Grenze International Journal of Engineering and Technology, June Issue

Marathi Paraphrase Detection using Lexical Similarity and MahaBERT - based Semantic Analysis

Sheetal R. Dehare[1], Aarti P. Raut[2], Rajashri G. Kanke[3] and C. Namrata Mahender[4]
[1-4]Department of CS and IT, Dr. Babasaheb Ambedkar Marathwada University, Chh. Sambhanjinagar, India
Email: dehareshеetal04@gmail.com, rautaarti443@gmail.com, csit.rgk@bamu.ac.in, cnamrata.csit@bamu.ac.in

***Abstract*— Paraphrase detection is an essential problem in Natural Language Processing (NLP) concerned with identifying if two sentences have similar meanings, but are said in different ways or presented in different structures. Although tremendous strides have been made for languages with abundant resources like English, research in Indian languages like Marathi is still limited, as there are few annotated datasets and language-specific models available. In this study, a Marathi paraphrase detection system has been created with both lexical and deep learning techniques. A custom dataset of 3,540 sentences was developed from the textbooks from Balbharati, with each sentence having three paraphrased variants created manually. The system compares two approaches: lexical-based Jaccard similarity and contextual semantic-based cosine similarity using a transformer-based approach that captures contextual semantic representations, MahaBERT. Experimental results show that Jaccard similarity would only compare on the surface level and not recognize paraphrases when different words were used. On the other hand, MahaBERT demonstrates a remarkable ability to represent semantic links between sentences and substantially outperforms the lexical approach. The accuracy of this model is 88%, with a precision of 87%, a recall of 89% and F1 of 88%. The results highlight the importance of deep learning-based contextual models for paraphrase detection in morphologically rich and low-resource languages like Marathi. This work aims at promoting Marathi NLP and act as a stepping stone for further research on semantic similarity and other applications.**

***Index Terms*— Natural Language Processing (NLP), Paraphrase Detection, Semantic Similarity, Transformer Models, MahaBERT, Cosine Similarity.**

I. INTRODUCTION

During conversations, individuals may use different words or phrases to convey the same meaning. For instance, the sentence for example, "ती बाजारात गेली." and "तिने बाजाराला भेट दिली" have the same meaning. Identification of Paraphrase is one of the basic tasks in Natural Language Processing (NLP), which is the problem of detecting whether two sentences are paraphrased or not, despite the differences in their expression. Humans can easily: recognize such semantic similarities, it remains a challenging problem for machines. This ability is vital for several applications such as question answering systems, chatbots, machine translation, content moderation and plagiarism detection. [1]

Paraphrase detection has been a well-studied problem in English but the study is comparatively less in Indian languages like Marathi.

Grenze ID: 01.GIJET.12.2.334_26

The rich morphology and flexible word order, dialectal variations, compound word formation, and the frequent use of informal expressions are some of the challenges of Marathi language for computational processing, which is an Indo Aryan language widely spoken. These characteristics Improve semantic knowledge of autonomous systems. [2]

The methods used so far for paraphrase detection are lexical similarity-based methods which compare sentences with respect to the word similarity. One popular approach is called Jaccard similarity and is based on the number of common words found between two sentences, divided by the number of all words in both sentences. But the approaches described above are not always successful in terms of semantic equivalence when the same meaning is formulated in different words or combinations of words.[3][4] For this reason, to capture the linguistic nuances of low-resource and morphologically rich languages like Marathi, recent advances in deep learning, particularly transformer-based models like BERT (Bidirectional Encoder Representations from Transformers), are used to train a model specifically for Marathi that can generate more representative and context-aware semantic representations.[5][6]

To overcome this limitation, MahaBERT, a transformer-based model specifically pretrained on large-scale Marathi text corpora, is employed to generate more accurate and context-aware semantic representations. Furthermore, due to the scarcity of annotated datasets for Marathi, this study constructs a small yet carefully curated and manually verified paraphrase dataset. The dataset is derived from Marathi Balbharati textbooks, where each source sentence was paired with three diverse paraphrased variants to ensure both linguistic richness and quality.

