection challenges via CNN, LSTM, and BERT variants, demonstrating the utility of machine learning approaches in multilingual and code-mixed settings. Their study emphasized the necessity of preprocessing and feature representation for refining classification accuracy [13].

Gajbhiye (2021) explored the usage of Logistic Regression and Random Forest models with TF-IDF features for hate speech recognition on Twitter data [14]. Collectively, these works underscore classical machine learning approaches—SVM, Naïve Bayes, and Logistic Regression paired with TF-IDF or n-grams—as robust baselines for low-resource Indic scenarios, increasingly complemented by transformer architectures to handle code-mixing complexities.

Several studies have explored the hate speech detection with machine learning techniques. Traditional approaches such as TF-IDF combined with Naïve Bayes and SVM have shown strong performance in text classification tasks by Meetei et al. (2021) [15].

Joshi (2022) and Patil (2022) contributed to dataset development in resources such as L3Cube-MahaHate which was significantly to Marathi natural language processing and pre-trained models [16] [17].

According to Modha, S. (2021) HASOC 2021 provides evaluation frameworks and benchmark datasets for offensive and hate speech results in Indian languages including Marathi and enabling comparative analysis of different approaches [18]. Chavan (2022) proposed transformer-based methods for hate and offensive language recognition in Marathi using Twitter data, emphasizing contextual understanding for detecting implicit hate speech [19].

Further research by Velankar et al. (2022) investigated the performance of monolingual and multilingual transformer-based models for Marathi hate speech detection, highlighting that language-specific models often give better results compared to generalized multilingual approaches [20].

Jahan and Oussalah (2023) highlighted that while significant progress has been achieved for high-resource languages such as English, research in low-resource languages like Marathi is still evolving due to challenges such as limited annotated datasets, informal and noisy text, and code-mixing [21].

Hate speech detection is a motivating task due to the complexity and variability of language, cultural differences, and evolving nature of online content. The different methods and techniques have been developed to identify and combat hate speech online.

Seble et al. (2023) explored machine learning techniques such as Random Forest, Naïve Bayes, Support Vector Machine (SVM) and Logistic Regression are effective for hate speech detection. Naïve Bayes performs well with text-based features such as TF-IDF, while SVM performed strongly with TF-IDF or embedding-based representations, particularly when the dataset is balanced [22]. Overall, the literature review indicated that while deep learning approaches were increasingly being explored, classical machine learning models merged with effective feature extraction techniques and remain highly relevant, especially for low-resource languages like Marathi. However, there is still a need for focused comparative analysis of SVM and Naïve Bayes using TF-IDF features specifically for Marathi hate speech detection study.

Methodology

The planned methodology follows a systematic pipeline including preprocessing of Marathi text, feature extraction using TF-IDF and classification with machine learning models such as SVM and Naïve Bayes. The system is reviewed using standard performance metrics to ensure actual hate speech detection.

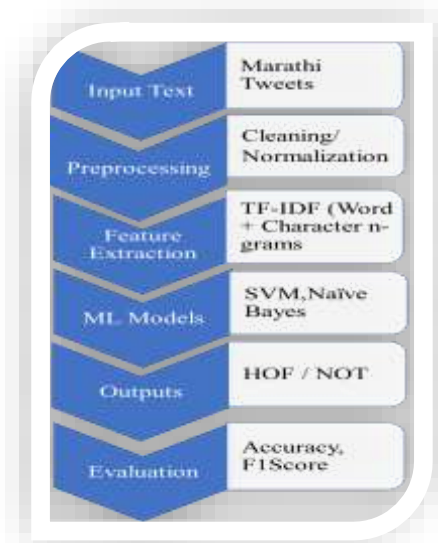


Figure1: Proposed Architecture

Dataset

The planned study employs the L3Cube-MahaHate¹ dataset, a large-scale interpreted corpus precisely curated for hate speech and offensive language detection in Marathi. The dataset consists of Marathi social media textual contents collected from Twitter and representing certainly occurring user-generated social media content. It captures the informal and diverse linguistic nature of social media, including, spelling variations, slang usage, code-mixed text (Marathi-Hindi-English) and context-dependent expressions.

