Grenze International Journal of Engineering and Technology, June Issue

A Neural Universal Dependencies-based Framework for Analyzing POS Divergence in Marathi Automated Short Answer Evaluation

Aarti P. Raut[1], Sheetal R. Dehare[2] and C. Namrata Mahender[3]
[1-3]Dr. Babasaheb Ambedkar Marathwada University, Department of CS and IT, Chh. Sambhajinagar, India
Email: rautaarti443@gmail.com, dehare sheetal04@gmail.com, cnamrata.csit@bamu.ac.in

***Abstract*— Part of Speech (POS) tagging is important in numerous Natural Language Processing (NLP) applications, including Automated Short Answer Evaluation (ASAE). Most traditional approaches assume that semantically identical concepts receive identical POS tags despite variations in context. However, this principle does not apply to morphologically diverse languages such as Marathi, where the same concept could exist in different grammatical forms. In this paper, POS divergence in ASAE systems is investigated via a neural UD approach based on the Stanza NLP library. POS divergence is described as a situation where identically spelled or semantically similar words receive different POS tags due to morphological variations. For the research, a manually curated short answer dataset with one-word model responses and students' replies in Marathi language is utilized. According to the results obtained via experimental testing, POS divergence occurs in most cases of Marathi ASAE. More precisely, around 23% of lemma-based matching showed POS divergence, whereas a significantly higher rate was noted among semantically accurate responses. Upon closer examination of manually processed data, it becomes clear that most differences were linguistically justified, being the result of the grammar change in the context, rather than incorrect tagging. Consequently, strict dependency on POS matching for evaluation purposes leads to false negative scores, which means that the system rejects the response, even if it is correct. Therefore, Marathi ASAE requires alternative methods of response evaluation that include not only grammatical but also semantic analysis.**

***Index Terms*—Part of Speech (POS) Divergence, Marathi NLP, Stanza, Universal Dependencies (UD), Automatic Short Answer Evaluation systems (ASAE), Morphological Linguistics.**

I. INTRODUCTION

Automated Short Answer Evaluation (ASAE) systems have become increasingly important in digital learning environments, online assessments, and intelligent tutoring systems. Unlike objective questions, short-answer evaluation is challenging because students can express the same concept using different words, grammatical structures, and sentence formulations. Therefore, effective ASAE systems must consider both semantic and linguistic variations when evaluating responses [1]. Most existing ASAE approaches rely on lexical matching, lemma matching, semantic similarity, or syntactic features. These methods generally assume that semantically equivalent words appearing in model answers and student responses will exhibit identical grammatical characteristics.

Grenze ID: 01.GIJET.12.2.687_26

However, this assumption is often invalid for morphologically rich languages such as Marathi, where the same semantic concept can be expressed through multiple grammatical forms due to inflectional morphology, derivational processes, and contextual language usage. One important phenomenon arising from such linguistic variability is Part-of-Speech (POS) divergence, where semantically equivalent or lemma-matched words receive different POS tags depending on their contextual grammatical role. With the adoption of neural POS taggers based on the Universal Dependencies (UD) framework, contextual grammatical analysis has become more accurate, but it has also revealed frequent POS divergence between model answers and student responses. The recent advances in Marathi Natural Language Processing (NLP), the impact of POS divergence on Automated Short Answer Evaluation remains largely unexplored. Existing evaluation systems often enforce strict POS agreement, which may incorrectly reject semantically correct responses and produce false-negative classifications. Consequently, there is a need to systematically investigate the occurrence of POS divergence and its influence on evaluation accuracy in Marathi ASAE systems. To address this challenge, this study proposes a Neural Universal Dependencies -based framework for analyzing POS divergence in Marathi Automated Short Answer Evaluation. The framework integrates lemmatization, contextual POS tagging using the Stanza NLP toolkit, lemma-based matching, and divergence analysis to identify grammatical variations between model answers and student responses. The study further quantifies POS divergence frequency, analyzes grammatical transformation patterns, and evaluates their impact on automated assessment performance. The findings contribute toward the development of divergence-aware ASAE frameworks capable of handling grammatical variability while preserving semantic correctness in morphologically rich languages.

