

Original Article

A Comparative Study on Deep Learning Approaches For Marathi Text Classification

Madhuri P. Narkhede¹, Harshali B. Patil²

¹Department of Computer Science, Smt. Devkiba Mohansinhji Chauhan College of Commerce & Sciences, Silvassa, UT of DD & Dnh, India

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

Manuscript ID:
IJERSD-2026-020305

ISSN: 3067-2325

Volume 2

Issue 3

Pp. 21-24

June 2026

Submitted: 12 May 2026

Revised: 19 May 2026

Accepted: 12 June 2026

Published: 30 June 2026

Correspondence Address:

Madhuri P. Narkhede
Department of Computer Science, Smt.
Devkiba Mohansinhji Chauhan College
of Commerce & Sciences, Silvassa, UT
of DD & DNH, India
Email:
madhurnarkhede.research@gmail.com

Quick Response Code:



Web: <https://rlgjaar.com>

DOI: 10.5281/zenodo.23057504

DOI Link:
<https://doi.org/10.5281/zenodo.23057504>



Attribution-NonCommercial-
ShareAlike 4.0 International



Abstract

Marathi is a morphologically rich and low-resource language, making automated text classification challenging because of complex linguistic structure and limited supervised data. This paper illustrates a comparative study of various deep learning models for Marathi text classification using a subset of the L3Cube MahaNews dataset comprising 6000 samples across 12 categories. Five architectures CNN, LSTM, BiLSTM, a hybrid CNN-LSTM, and BiLSTM with Attention are implemented and evaluated using a CLS token-based representation to capture global contextual semantics. The models are assessed using accuracy, F1-score, and training time to ensure a comprehensive evaluation of both effectiveness and computational efficiency. Among all the five models implemented BiLSTM with Attention model achieved the impressive performance, with of 89.01% accuracy and 89.10% F1-score. In contrast, the CNN model demonstrates the highest efficiency with the lowest training time, making it suitable for resource-constrained environments. The hybrid CNN-LSTM model provides a balanced trade-off between performance and efficiency. Overall, this study establishes a benchmark and shows that optimized deep learning models can deliver competitive results without relying on computationally intensive transformer-based approaches.

Keywords: Text Classification, NLP, Machine learning, Deep Learning, Hybrid models.

Introduction

Text classification is a crucial part of Natural Language Processing (NLP) that automatically organizes text into pre-defined groups. It has extensive applications in areas such as news classification, spam identification, Information retrieval and many more. During last few years Deep Learning methods which automatically extract the features and find complex semantic relationships in the text have enhanced text classification systems [1, 2]. Despite these advancements, most research in NLP has primarily focused on high-resource languages such as English, leaving many regional languages, including Marathi, relatively underexplored. Marathi is a morphologically rich language written in the Devanagari script and presents unique challenges such as complex word formations, free word order, and limited availability of annotated datasets [3]. These challenges make it difficult to directly apply existing models developed for high-resource languages to low resource Marathi language text classification tasks. Naïve Bayes and Support Vector Machines are two common traditional machine learning methods that have been used to classify text. However, their performance is limited due to their reliance on handcrafted features [4]. CNN and LSTM Deep learning models have shown better performance by learning hierarchical and sequential representations of text data [1]. Furthermore, advanced architectures such as BiLSTM and attention-based models have been presented to improve classification accuracy by capturing contextual dependencies more effectively [5]. Recent developments in transformer-based models, such as multilingual BERT and Indic-specific models, have further advanced text classification performance. However, these models are computationally expensive and require large-scale datasets for effective training, which are often unavailable for low-resource languages like Marathi [6].