The dataset is provided in both binary and multi-class formats. In this research, the binary classification version is utilized, where the data is separated into two categories:

HOF (Hate/Offensive): The HOF class includes hate speech, offensive language, and profane expressions with individuals or groups based on attributes such as caste, gender or religion,

NOT (Non-hate): The NOT class represents neutral or non-offensive content without harmful intent.

Table 1: Training and Test Dataset Statistics for Marathi Language

Language	Dataset Type	Tweet Type	Count (%)	Total
Marathi	Training	HOF	15,000 (50%)	30,000
Marathi	Training	NOT	15,000 (50%)	
Marathi	Test	HOF	1,875 (50%)	3,750
Marathi	Test	NOT	1,875 (50%)	

The Table1 represents L3Cube-MahaHate dataset is balanced, with equal representation of HOF and NOT classes. This helps avoid bias in the model and ensures fair results. It also includes different topics such as religion, politics, caste, and social issues. The dataset also includes abbreviations, real social media content, short and noisy text. These texts depend on context, improper grammar, and making detection more difficult. Therefore, the dataset is more challenging for machine learning models.

Overall, the L3Cube-MahaHate dataset offers a reliable and comprehensive foundation for emerging and evaluating hate speech detection systems in Marathi. Its real-world relevance, manual annotation quality, and linguistic diversity make it a valuable resource for proceeding study in low-resource natural language processing.

Preprocessing

The data is often in the typical human format, which is in sentences or paragraphs. The information needs to be rehabilitated and cleaned up before analysing it, to confirm that the textual data is transformed into a strategic format which can be efficiently processed by the system.

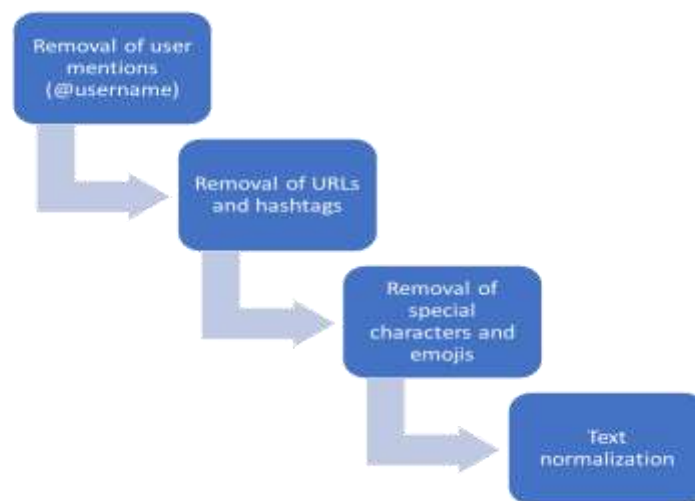


Figure2: Preprocessing Steps

Preprocessing is an important stage in natural language processing and helps to deal with social media data, which is often informal, noisy and unstructured. The L3Cube-MahaHate dataset contains user-generated Marathi text that includes code-mixing, spelling variations, emojis, and informal expressions. Therefore, a structured preprocessing process was applied to improve data quality, enhance model performance and accuracy.

The following preprocessing steps were performed:

- **Removal of user mentions (@name, @id):**
Social media texts often include user allusions, which do not contribute to the semantic meaning of the sentence. Therefore, these user mentions are removed to improve the quality and minimize noise of text.
- **Removal of hashtags and URLs:**
Hashtags were cleaned by removing special symbols while retaining meaningful words wherever applicable. URLs and hyperlinks were eliminated.

- **Removal of emojis and special characters:**

Special characters, Emojis and other symbols were removed to normalize the text. The inconsistent usage of sentimental emojis can introduce noise in traditional machine learning models [15].

- **Text normalization:**

This step supports in reducing unpredictability caused by inconsistent writing styles and improves feature consistency. The text was standardized by removing extra spaces and normalizing the format.

These preprocessing steps confirmed that noisy and irrelevant elements were removed while preserving meaningful textual content. A clean and structured text representation was obtained through the preprocessing phase, which implicitly improved the efficiency of TF-IDF feature extraction and subsequent classification models.