II. Related Work

Marathi NLP has experienced significant growth despite being a low-resource language with complex morphology and rich inflectional structures. Early research focused on rule-based approaches for tasks such as POS tagging, while recent studies have increasingly adopted machine learning and deep learning techniques supported by stemming and lemmatization methods for improved linguistic processing [1][2][3]. However, the scarcity of annotated datasets remains a major challenge, limiting the development of robust Marathi NLP systems compared to high-resource languages such as English [4]. To address this issue, researchers have developed language resources including stopword lists, annotated corpora, and transformer-based models for applications such as sentiment analysis, Named Entity Recognition (NER), and hate speech detection [5][7][8]. Recent advancements have been driven by large-scale datasets and transformer architectures. Resources such as L3Cube-MahaCorpus have enabled the development of pretrained models including MahaBERT, MahaAlBERT, and MahaRoBERTa, which have achieved strong performance in text classification, sentiment analysis, and NER tasks [10][11]. Domain-specific datasets such as MahaNER, MahaHate, and MahaSocialNER further support Marathi language applications [12][13][14][15]. For code-mixed Marathi-English text, resources such as MeCorpus along with MeBERT and MeRoBERTa have been introduced [16], while sentence-level semantic models such as MahaSBERT and HindSBERT improve semantic similarity and matching tasks [17]. Additional efforts include sentiment lexicon development and parameter-efficient transformer training approaches for Marathi text processing [18]. The emergence of transformer-based architectures has transformed Marathi NLP by shifting from traditional rule-based systems to deep learning frameworks. Research has also explored techniques such as Chain-of-Translation Prompting (CoTR) for low-resource language processing [9] and long-range named entity recognition for complex Marathi documents. Despite these advances, challenges related to data scarcity, linguistic diversity, tokenization, and resource availability continue to hinder progress [9]. Consequently, further research is required to develop robust Marathi NLP systems and address language-specific challenges in tasks such as semantic analysis, text classification, summarization, and automated evaluation.

Existing Marathi NLP research has primarily focused on POS tagging, sentiment analysis, named entity recognition, text classification, and semantic representation learning. Similarly, Automated Short Answer Evaluation systems have largely emphasized lexical matching, semantic similarity computation, and embedding-based scoring approaches. However, current literature reveals three important gaps,Existing ASAE systems rarely analyze the effect of contextual grammatical variation on answer evaluation,The impact of Universal Dependencies based contextual POS tagging on evaluation outcomes remains unexplored for Marathi,No prior study has quantitatively measured POS divergence between model answers and student responses or investigated its contribution to false-negative classifications. Consequently, the relationship between grammatical variation and evaluation accuracy remains insufficiently understood in Marathi ASAE systems. Addressing this gap is essential for developing linguistically informed and robust evaluation frameworks capable of handling morphological complexity.

III. PROPOSED METHODOLOGY

The proposed methodology focuses on analyzing POS divergence in Marathi short answer evaluation, especially in cases where a one-word model answer appears within a descriptive student response but receives a different POS tag during neural parsing. Since Marathi is morphologically rich, semantically similar responses often show grammatical variation because of inflectional changes, derivational forms, and contextual sentence usage. To study this phenomenon, the methodology uses the Stanza NLP toolkit based on the Universal Dependencies framework. Unlike traditional rule-based systems, the neural parser assigns POS tags according to the contextual grammatical role of a word rather than treating words as fixed lexical units. This allows the system to capture structural and grammatical variation between model answers and student responses. The overall framework was developed as a multi-stage processing pipeline consisting of dataset preparation, Preprocessing, neural POS tagging, divergence detection, and result analysis. During experimentation, we observed that many semantically correct student responses contained contextual grammatical transformations even when the core meaning remained unchanged. Therefore, the methodology was designed to systematically identify such divergence patterns and examine their impact on automated short answer evaluation.