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:

Narkhede, M. P., & Patil, H. B. (2026). A Comparative Study on Deep Learning Approaches For Marathi Text Classification. *International Journal of Engineering Research for Sustainable Development*, 2(3), 21–24. <https://doi.org/10.5281/zenodo.23057504>

As a result, there is a growing need to explore efficient deep learning architectures that can achieve competitive performance with limited data and computational resources. The availability of datasets such as the L3Cube MahaNews corpus has facilitated research in Marathi text classification by providing labeled data across multiple categories [7]. However, existing studies often focus on a limited set of models or do not provide a comprehensive comparison of different deep learning architectures under a unified experimental setup. Additionally, factors such as computational efficiency and training time are often overlooked in performance evaluation. To address these limitations, this paper presents a comparative study of five deep learning models for classification of Marathi text data using a subset of the L3Cube MahaNews dataset. Specifically, five architectures CNN, LSTM, BiLSTM, CNN-LSTM, and BiLSTM with Attention are implemented and evaluated using a CLS token-based representation to capture global semantic features of the input text. The models are compared based on accuracy, F1-score, and training time to provide a comprehensive evaluation of both performance and efficiency.

This study presents a comparative evaluation of five deep learning architectures for Marathi text classification under a unified framework, utilizing a CLS token-based representation for capturing global contextual semantics. It further analyzes model performance in terms of both classification accuracy and computational efficiency. The other part of the article is arranged as follows: Section II discusses the related work, Section III describes the suggested method to be implemented, Section IV shows the findings and analysis, and Section V formally wraps up the article with directions for further study.

Related Work

Deep learning techniques have significantly advanced text classification tasks across multiple languages, including low-resource languages such as Marathi. Initial studies in this area mainly concentrated on assessing baseline deep learning architectures such as CNN, LSTM, and ULMFiT. A comparative study demonstrated that CNN and LSTM models, when combined with pretrained embeddings such as FastText, achieved performance comparable to transformer-based models like BERT, particularly in scenarios with limited training data [8]. This observation highlights that relatively simple architectures can still deliver competitive results in low-resource settings where data scarcity is a critical constraint. Building on these findings, subsequent studies explored the role of embedding strategies in improving classification performance. In particular, the use of monolingual and multilingual embeddings has been extensively investigated. Approaches incorporating multilingual transformer models such as IndicBERT and MuRIL, combined with sequential architectures like LSTM, have shown improvements in precision, recall, and F1-score compared to traditional embeddings [9]. However, these models still face challenges in adapting across domains, especially when handling informal, noisy, or user-generated content common in real-world applications.

The availability of large-scale datasets has further accelerated research in Marathi text classification. The L3Cube-MahaNews dataset, comprising over 100,000 labeled samples across multiple categories, has enabled more robust benchmarking of deep learning models. Experimental studies on this dataset indicate that monolingual transformer models such as MahaBERT consistently outperform multilingual counterparts like IndicBERT and MuRIL across varying text lengths [7]. This underscores the importance of language-specific pretraining in achieving superior performance for regional languages. In addition, researchers have explored hybrid architectures to harness the combined capabilities of various deep learning methods. For example, models that combine CNN and BiLSTM have demonstrated improved classification accuracy by effectively capturing both local features and long-range dependencies in text [10]. Such hybrid approaches are particularly well-suited for morphologically rich languages like Marathi, where both contextual semantics and structural patterns play a crucial role in text understanding. Deep learning-based POS tagging methods utilizing architectures such as RNN, LSTM, GRU, and BiLSTM have reported accuracy levels exceeding 97%, with BiLSTM models achieving the best performance [11]. Similarly, transformer-based approaches applied to Marathi NER tasks have shown promising results, although challenges remain in handling informal and noisy textual data [12]. These developments further reinforce the effectiveness of deep learning techniques for Marathi NLP tasks.

Overall, it is analysed that Deep Learning models constantly improving the performance of text classification systems as compared to traditional Machine Learning Models [13]. However, despite these advancements, there remains a lack of comprehensive studies that evaluate multiple deep learning architectures including hybrid and attention-based models under a unified experimental framework on balanced datasets. A systematic comparison of CNN, LSTM, BiLSTM, CNN-LSTM, and BiLSTM with Attention models using the MahaNews dataset is presented to address this gap.

Methodology

This study presents a comparative analysis of multiple deep learning architectures for Marathi text classification using the L3Cube MahaNews dataset. The overall pipeline consists of data preprocessing, feature representation using a CLS-based encoding strategy, model development, and performance evaluation.