Feature Extraction

Feature extraction is a significant step in transmuting textual data into a structured numerical format suitable for machine learning models. In this study, the Term Frequency–Inverse Document Frequency (TF-IDF) technique was employed to signify Marathi text as numerical feature vectors. TF-IDF allocates higher weights to words that are important within a document while reducing the influence of commonly occurring words across the dataset.

Term Frequency (TF) measures a word appears in a document, and Inverse Document Frequency (IDF) reduces the weight of words that appear frequently across multiple documents. TF-IDF helps identify important words in text classification [22]. This combination helps in identifying discriminative relations that are suitable for classification tasks. TF-IDF was used to convert text into numerical features. The following feature extraction configurations were used:

- **Unigrams and Bigrams:**

Unigrams capture individual word importance. Bigrams help to understand phrase-level meaning and local context, which is beneficial in detecting hate speech expressions.

- **Hybrid Features (Word + Character n-grams):**

The combination of word-level TF-IDF and character-level n-grams (3–5 characters)

captured spelling variations, morphological richness and subword patterns of the Marathi language. These features are fit to handle noisy social media text and improving robustness in low-resource language settings.

The TF-IDF-based feature representation significantly enhanced the performance of machine learning models by effectively capturing both lexical importance and contextual patterns. The researchers have also demonstrated the effectiveness of TF-IDF in hate speech detection tasks and text classification [23].

Results

The evaluation of hate speech detection methods is typically based on precision, recall, and F1 score. Precision is the fraction of true positive examples among the instances that the model classified as positive for a specific class. Recall represents the fraction of

examples classified as positive among the total number of positive examples for each class. The F1 score is then calculated on the basis of precision and recall [23]. Here, TN, TP, FN, and FP stand for true negatives, true positives, false negatives, and false positives.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN} \quad F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The experimental results indicated that both Naïve Bayes and SVM perform effectively for Marathi hate speech detection using TF-IDF features.

Table2: Experimental Results

Algorithms	Accuracy	Precision	Recall	F1-Score
Naive Bayes + TF-IDF	85.01%	85%	85%	85%
SVM + TF-IDF	88.45%	88%	88%	88%

Naïve Bayes performs well due to its probabilistic nature and efficiency in handling high-dimensional sparse data, which is typical in text classification tasks. SVM demonstrates better performance and identify optimal decision boundaries with handle high-dimensional feature spaces. In this results SVM achieves superior performance compared to Naive Bayes, with an accuracy of 88.45% and an F1-score of 88%.

Table 3 demonstrates a confusion matrix performance of a classification models. It compares four components TN, FP, FN and TP which help in understanding model performance beyond overall accuracy.

Table3: Confusion Matrix Results

Model	TN	FP	FN	TP
Naive Bayes	1680	195	357	1518
SVM	1653	222	210	1665

Naive Bayes shows a higher number of false negatives (357) indicating that it fails to detect a greater portion of hate speech occurrences. SVM has fewer false negatives (210) signifying better capability in finding actual hate speech. Although SVM shows slightly higher false positives, it achieves superior overall performance improved detection effectiveness.

Conclusion

The TF-IDF effectively grabbing the meaningful linguistic forms from Marathi social media text. The experimental outcomes clearly indicate that SVM outperforms Naive Bayes in terms of accuracy and F1-score, representing its higher capability in handling sparse and high-dimensional textual data.

The proposed approach founds a vigorous and computationally efficient baseline for hate speech detection in low-resource language like Marathi. This approach highlights the critical role of feature engineering and appropriate model selection in achieving improved classification performance. This research contributes to the development of automated content moderation systems and support advancements in regional language natural language processing. Future work can emphasis on enhancing model performance by including deep learning methods such as Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN). Furthermore, the addition of linguistic features, including semantic context, morphological patterns, and domain-specific lexicons, may significantly improve detection accuracy.

Acknowledgment

The authors express their sincere gratitude to their respective institutions, K. M. Agrawal College, Kalyan, and Dr. Annasaheb G. D. Bendale Mahila Mahavidyalaya, Jalgaon, for providing the academic environment, encouragement, and necessary facilities to carry out this research work.