In this paper, a comparative study on paraphrase detection with a traditional lexical similarity-based approach, Jaccard similarity and MahaBERT-based contextual embeddings is carried out. The main aim is to compare the effectiveness of deep learning based semantic representations to represent the meaning equivalence with the word overlap methods at the surface level.

II. Literature Review

Over the past several years, the field of paraphrase detection in Marathi and other Indian languages has progressed a long way, from statistical to transformer-based methods. The initial research work on this area was mainly with rule-based and semantic graph methods.

Shruti Srivastava and Sharvari Govilkar in 2020, investigated how to detect paraphrases in Marathi using statistical and semantic similarity measures with a representation based on UNL (Universal Networking Language) graphs. Their method showed the feasibility of semantic representations, but limited to the scope of their small data set, Handcrafted features are a reliance on handmade features. [7] In the same period, transformer-based multilingual models began to emerge. Shivam Gupta and colleagues introduced IndicBERT, trained on large-scale Indic corpora. Although IndicBERT improved cross-lingual understanding, its multilingual design limited its ability to capture fine-grained linguistic nuances specific to Marathi. [8][9]

A significant advancement came in 2021 with the development of MahaBERT by researchers from L3Cube Pune. MahaBERT provided contextual embeddings tailored to Marathi by training on a large monolingual corpus. However, its application to paraphrase detection remained underexplored at the time. Building on multilingual sentence representation, Raviraj Joshi and his team later introduced IndicSBERT in 2023[10] [11], focusing primarily on cross-lingual similarity rather than Marathi-specific paraphrase tasks. Similarly, Vidula Magdum et al. developed an NLP toolkit for Marathi language based on transformer architecture model with MahaBERT embeddings for various NLP tasks such as classification and tagging without any specific focus on paraphrase datasets. More recent studies show a change in focus to task-specific implementations. [12]

In 2025, Suramya Jadhav et al. used BERT-based models to classify paraphrase of sentences in Marathi using a huge corpus of thousands of sentence pairs. This was a step forward towards practical in applications, the data set size was still not large when compared to high resource languages. At the same time, Dianeliz Ortiz Martes and co-authors tested transformer models like BERT and RoBERTa on benchmark datasets of semantic similarity tasks. The aim of their work was more global, however, and they did not address the challenges of low resource Indian languages such as Marathi. However, their work is mainly global, and they did not consider the challenges of low resource Indian languages such as Marathi. [13] [14]

Overall, the transformer-based models have yielded remarkable performance in paraphrase detection, but there are still significant challenges, such as the lack of large-scale high-quality annotated datasets and Marathi-specific fine-tuning approaches. Nevertheless, there are a number of studies that are based on multilingual models or bring in datasets without extensive evaluation. Hence, it is clear that there is a need to develop and evaluate systematically Marathi specific Paraphrase detection approaches. To tackle this a study has been

conducted to assess MahaBERT on a custom-built Marathi paraphrase database from Balbharati textbooks. The goal of this work is to gain deeper insights in semantic similarity modelling for Marathi and to help toward better language-specific solutions in NLP.

III. Proposed Methodology

Paraphrase Detection: Given two sentences, determine whether they are paraphrases of each other or not, even if they are written in different words or with different structures. Two sentences that have the same meaning are paraphrases; two sentences that do not have the same meaning are non-paraphrases.

Since there was not enough access to publicly available datasets to build paraphrase detection models for Marathi, a custom dataset for the purpose of study was created. The Corpus comprises of 3540 Marathi sentences from Balbharati Textbooks. For each original sentence there are three paraphrased sentences and for each original sentence there are three pairs of sentences that convey the same meaning but with different lexical and syntactic structures. Each paraphrase Slightly alters the meaning and/or the words or the grammar. This dataset allows a detailed study of semantic similarity, by providing information on several distinct paraphrase pairs of a single source sentence. The proposed methodology of the system is discussed below.