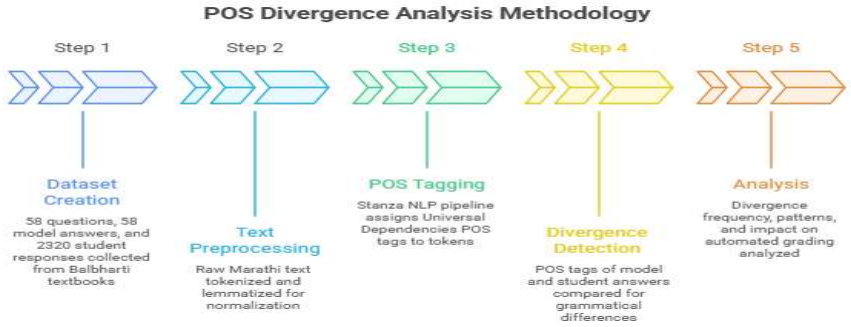

Figure 1: POS Divergence Analysis Methodology

Step 1: Marathi Short Answer Dataset: The methodology begins with the construction of a Marathi short answer dataset, specifically curated for POS divergence analysis in automated short answer evaluation. The dataset is derived from Balbharati textbooks for primary education (standards 2nd to 4th).A key characteristic of the dataset is that model answers consist of single-word reference terms, while student responses are typically short descriptive phrases or one-line answers. This structural difference introduces natural grammatical variation, where the model answer may appear in an expanded or morphologically transformed form within the student response. This dataset design supports POS divergence analysis by enabling the examination of semantically identical content expressed through different grammatical structures.

Step 2: Text Preprocessing:The Preprocessing stage converts raw Marathi text into a normalized form suitable for linguistic analysis. This stage primarily consists of Tokenization and lemmatization, which together reduces variability caused by morphological complexity and prepare the data for accurate POS tagging and comparison.

TABLE I: TEXT PREPROCESSING TECHNIQUE WITH EXAMPLE

Technique	Description	Example
Tokenization	Tokenization is the process of dividing a sentence into smaller meaningful units called tokens, which typically correspond to individual words. This step converts raw text into a structured format suitable for NLP processing.	Sentence: बुधवार पेठ पुणे येथे आहे Tokenization Output: [बुधवार, पेठ, पुणे, येथे, आहे[
Lemmatization	Lemmatization is the process of reducing a word to its base or root form (lemma). This is particularly important in Marathi due to its rich morphological structure, where words change form based on grammatical context.	Token:मुलांना →Lemma:मुल, मुलांना =आंना + मुल Here, "आंना "indicates plural and dative case (to/for children).

Step 3: Neural UD based POS Tagging (Stanza): After Preprocessing, the normalized text is processed using the Stanza NLP toolkit, which performs contextual POS tagging based on the Universal Dependencies framework. Unlike traditional rule-based taggers, the Stanza neural model assigns POS tags according to the syntactic role of words within the sentence context. Therefore, identical words may receive different POS tags when used in different grammatical environments. Some common POS tags include: NOUN – Noun ,VERB – Verb ,ADJ –

Adjective ,ADV – Adverb ,ADP – Adposition,PROPN-Proper Noun. Following POS tagging, the grammatical annotations generated by the Stanza NLP toolkit model are extracted and organized into a structured representation for both model answers and student responses. Each token is represented using, (Token, Lemma, POS) format.

Where, Token represents the original word in the sentence, Lemma denotes the root or base form of the word, and POS indicates the grammatical category assigned by the UD parser. The extracted POS information for the sentence is shown in Table 2:

TABLE II: EXAMPLE OF MARATHI LANGUAGE USING NEURAL UD BASED POS TAGGING

Token	Lemma	POS Tag	Morphological /Contextual Complexity
बुधवार	बुधवार	ADV	Contextual ambiguity; used as part of the compound place name "बुधवार पेठ"
पेठ	पेठ	NOUN	Common noun representing locality/market region
पुण्यात	पुणे	NOUN	Locative inflection ("पुणे + त") indicating location
आहे	आहे	AUX	Auxiliary verb used for sentence completion and tense structure

The sentence "बुधवार पेठ पुण्यात आहे" was processed using the Stanza NLP Toolkit based on the Universal Dependencies Framework.The neural model performs contextual grammatical analysis and assigns POS tags according to the syntactic role of each token.

- बुधवार demonstrates contextual complexity because it appears as part of a compound location expression and may receive different POS interpretations depending on sentence usage.
- पुण्यात illustrates morphological richness in Marathi, where the locative suffix "-त" modifies the base lemma "पुणे".
- आहे functions as an auxiliary verb supporting sentence structure and tense.