Financial support and sponsorship

Nil.

Conflicts of interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

1. T. Ranasinghe and M. Zampieri, "Multilingual offensive language identification for low-resource languages," arXiv preprint arXiv:2105.05996, 2021.
2. P. Deshmukh, N. Kulkarni, S. Kulkarni, K. Manghani, and R. Joshi, "Leveraging parameter efficient training methods for low resource text classification: A case study in Marathi," in Proc. IEEE 9th Int. Conf.
3. Convergence Technology (I2CT), 2024, pp. 1–6.
4. T. Mandl, S. Modha, P. Majumder, and D. Patel, "Overview of the HASOC track at FIRE 2020: Hate speech and offensive content identification," CEUR Workshop Proc., pp. 87–111, 2020.
5. G. L. De la Peña Sarraacén, "Multilingual and multimodal hate speech detection in social media," 2020. [5] Ç. Çöltekin, "A corpus of Turkish offensive language on social media," in Proc. 12th Language Resources and Evaluation Conf. (LREC), 2020, pp. 6174–6184.
6. T. X. Moy, M. Raheem, and R. Logeswaran, "Hate speech detection in Indonesian language using machine learning approach," Journal of Physics: Conf. Series, vol. 1952, no. 1, 2021.
7. M. M. Khan, K. Shahzad, and M. K. Malik, "HS-RU-20: Hate speech Roman Urdu corpus," 2020. [8] B. R. Chakravarthi et al., "Overview of the track on sentiment analysis for Dravidian languages in code-
8. mixed text," 2020.
9. Hande, R. Priyadarshini, and B. R. Chakravarthi, "KanCMD: Kannada code-mixed dataset for sentiment analysis and offensive language detection," 2020.
10. N. Romim, M. Ahmed, H. Talukder, and M. S. Islam, "Hate speech detection in the Bengali language: A dataset and its baseline evaluation," 2021.
11. D. Saha et al., "Hate-Alert@DravidianLangTech-EACL2021: Ensembling strategies for transformer-based offensive language detection," 2021.
12. C. Vasantharajan and U. Thayasivam, "Hypers@DravidianLangTech-EACL2021: Offensive language
13. identification in Dravidian code-mixed YouTube comments," 2021.
14. Velankar, H. Patil, A. Gore, S. Salunke, and R. Joshi, "Hate and offensive speech detection in Hindi and Marathi," arXiv preprint arXiv:2108.13230, 2021.
15. D. Gajbhiye, S. Deshpande, P. Ghante, A. Kale, and D. Chaudhari, "Machine learning models for hate speech identification in Marathi language," 2021.
16. L. S. Meetei et al., "Feature engineering for low-resource languages," 2021.
17. H. Patil, A. Velankar, A. Gore, S. Salunke, and R. Joshi, "L3Cube-MahaHate: A tweet-based Marathi hate speech detection dataset and BERT models," in Proc. TRAC, 2022, pp. 1–9.
18. R. Joshi, "L3Cube-MahaNLP: Marathi natural language processing datasets, models, and library," arXiv preprint arXiv:2205.14728, 2022.

19. S. Modha, T. Mandl, P. Majumder, D. Patel, and M. Dave, "Overview of the HASOC 2021 shared task,"
20. FIRE, 2021.
21. T. Chavan et al., "A Twitter BERT approach for offensive language identification in Marathi," arXiv preprint arXiv:2212.10039, 2022.
22. Velankar, H. Patil, A. Gore, S. Salunke, and R. Joshi, "Mono vs multilingual BERT: A case study in Marathi hate speech detection," arXiv preprint arXiv:2204.12233, 2022.
23. M. S. Jahan and M. Oussalah, "A systematic review of hate speech automatic detection using natural
24. language processing," Neurocomputing, vol. 535, 2023.
25. H. Seble et al., "Hate speech detection using machine learning: A survey," AJSE, vol. 20, no. 1, 2023.
26. M. P. Narkhede and H. B. Patil, "Impact of pre-processing for Marathi text classification using SVM and NB," in Proc. ICCSCE, 2025.