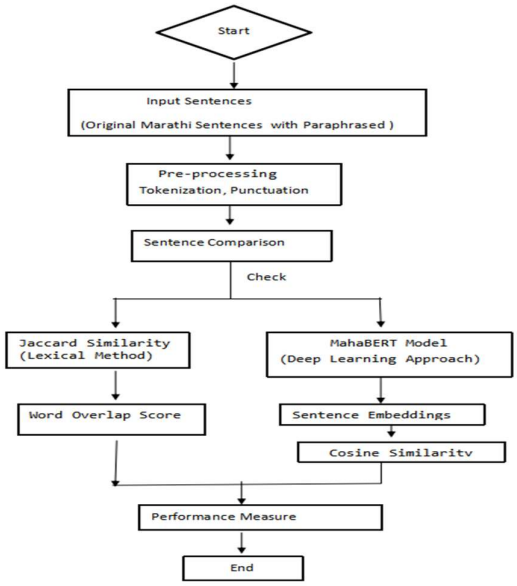

Figure 1: Flowchart of Marathi Paraphrase Detection Using Jaccard Similarity and MahaBERT Model

A. Step 1: Start

The process begins with the initialization of the system for analyzing sentence similarity in Marathi.

B. Step 2: Input Sentences

In this step, the system takes Marathi sentences as input from a custom-created dataset. The dataset was constructed using sentences collected from Marathi BalBharati textbooks. Each entry consists of one original sentence along with three manually created paraphrased versions. These paraphrases preserve the original meaning while introducing variations in wording and sentence structure.

Table I: Sample of Original Marathi Sentences with Three Paraphrase Sentences

Sr. No	Original Sentences	Paraphrase1	Paraphrase 2	Paraphrase 3
1	देश तुझा किती हे या पाढ्याचे नाव आहे.	या धड्याचे शीर्षक 'देश तुझा किती' असे आहे.	हे शीर्षक या पाठाला दिलेले आहे.	धड्याच्या शीर्षकाचे नाव 'देश तुझा किती' असे आहे.
2	देश खूप सुंदर आहे.	आपला देश अत्यंत आकर्षक आहे.	देशाचे सौंदर्य फार मोठे आहे.	हा देश मनमोहक व सुंदर आहे.
3	आपण भारतात राहतो.	आपले निवासस्थान भारत आहे.	आपले वास्तव्य भारत देशात आहे.	आपण भारतीय आहोत आणि इथेच राहतो.

4	ह्या देशात झाडे, फुले, चांगले गाणी आणि खेळ असतात.	या देशात हिरवीगार झाडे, सुंदर फुले, तसेच उत्तम गाणी आणि खेळ उपलब्ध आहेत.	देशात निसर्गरम्य झाडे व फुले, तसेच मनोरंजनासाठी गाणी-खेळ आहेत.	इथे वनस्पती, पुष्पं, चांगल्या कला (गाणी) आणि क्रीडा पाहायला मिळतात.
5	हो, आपल्याला आपल्या देशावर प्रेम आहे.	नक्कीच, आपल्या देशावर आपले निस्सीम प्रेम आहे.	आपली देशभक्ती खूप प्रबळ आहे.	होय, आपण आपल्या देशाबद्दल आकर्षण बाळगतो.

As observed, both original and paraphrased sentences convey the same meaning with slight lexical and structural variations.

C. Step 3: Pre-processing

Pre-processing is an essential step to clean and prepare the text for accurate similarity analysis. The below steps used to preprocess the data:

Punctuation Removal

Punctuation marks (such as commas, full stops, etc.) are removed to retain only meaningful words, improving similarity computation.

Tokenization

Tokenization is the process of breaking a sentence or text into smaller units called tokens. These tokens are usually words, but they can also be characters or sub-words.