This example highlights how neural UD-based POS tagging captures contextual and morphological variations in Marathi, which is important for analyzing POS divergence in automated short answer evaluation systems. This stage produces structured, context-aware grammatical annotations for each token. These annotations serve as the foundation for subsequent POS comparison and divergence detection between model answers and student responses.

Step 4: POS Divergence Detection: After extracting the grammatical annotations generated by the Stanza NLP toolkit, the next stage is POS Divergence Detection. In this stage, the POS tags assigned to the model answer and the corresponding tokens in the student response are systematically compared to identify grammatical variations between semantically equivalent lexical items. The objective is to determine whether the same semantic concept is expressed using different grammatical categories due to contextual, morphological, or syntactic variations. To formally identify such cases, the following mathematical definition of POS divergence is introduced.

A. Formal Defination of POS Divergence

Let,

- (M) denote the model answer
- (S) denote the student response
- (Lemma(M)) denote the lemma of the model answer
- (Lemma(S_i)) denote the lemma of the (i^{th}) token in the student response
- (POS(M)) denote the Part-of-Speech tag of the model answer
- (POS(S_i)) denote the Part-of-Speech tag of the corresponding student-response token

A semantic correspondence exists when: Lemma(M) = Lemma(S_i)

A POS match occurs when: Lemma(M) = Lemma(S_i) and POS(M) = POS(S_i)

A POS divergence occurs when: Lemma(M) = Lemma(S_i) and POS(M) ≠ POS(S_i)

This definition enables systematic identification of semantically equivalent lexical items that exhibit different grammatical categories due to contextual, morphological, or syntactic variation.

B. POS Divergence Detection Framework

Based on the formal definition of POS divergence, a rule-based detection framework is developed to identify grammatical mismatches between semantically equivalent lexical items. The framework combines lemma matching and contextual POS comparison using the Stanza Universal Dependencies parser. For each token in the student response, semantic correspondence is first established through lemma matching. Subsequently, the POS categories of the matched lexical items are compared to determine whether the instance represents a POS match or a POS divergence. The complete workflow of the proposed framework is presented in Algorithm 1:

Algorithm 1: POS Divergence Detection Framework
Input:
M = Model Answer
S = Student Response
Output:
Divergence Status and Transformation Pattern
Step 1: Preprocess M and S
a. Tokenize the text
b. Remove unnecessary symbols
c. Perform lemmatization
Step 2: Apply Stanza UD-based POS Tagger
a. Extract lemma and POS tag for M
b. Extract lemma and POS tags for all tokens in S
Step 3: For each token Si in S do
    If Lemma(M) = Lemma(Si) then
        If POS(M) = POS(Si) then
            Mark as POS Match
        Else
            Mark as POS Divergence
            Record Transformation:
            POS(M) → POS(Si)
        End If
    End If
Step 4: Count Total Matched Cases
Step 5: Count Total Divergent Cases
Step 6: Calculate Divergence Frequency
    D = (Number of Divergent Cases / Total Matched Cases) × 100
Step 7: Generate Divergence Analysis Report
End Algorithm

Initially, the system identifies semantically matched words between the model answer and the student response using lemma-based matching after preprocessing and lemmatization. During this process, each token is reduced to its root or base form (lemma), allowing morphologically different surface forms to be treated as semantically equivalent. For example, the tokens "पुण्यात" and "पुणे" are matched because both share the same lemma "पुणे". The POS tags assigned by the Stanza neural UD parser are then compared to determine whether the matched lexical items represent a POS match or a POS divergence according to the formal definition presented in Section A. This condition indicates that although the lexical meaning remains unchanged, the grammatical role of the word differs due to contextual, morphological, or syntactic variation, as defined in the Universal Dependencies framework.

TABLE III: EXAMPLE OF POS DIVERGENCE

Model Answer	POS Tag	Student Answer	POS Tag
पुण्यात	ADV	पुणे	NOUN
पणती	NOUN	पणती	VERB
खारट	ADJ	खारट	NOUN
सडक	NOUN	सडक	VERB
खाऊ	NOUN	खाऊ	VERB

These examples demonstrate that the same lexical item may receive different POS tags depending on sentence context, even when the semantic meaning remains unchanged.

