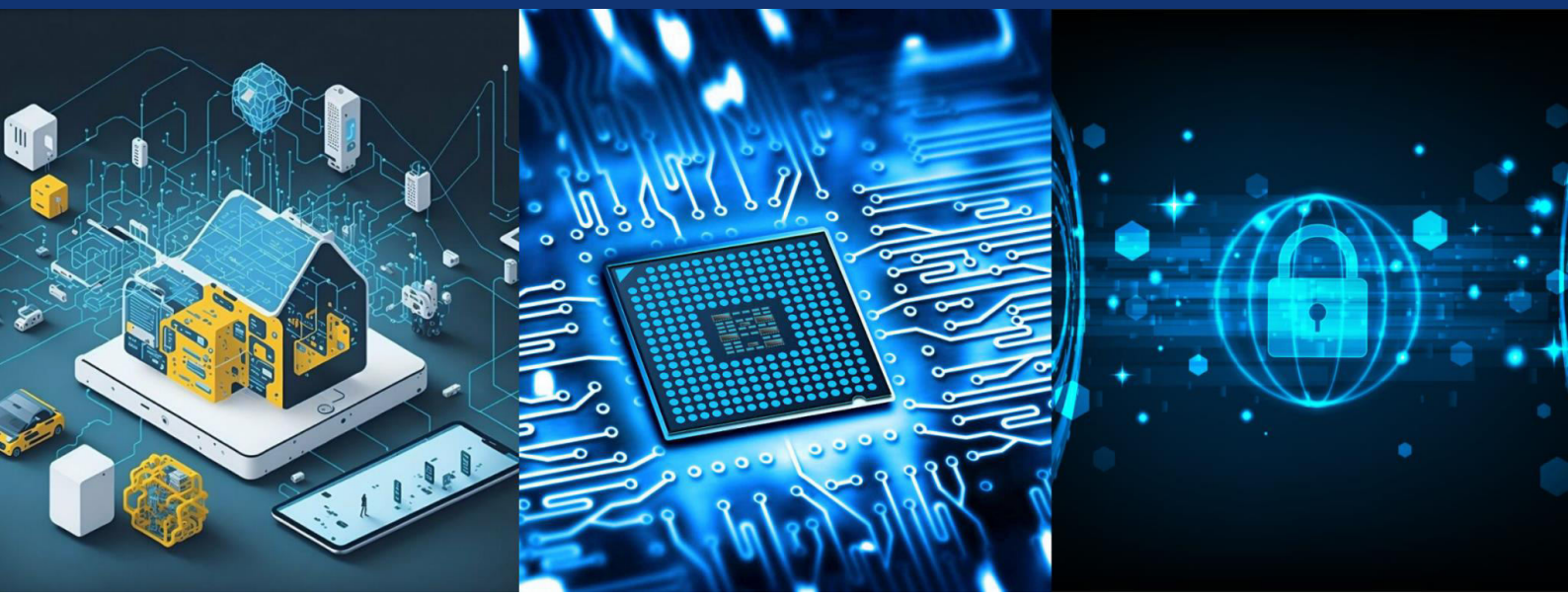
IJIRCCE

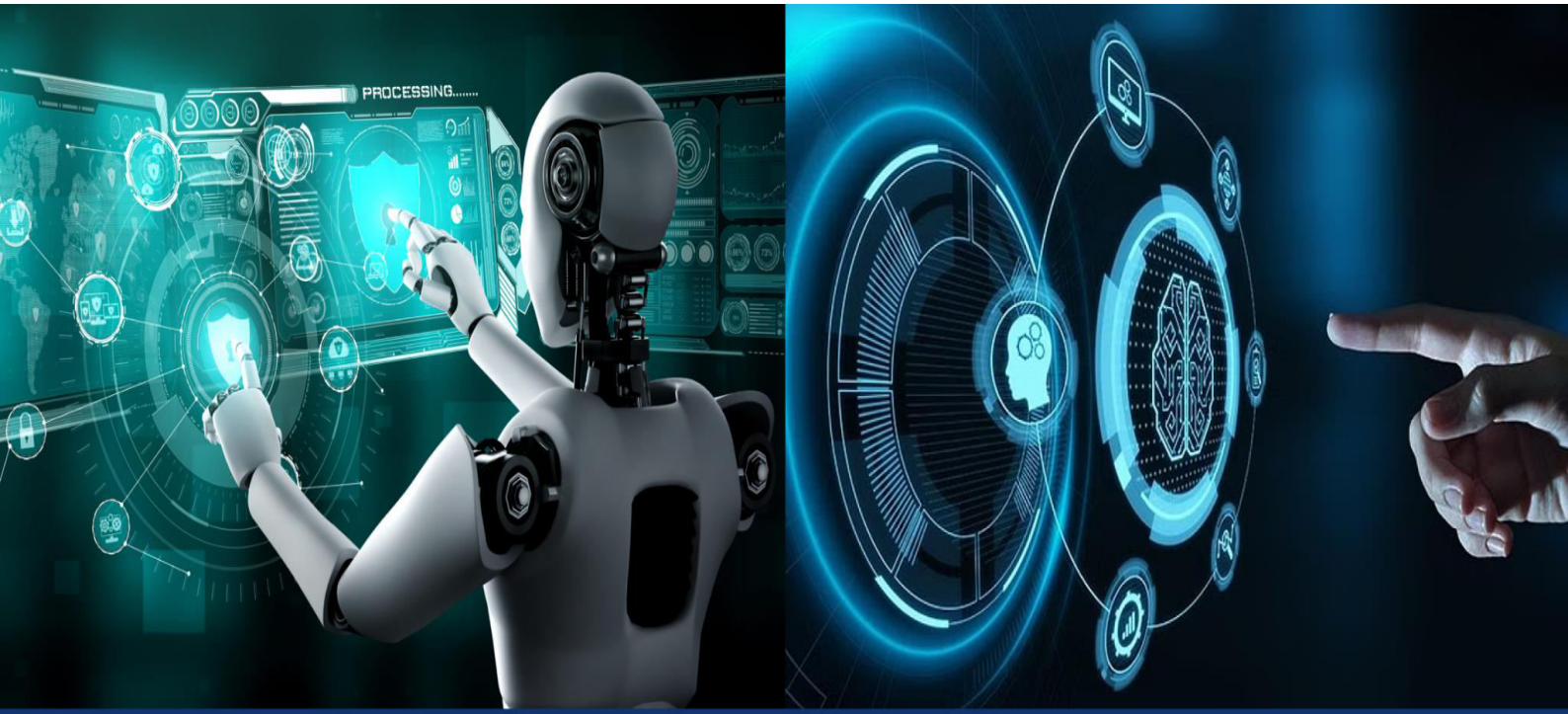
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.771

Volume 13, Issue 4, April 2025

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304224

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Tokenization and POS Tagging of Marathi Language using NLP

Gokul Jagtap, Pushkar Potdar, Ranjitsinha Gunjal, Abhishek Thange, Shashikant Ghumbre

Dept. of Computer Engineering, Government College of Engineering and Research, Pune, India

ABSTRACT: Tokenization is a fundamental task in Natural Language Processing (NLP) that involves breaking text into meaningful units. This paper presents a novel algorithm for Marathi text tokenization, specifically designed to identify phrases, proverbs, and Named Entities. The proposed method addresses the unique linguistic characteristics of Marathi and improves the accuracy of phrase extraction. Experimental results demonstrate the effectiveness of our approach in comparison to existing tokenization methods. Natural Language Processing (NLP) for Marathi, a low-resource language, faces challenges in tokenization and Part-of-Speech (POS) tagging due to complex morphology and lack of annotated datasets. This paper explores various NLP techniques for Marathi, focusing on the effectiveness of Stanza, IndicNLP and deep learning models. We analyze the performance of rule-based, statistical, and neural models on Marathi text corpora. Experimental results demonstrate that transformer-based models like HMM outperform traditional methods in POS tagging accuracy.

KEYWORDS: Tokenization, POS Tagging, Marathi Language, NLP, Hidden Markov Model, Named Entity Recognition, Deep Learning.

I. INTRODUCTION