Example:

Sentence: राम रोज शाळेत जातो and Tokenization: राम | रोज | शाळेत | जातो.

This is an important pre-processing step in NLP because it helps the system understand and analyze the text more easily.

D. Step 4: Sentence Comparison

The proposed Marathi paraphrase detection system consists of two parallel similarity computation modules. After pre-processing, sentence pairs are processed independently through the lexical similarity module and the semantic similarity module. The lexical similarity method focuses on the surface-level comparison of words by measuring the overlap between the words in both sentences. In contrast, the deep learning method uses advanced language models to understand the contextual and semantic meaning of the sentences.

Approach 1: Lexical Similarity (Jaccard Similarity)

The pre-processed sentences are converted into token sets. Jaccard similarity is computed using the ratio of common words to total unique words. A predefined threshold is applied to classify sentence pairs as paraphrases. It is important because it is simple, fast, and easy to compute, making it useful for large datasets. Although it does not capture deep meaning, it provides a good basic comparison. Therefore, it is used in this work as a baseline method to compare with advanced deep learning approaches.

Jaccard similarity is a method used to measure similarity between two sentences based on common words. It compares the number of shared words with the total number of unique words in both sentences. In this work, it is used to calculate how similar the original sentence is to its paraphrased sentences. The following formula is used to calculate lexical similarity between two sets of words.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \qquad (1)$$

Where,

A= words in sentence 1

B= words in sentence 2

$A \cap B$ = common words of sentence 1 and sentence 2

$A \cup B$ = total unique words of sentence 1 and sentence 2

Approach 2: Deep Learning (MahaBERT Model)

For semantic similarity computation, each sentence is passed through the pretrained MahaBERT model. The contextual embedding corresponding to the sentence representation is extracted. Cosine similarity is then calculated between the embedding vectors of the sentence pair. If the cosine similarity score exceeds the selected threshold value, the pair is classified as a paraphrase.

MahaBERT is a transformer-based model designed specifically for Marathi. It captures contextual meaning rather than relying only on word overlap. Each sentence is converted into a numerical vector (embedding), representing its semantic meaning.

Cosine Similarity

To measure the similarity between sentence embeddings, cosine similarity is used. It calculates how close the meanings of two sentences are by measuring the angle between their vectors.

$$\textbf{Cosine Similarity} = \frac{A.B}{||A||X||B||} \quad (2)$$

Where,
A and B are sentence vectors, A.B is the dot product and ||A|| and ||B|| are vector magnitudes.

E. Step 5: Performance Measures

In this step, the performance of both approaches; Jaccard similarity (lexical-based) and MahaBERT with cosine similarity (semantic-based) was evaluated. After computing similarity scores for sentence pairs, each pair is classified as Paraphrase (Similar). To measure how well the system performs, standard evaluation metrics are used.

Accuracy

Accuracy measures the overall correctness of the sentences. It tells us how many predictions are correct out of all predictions.

$$\text{Accuracy} = \frac{TN+TP}{TN+FP+TP+FN} *100 \quad (3)$$

True Positive (TP): Correctly identified paraphrase pairs
True Negative (TN): Correctly identified non-paraphrase pairs
False Positive (FP): Incorrectly predicted as paraphrase
False Negative (FN): Missed paraphrase (predicted as non-paraphrase)

Precision

Precision is defined as the ratio of True Positives to Total Positives. It demonstrates how many "yes" predictions generated by the system are right. It aids in reducing incorrect "yes" guesses, often known as false positives (FP). The precision is determined as,

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)+FalsePositive (FP)}} *100 \quad (4)$$

Precision measures how accurately the model predicts positive cases. It is calculated as the ratio of correctly predicted positive instances (TP) to all instances predicted as positive (TP +FP). A high precision means that when the model predicts sentences as similar, it is usually correct and makes fewer false positive errors.