Step 5: Divergence Analysis: After identifying all POS divergence cases, the final stage of the methodology is Divergence Analysis, in which the detected grammatical mismatches are systematically examined to understand their frequency, structural patterns, and impact on automated short answer evaluation. This stage validates the central hypothesis that grammatical mismatch does not necessarily imply semantic incorrectness, particularly in morphologically rich languages such as Marathi. The divergence analysis is divided into three key components:

Frequency Analysis: Frequency analysis quantifies the extent to which POS divergence occurs within the dataset.
It measures the proportion of semantically matched instances in which the same lemma is assigned different POS tags. The divergence frequency is computed as:

$$Divergence\ Frequency = \frac{Number\ of\ Divergent\ cases}{Total\ Matched\ Cases} * 100 \quad (1)$$

Where: Number of Divergent Cases = Total instances where the same lemma has different POS tags and Total Matched Cases = Total instances where the model answer lemma appears in the student response.This metric provides a quantitative measure of the prevalence of grammatical variation in semantically correct answers.
Pattern Analysis: Pattern analysis identifies the most frequent grammatical transformation patterns responsible for POS divergence. It examines how the grammatical category of a lexical item changes between the model answer and the student response.
Common divergence transformation patterns include, NOUN → VERB, VERB → NOUN,ADJ → NOUN, ADV->NOUN,NOUN->ADP,etc.These transformation patterns highlight the inherent grammatical flexibility of Marathi, where the same semantic concept can be expressed through different syntactic forms. Such variations often arise due to processes like Nominalization, verbalization, and morphological inflection, rather than tagging inconsistencies.
Impact Analysis: Impact analysis evaluates the effect of POS divergence on the performance of automated short answer evaluation systems. Specifically, it examines whether semantically correct responses are incorrectly classified due to strict POS matching constraints.The impact is assessed through:

- False Negative Cases – Correct answers wrongly classified as incorrect
- Incorrectly Rejected Responses – Valid responses penalized due to POS mismatch
- Misclassification due to POS Rigidity – Errors caused by strict grammatical matching rules

This analysis demonstrates that rigid POS-based evaluation frameworks can significantly reduce system reliability by failing to accommodate valid grammatical variation. It therefore underscores the necessity of divergence-aware evaluation strategies that integrate both syntactic and semantic information.

IV. RESULT

This section presents the experimental findings obtained from the POS divergence analysis conducted on the Marathi short answer dataset derived from Balbharati textbooks (standards 2nd to 5th) with 58 one word questions and 2320 student responses.
The evaluation focuses on cases where the model answer lemma is present in the student response and examines how grammatical variation affects automated short answer evaluation.The results are organized into three components: Frequency Analysis, Pattern Analysis, and Impact Analysis, providing a comprehensive understanding of POS divergence.

A. POS Divergence Frequency Analysis

The frequency analysis measures how often Part-of-Speech (POS) divergence occurs in semantically matched responses. A total of 1840 matched cases were identified, where the model answer lemma appears in the student response. Among these, 419 cases exhibit POS divergence, where the same lemma is assigned different POS tags by the Stanza NLP toolkit based on the Universal Dependencies framework. The divergence frequency is calculated as:

$$Divergence\ Frequency = \frac{419}{1840} * 100 \approx 22.77\%(\approx 23\%) \quad (2)$$

The observed divergence rate of 23% demonstrates that semantic equivalence in Marathi does not necessarily imply grammatical uniformity, highlighting the prevalence of contextual and morphological variation in student responses.

B. POS Transformation Pattern Analysis

Pattern analysis was conducted to identify the most common grammatical transformations responsible for divergence.
The results show that divergence primarily occurs due to contextual and morphological variations, where words shift their grammatical roles depending on sentence structure. Below table shows the POS Transformation Pattern with their frequency in percentage:

TABLE IV: POS TRANSFORMATION PATTERN

Transformation Pattern	Frequency (%)
VERB->NOUN	34.60
ADV->NOUN	15.99
NOUN->ADP	15.03
ADJ->NOUN	6.44
NOUN->VERB	2.38
VERB->ADP/ADV/ADJ	5.48
NOUN->ADV	4.53
Other Patterns	15.55