3.1 Dataset Description:

The experiments are conducted on a subset of the L3Cube MahaNews Marathi LDSC dataset, consisting of 6000 news samples distributed across 12 categories. The dataset provides a balanced multi-class classification setting suitable for benchmarking deep learning models.

3.2 Data Preprocessing

To ensure data consistency and quality, the following preprocessing steps were applied:

- Unicode normalization for Devanagari text
- Removal of special characters, punctuation, and noise
- Removal of Non Marathi characters
- Tokenization

3.3 Feature Representation using CLS Token

A CLS (classification) token-based representation is used to capture the semantic meaning of the entire input sequence. The CLS token is appended at the beginning of each input sequence and its corresponding embedding is used as a fixed-length representation of the text.

Model Architectures

Five deep learning models were implemented and compared:

- LSTM Model

Captures sequential dependencies in text using memory cells.

- CNN Model

Extracts local n-gram features using convolution filters.

- BiLSTM Model

Process the in data both forward and backward directions using two LSTM layers.

- CNN-LSTM Hybrid Model

Combines feature extraction using CNN and implements sequence modeling using LSTM.

- BiLSTM with Attention Mechanism

Assigns importance weights to relevant words in the sequence.to enhances the performance of BiLSTM.



Figure 1 Proposed System Architecture for Marathi Text Classification

The models were trained using an 80:20 train-test split to ensure a reliable evaluation of generalization performance. Early stopping was incorporated to prevent overfitting by monitoring validation loss. The implemented models are evaluated using standard classification metrics such as accuracy, precision, recall and F1-score, and are considered as the key metric due to the multi-class nature of the Text classification problem. In addition to predictive performance, training time was recorded as a measure of computational efficiency, enabling a comprehensive comparison of both effectiveness and resource utilization across different deep learning architectures.

Results & Discussion

The models proposed are tested on the test dataset and the metrics used are accuracy, F1-score and training time. A comparative evaluation of all the five architectures is shown in Table 1.

Model	Test Accuracy	F1 Score	Training Time (s)
LSTM	87.27%	87.17%	638.3576019
BiLSTM	86.73%	86.72%	1527.270632
CNN	88.52%	88.61%	209.8458755
CNN_LSTM	87.74%	87.70%	360.0126126
BiLSTM_Attention	89.01%	89.10%	776.8735383

Table 1 Performance Comparison of Various DL Models

The findings of the experiment show that BiLSTM with Attention model provides the best classification result with the accuracy of 89.01% and the F1-score of 89.10%. The attention method provides a great improvement since it helps the model to focus on the most significant parts of the input sequence increasing the contextual representation. The CNN model demonstrates the second-best performance, achieving an F1-score of 88.61%, while also exhibiting the lowest training time (209.85 seconds). This indicates that CNN is highly efficient in extracting local features and is well-suited for scenarios where computational resources or training time are limited.

The CNN-LSTM hybrid model provides a balanced trade-off between performance and efficiency, achieving an F1-score of 87.70% with moderate training time. This confirms that combining convolutional feature extraction with sequential modeling improves performance compared to standalone LSTM. The LSTM model shows stable and competitive performance, achieving an F1-score of 87.17%, but is outperformed by CNN and hybrid architectures. The BiLSTM model, despite its ability to capture bidirectional context, does not significantly outperform LSTM and incurs the highest training time (1527.27 seconds), making it computationally expensive without proportional performance gains.

A significant variation in training time is observed across models. CNN is the most efficient model, requiring **less training time** compared to sequential models such as LSTM and BiLSTM. In contrast, BiLSTM exhibits the highest computational cost due to its bidirectional processing. The addition of attention in BiLSTM further increases training time but results in noticeable performance improvement, highlighting a trade-off between accuracy and computational overhead.

These results confirm the effectiveness of attention-based and hybrid deep learning approaches for Marathi text classification and establish a strong benchmark for future research in low-resource language processing.