Recall

Recall shows how effectively a model can identify all the right "yes" cases in the data. It counts the number of real positive cases that the model was able to find. The formula for figuring out recall is,

$$\text{Recall} = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP)+False\ Negatives(FN)} *100 \quad (5)$$

Recall measures how well the model identifies all actual positive cases. It is calculated as the ratio of correctly predicted positive instances (TP) to all actual positive instances (TP + FN). A high recall means the model successfully detects most of the similar (paraphrase) sentences and misses fewer true cases.

F1 score

The F1 score measures the balance between precision and recall. It is calculated as the harmonic mean of these two metrics, offering a single value that effectively accounts for both false positives and false negatives.

$$\text{F1 Score} = 2 \text{ x } \frac{Precision * Recall}{Precision+Recal} *100 \quad (6)$$

F. Step 6: End

The process ends after generating the similarity scores and comparing the results of both methods.

IV. Result and Analysis

The performance of the proposed Marathi paraphrase detection system was evaluated using two approaches: a lexical-based method (Jaccard similarity) and a deep learning-based method (MahaBERT with cosine similarity).

The evaluation was conducted on a custom dataset consisting of 3,540 Marathi sentences, where each original sentence is paired with three paraphrased variants.

The experimental results reveal clear differences in the behavior of the two approaches. The Jaccard similarity method computes similarity based on word overlap between sentences. It performs adequately when sentences share common words; however, it fails to capture semantic equivalence when different lexical items or syntactic structures are used. A detailed examination of sentence pairs reveals the limitations of lexical similarity methods. Consider the following example:

Original Sentence: "आपण भारतात राहतो."

Paraphrase: "आपले वास्तव्य भारत देशात आहे."

The Jaccard similarity score for this pair is very low because the two sentences share very few common words. However, both sentences convey the same meaning. MahaBERT assigns a high semantic similarity score because it captures contextual meaning rather than relying solely on word overlap.Similarly, sentences containing synonyms, reordered phrases, or grammatical transformations often receive low Jaccard scores but high MahaBERT similarity scores. This demonstrates the capability of transformer-based models to capture deeper semantic relationships in Marathi.

As a result, many semantically similar sentence pairs receive low similarity scores. This limitation is particularly evident in Marathi due to its rich morphology and flexible word order. The table below shows the Sample score of Jaccard similarity:

TABLE II: SAMPLE SCORE OF JACCARD SIMILARITY

Original	Jaccard_P1	Jaccard_P2	Jaccard_P3
01	0.230769231	0.272727273	0.230769231
02	0.285714286	0.125	0.428571429
03	0	0	0.285714286
04	0.375	0.125	0.055555556
05	0.444444444	0.1	0.090909091

In contrast, the MahaBERT-based approach generates contextual embeddings that represent the semantic meaning of sentences. By applying cosine similarity to these embeddings, the model effectively captures deeper relationships between sentences, even when they lexically different. Consequently, MahaBERT produces consistently higher similarity scores for paraphrased sentence pairs and demonstrates superior capability in identifying semantic equivalence.

The table below shows the sample score of the MahaBERT Similarity:

TABLE III: SAMPLE SCORE OF MAHABERT SIMILARITY

Original	BERT_Sim_P1	BERT_Sim_P2	BERT_Sim_P3
01	0.8860804	0.8068156	0.856563
02	0.80747634	0.8157744	0.86308306
03	0.7364261	0.8232223	0.89697254
04	0.8609422	0.8836241	0.8844763
05	0.852957	0.8030334	0.8002645

A comparative analysis of similarity scores shows that Jaccard values are often low (ranging approximately between 0.1 and 0.4) for valid paraphrases, whereas MahaBERT scores remain high (typically between 0.8 and 0.9), indicating strong semantic alignment. This highlights the inability of lexical methods to handle synonymy and structural variation, while transformer-based models successfully address these challenges.The performance of both methods was further evaluated using standard metrics, including Accuracy, Precision, Recall, and F1-score. The results were presented in Table 4.