The dominance of the VERB→NOUN transformation pattern (34.60%) indicates that nominalization is a major source of POS divergence in Marathi student responses. Similarly, ADV→NOUN, ADJ→NOUN, and NOUN→ADP transformations highlight the influence of contextual usage, category shifts, and morphological case-marking mechanisms. These findings demonstrate the grammatical flexibility of Marathi, where the same semantic concept can be expressed through different syntactic structures. Importantly, such variations represent linguistically valid transformations rather than tagging errors. Consequently, evaluation systems that rely solely on grammatical matching may incorrectly classify semantically correct responses, emphasizing the need to treat semantic and grammatical equivalence as independent dimensions in automated answer evaluation.

C. Impact Analysis of POS Divergence on ASHE

The impact analysis evaluates how POS Divergence affects the performance of Automated Short Answer Evaluation (ASAE) systems for Marathi. Traditional evaluation methods assume that semantically matched words in the model answer and student response will have identical POS tags. However, due to the morphological richness and contextual flexibility of Marathi, semantically correct answers often appear in different grammatical forms. In the experiment, 1840 semantically matched cases were identified, among which 419 cases exhibited POS divergence. These responses were semantically correct but were incorrectly rejected because of POS mismatch, resulting in false negatives.

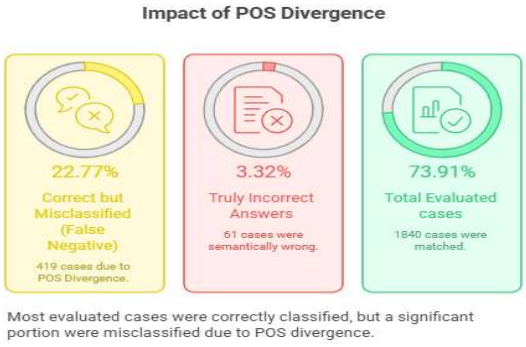

Figure 3: Impact of POS Divergence

The analysis confirms that POS divergence is a linguistically valid phenomenon arising from normalization, verbalization, inflectional morphology, and contextual grammatical shifts in Marathi. These grammatical transformations allow semantically correct responses to appear in different syntactic forms while preserving the same meaning. Therefore, strict POS-based evaluation becomes unreliable for Marathi Automated Short Answer Evaluation (ASAE) systems, as it incorrectly rejects many valid student responses due to grammatical mismatch. The findings emphasize the need for divergence-aware evaluation frameworks that integrate lemma-based matching, flexible POS alignment, and semantic similarity analysis to improve evaluation accuracy and reduce false negatives in morphologically rich languages like Marathi.

V. CONCLUSION

This study investigated the phenomenon of Part-of-Speech (POS) divergence in Marathi Automated Short Answer Evaluation (ASAE) using a neural Universal Dependencies (UD)-based framework implemented through the Stanza NLP toolkit. The objective was to examine how contextual grammatical variation influences automated evaluation when semantically correct responses are expressed using different linguistic structures.

The experimental analysis revealed that POS divergence is a frequent characteristic of Marathi student responses, occurring in approximately 22.77% of lemma-matched cases. The results indicate that semantically equivalent words may receive different grammatical classifications due to contextual usage, morphological variation, and syntactic transformations. The identified divergence patterns, particularly VERB→NOUN and NOUN→ADP transformations, demonstrate the significant role of nominalization and case-marking processes in Marathi. The findings further show that POS divergence is a linguistically valid outcome of context-sensitive grammatical annotation rather than a tagging error. Consequently, evaluation systems that rely solely on strict POS agreement may incorrectly reject semantically correct responses, leading to false-negative classifications and reduced evaluation reliability. Overall, this study highlights the limitations of rigid POS-based evaluation strategies for morphologically rich languages and provides empirical evidence for the need to incorporate contextual linguistic information into Automated Short Answer Evaluation frameworks. The proposed analysis contributes to a better understanding of grammatical variation in Marathi educational assessment and establishes a foundation for divergence-aware evaluation methods. Future work will focus on integrating semantic similarity measures, dependency-based linguistic features, and contextual language representations to develop more robust ASAE systems capable of handling both semantic and grammatical variability in Marathi student responses.