Conclusion

This study presented a comparative evaluation of multiple deep learning architectures for Marathi text classification using the L3Cube MahaNews dataset. The results indicate that the BiLSTM with Attention model achieves the best classification performance by effectively capturing contextual dependencies and highlighting the most relevant textual features. The CNN model, on the other hand, offers superior computational efficiency with significantly reduced training time, making it well-suited for real-time and resource-constrained environments. The hybrid CNN-LSTM model provides a balanced approach, combining feature extraction and sequential learning to achieve competitive performance. The study also highlights that increasing model complexity does not always lead to proportional performance gains, as observed in the case of the BiLSTM model without attention. The use of CLS token-based representation contributes to consistent performance across all models by enabling effective global semantic encoding of input text. Overall, the results confirm that deep learning architectures, particularly attention-based and hybrid models, are well-suited for Marathi text classification even with relatively limited data.

Despite these contributions, certain challenges remain, including limited dataset size, lack of code-mixed text handling, and the need for efficient deployment in low-resource environments. Future work may focus on incorporating transformer-based models with parameter-efficient fine-tuning techniques, exploring data augmentation strategies, and improving cross-domain generalization. These directions can further enhance the robustness and scalability of Marathi text classification systems.

Acknowledgment

The authors express their sincere gratitude to their respective institutions, Smt. Devkiba Mohansinhji Chauhan College of Commerce & Sciences, Silvassa, and Dr. Annasaheb G. D. Bendale Mahila Mahavidyalaya, Jalgaon, for providing the necessary academic environment, encouragement, and support to carry out this research. The authors also acknowledge the developers of the L3Cube MahaNews dataset for making valuable Marathi language resources available to the research community, which significantly contributed to the successful completion of this study.

Financial support and sponsorship

Nil.

Conflicts of interest

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

Reference

1. Kim, Y. *Convolutional neural networks for sentence classification*. in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.
2. Zhang, X., J. Zhao, and Y. LeCun, *Character-level convolutional networks for text classification*. Advances in neural information processing systems, 2015. **28**.
3. Kapse, V.S. and S.S. Deshmukh. *Text Preprocessing Technique of Marathi Documents for Text Summarization*. in *2025 International Conference on Applications of Machine Intelligence and Data Analytics (ICAMIDA)*. 2025. IEEE.
4. Sebastiani, F., *Machine learning in automated text categorization*. ACM computing surveys (CSUR), 2002. **34**(1): p. 1-47.
5. Yang, Z., et al. *Hierarchical attention networks for document classification*. in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*. 2016.
6. Mohammed, A.H. and A.H. Ali. *Survey of bert (bidirectional encoder representation transformer) types*. in *Journal of Physics: Conference Series*. 2021. IOP Publishing.
7. Mittal, S., et al. *L3cube-mahanews: News-based short text and long document classification datasets in marathi*. in *International Conference on Speech and Language Technologies for Low-resource Languages*. 2023. Springer.
8. Kulkarni, A., et al. *Experimental evaluation of deep learning models for marathi text classification*. in *Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications: ICMISC 2021*. 2022. Springer.
9. Eranpurwala, F., P. Ramane, and B.K. Bolla. *Comparative study of Marathi text classification using monolingual and multilingual embeddings*. in *International conference on advanced network technologies and intelligent computing*. 2021. Springer.
10. Agarwal, M., et al., *Hybrid CNN-BiLSTM Model for Enhancing Sentiment Analysis using Text Classification on WhatsApp Group*.
11. Deore, P.R., N.V. Patil, and A.S. Patil, *Deep Learning-Based Parts-of-Speech Tagging in Marathi Language*. Procedia Computer Science, 2025. **258**: p. 3771-3780.
12. Litake, O., et al. *Mono Versus Multilingual BERT: A Case Study in Hindi and Marathi Named Entity Recognition*. in *Proceedings of 3rd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications: ICMISC 2022*. 2023. Springer.
13. Zulqarnain, M., et al., *Text classification using deep learning models: a comparative review*. Cloud Computing and Data Science, 2024: p. 80-96.