TABLE IV: PERFORMANCE COMPARISON OF JACCARD AND MAHABERT METHODS FOR MARATHI PARAPHRASE DETECTION IN PERCENTAGE

Method	Accuracy	Precision	Recall	F1 Score
Jaccard	55	50	48	49
MahaBERT	88	87	89	88

The experimental results indicate a substantial improvement when using MahaBERT over the traditional Jaccard similarity approach. MahaBERT achieved an accuracy of 88%, compared to only 55% for Jaccard similarity, representing a 33% absolute improvement. Similarly, precision increased from 50% to 87%, recall improved from 48% to 89%, and the F1-score increased from 49% to 88%.

These improvements demonstrate that contextual semantic representations generated by MahaBERT are significantly more effective in identifying paraphrases than lexical overlap-based methods. The large gain in recall suggests that MahaBERT successfully detects semantically equivalent sentence pairs even when they contain different vocabulary and sentence structures.

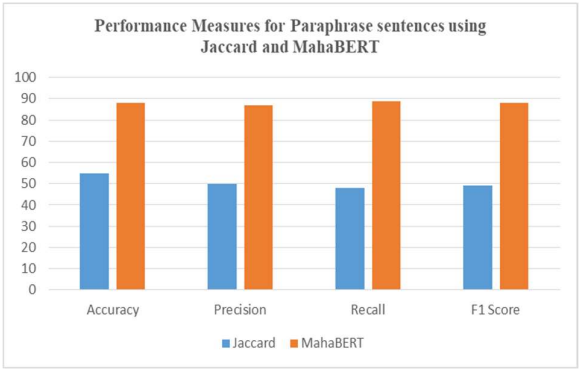

Figure 2: Performance Measures of Paraphrase Sentences using Jaccard and MahaBERT

Overall, the experimental findings clearly demonstrate that deep learning-based approaches, particularly those leveraging contextual embeddings such as MahaBERT, are significantly more effective than traditional lexical similarity methods for Marathi paraphrase detection. The results emphasize the importance of semantic understanding in processing morphologically rich and low-resource languages.

V. Conclusion and Future Scope

This study presented a Marathi paraphrase detection system using both lexical and deep learning approaches. A custom dataset consisting of 3,540 sentences was developed from Balbharati textbooks, where each sentence was paired with three paraphrased variants to ensure diversity in lexical and syntactic structures. The performance of two methods—Jaccard similarity and MahaBERT with cosine similarity—was evaluated and compared. The results clearly demonstrate that Jaccard similarity, being a lexical-based approach, is limited to surface-level word matching and fails to capture semantic equivalence when different words or sentence structures are used. This limitation is particularly significant in Marathi due to its rich morphology and flexible linguistic patterns.

In contrast, the MahaBERT-based approach effectively captures contextual and semantic relationships between sentences through deep learning embeddings. It consistently produces higher similarity scores and achieves superior performance across all evaluation metrics, including accuracy, precision, recall, and F1-score. Therefore, it can be concluded that deep learning-based contextual models significantly outperform traditional lexical methods for Marathi paraphrase detection. The study highlights the importance of semantic understanding in handling low-resource and morphologically rich languages.

Although MahaBERT achieved strong performance, some errors were observed. Certain sentence pairs containing highly domain-specific vocabulary, rare expressions, or ambiguous linguistic structures produced lower similarity scores. In a few cases, sentences sharing several common words but expressing different meanings were incorrectly classified as paraphrases. These observations indicate that further fine-tuning of MahaBERT on larger Marathi paraphrase datasets may improve performance and reduce classification errors.