ACKNOWLEDGMENT

The authors extend their gratitude to Chhatrapati Shahu Maharaj Research Training and Human Development Institute (SARTHI) and Dr. Babasaheb Ambedkar Research and Training Institute (BARTI), 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] Sarda, P., Kokare, D., Mokashi, R., Patkar, T., Deshpande, P., &Jahirabadkar, S. (2025). "ShabdaAnveshak": Comparative Review and Analysis of Marathi Part of Speech Tagging Approaches. 2025 1st International Conference on AIML-Applications for Engineering & Technology (ICAET).
[2] Deore, P. R., Patil, N. V., &Patil, A. S. (2025). Deep Learning-Based Parts-of-Speech Tagging in Marathi Language. Procedia Computer Science.
[3] Kadam, P. V., Khandale, B. K., & Mahender, C. N. (2022). Design and Development of Marathi Word Stemmer. Proceedings of Second International Conference on Advances in Computer Engineering and Communication Systems.
[4] Dani, A., &Sathe, S. R. (2024). A Review of the Marathi Natural Language Processing.
[5] Joshi, R. (2022). L3Cube-MahaNLP: Marathi Natural Language Processing Datasets, Models, and Library. arXiv-Computation and Language.
[6] Deshmukh, P., Kulkarni, N., Kulkarni, S., Manghani, K., & Joshi, R. (2024). L3Cube-MahaSum: A Comprehensive Dataset and BART Models for Abstractive Text Summarization in Marathi. arXiv-Computation and Language.
[7] Chavan, R., Patil, G., Madle, V., & Joshi, R. (2024). Curating Stopwords in Marathi: A TF-IDF Approach for Improved Text Analysis and Information Retrieval. arXiv-Computation and Language.
[8] Magdum, V., Dhekane, O., Hiwarkhedkar, S., Mittal, S., & Joshi, R. (2023). mahaNLP: A Marathi Natural Language Processing Library. arXiv-Computation and Language.
[9] Deshpande, T., Kowtal, N., & Joshi, R. (2024). Chain-of-Translation Prompting (CoTR): A Novel Prompting Technique for Low Resource Languages. arXiv-Computation and Language.
[10] Deshmukh, P., Kulkarni, N., Kulkarni, S., Manghani, K., & Joshi, R. (2024). Leveraging Parameter Efficient Training Methods for Low Resource Text Classification: A Case Study in Marathi.
[11] Joshi, R. (2022). L3Cube-MahaCorpus and MahaBERT: Marathi Monolingual Corpus, Marathi BERT Language Models, and Resources. arXiv-Computation and Language.
[12] Patil, P., Ranade, A., Sabane, M., Litake, O., & Joshi, R. (2022). L3Cube-MahaNER: A Marathi Named Entity Recognition Dataset and BERT models. arXiv-Computation and Language.
[13] Velankar, A., Patil, H., Gore, A., Salunke, S., & Joshi, R. (2022). L3Cube-MahaHate: A Tweet-based Marathi Hate Speech Detection Dataset and BERT models. arXiv-Computation and Language.
[14] Chaudhari, H., Patil, A., Lavekar, D., Khairnar, P., & Joshi, R. (2023). L3Cube-MahaSocialNER: A Social Media based Marathi NER Dataset and BERT models. arXiv-Computation and Language.
[15] Pingle, A., Vyawahare, A., Joshi, I., Tangsali, R., & Joshi, R. (2023). L3Cube-MahaSent-MD: A Multi-domain Marathi Sentiment Analysis Dataset and Transformer Models. arXiv-Computation and Language.

[16] Chavan, T., Gokhale, O., Kane, A., Patankar, S., & Joshi, R. (2023). My Boli: Code-mixed Marathi-English Corpora, Pretrained Language Models and Evaluation Benchmarks. arXiv-Computation and Language.
[17] Joshi, A., Kajale, A., Gadre, J., Deode, S., & Joshi, R. (2022). L3Cube-MahaSBERT and HindSBERT: Sentence BERT Models and Benchmarking BERT Sentence Representations for Hindi and Marathi. arXiv-Computation and Language.
[18] Kulkarni, P. V., &Thakre, K. S. (2024). Developing sentiment lexicon for Marathi: A comprehensive survey and analysis. Journal of Information and Optimization Sciences.