The proposed work can be extended by increasing the size and diversity of the Marathi paraphrase dataset to improve model performance. Further improvements can be achieved by fine-tuning MahaBERT on task-specific data and exploring advanced models such as IndicBERT and SBERT. Additionally, combining lexical and deep learning approaches, along with incorporating linguistic features like POS tagging and dependency parsing, may enhance accuracy. The approach can also be extended to other Indian languages and real-world applications such as chatbots and automated answer evaluation systems.

Acknowledgment

The authors extend their gratitude to the Dr. Babasaheb Ambedkar Research and Training Institute (BARTI) and Chhatrapati Shahu Maharaj Research Training and Human Development Institute (SARTHI), Pune, for awarding a fellowship. Additionally, under the university scheme SEED Money for Minor Research project for university

faculty sanction with 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 Sambhajinagar, Maharashtra, India.

REFERENCES

[1] Professor Smita Chunamari, Sahil Team, Bhavesh Sonawane , Yash Daund , Janhvi Pawar (2025) Paraphrase Detection in Indian Language.International Journal of Scientific Research & Engineering Trends Volume 11, Issue 3, May-June-2025, ISSN (Online): 2395-566X .

[2] Dani, A., & Sathe, S. R. (2024). A review of the Marathi natural language processing. arXiv preprint arXiv:2412.15471.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proc. NAACL-HLT*, 2019.

[4] B. Min, H. Ross, E. Sulem, et al.2021, "Recent advances in natural language processing via large pre-trained language models: A survey," *arXiv preprint arXiv:2111.01243.*

[5] A. Lahoti et al. 2022, "Indic NLP: Challenges and opportunities for low-resource Indian languages".

[6] S. Dani and M. Sathe 2024, "Challenges in processing morphologically rich Indian languages," R. Joshi, "L3Cube-MahaCorpus and MahaBERT: Marathi monolingual corpus and language models," *arXiv preprint arXiv:2202.01159*, 2022.

[7] Srivastava, S., & Govilkar, S. (2020). Paraphrase detection in Marathi using statistical and semantic similarity measures based on Universal Networking Language (UNL). *International Journal of Advanced Computer Science and* Applications *(IJACSA)*, 11(6), 1–7.

[8] Gupta, S., Goyal, P., Kumar, A., Singh, V., & Kumaraguru, P. (2020). IndicBERT: A pre-trained language model for Indian languages. *Proceedings of the Findings of the Association for Computational Linguistics (EMNLP Findings* 2020*)*, 1–10. https://doi.org/10.18653/v1/2020.findings-emnlp.139.

[9] Kulkarni, A., Mandhane, P., Likhitkar, M., Kshirsagar, A., & Joshi, R. (2021). *MahaBERT: A pre-trained* language *model for Marathi*. L3Cube Pune. https://doi.org/10.48550/arXiv.2103.03761.

[10] Kakwani, D., Kunchukuttan, A., Golla, S., Bhattacharyya, A., Khapra, M. M., & Kumar, P. (2020). IndicNLPSuite*: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages*. In Proceedings of LREC.
(Used as supporting reference for Indic models' context, including IndicBERT).

[11] Joshi, R., Kale, P., & Bhattacharyya, P. (2023). *IndicSBERT: Sentence embeddings for Indian languages using* Siamese *BERT networks*. arXiv preprint arXiv:2303.

[12] Magdum, V., Patil, S., & Joshi, R. (2022). *A transformer-based toolkit for Marathi natural language processing*. In Proceedings of the Workshop on NLP for South Asian Languages.

[13] Jadhav, S., Shanbhag, A., Thakurdesai, A., & Joshi, R. (2025). *Marathi paraphrase detection using BERT-based models*. In Proceedings of the Pacific Asia Conference on Language, Information and Computation (PACLIC).

[14] Ortiz Martes, D., Lopez, J., & Fernandez, A. (2022). *Evaluation of BERT and RoBERTa for semantic textual similarity tasks*. *Expert Systems with Applications*, 198, 116780. https://doi.org/10.1016/j.eswa.2022.116780.