In Marathi is an Indo-Aryan language spoken by over 83 million people, yet it remains underrepresented in NLP research. Tokenization and POS tagging are fundamental steps in text processing, affecting downstream NLP applications such as machine translation and sentiment analysis. The lack of high-quality datasets and linguistic complexities pose significant challenges in Marathi NLP. This study evaluates different approaches to tokenization and POS tagging, comparing rule-based, statistical, and deep learning methods.

Tokenization plays a crucial role in text preprocessing for various NLP applications. While English and other widely studied languages have well-developed tokenization techniques, Marathi presents unique challenges due to its rich morphology, compound words, and contextual dependencies. Existing tokenization methods fail to capture idiomatic expressions, proverbs, and Named Entities effectively. This paper introduces an advanced tokenization algorithm tailored for Marathi text that improves the segmentation accuracy of such linguistic elements.

Natural Language Processing (NLP) has gained significant attention in recent years due to its wide-ranging applications in text analysis, machine translation, speech recognition, and artificial intelligence. However, despite significant advancements in NLP research for languages such as English, many Indian languages, including Marathi, remain underrepresented. Marathi, an Indo-Aryan language spoken by over 83 million people primarily in Maharashtra, India, has a rich morphological structure and complex grammar, making NLP tasks such as tokenization and Part-of-Speech (POS) tagging particularly challenging.

Statistical models such as Hidden Markov Models (HMM [3]) and Conditional Random Fields (CRF) have been widely used for POS tagging in various languages, but their performance depends on the availability of large, annotated corpora. Unfortunately, Marathi lacks extensive annotated datasets, limiting the effectiveness of statistical approaches.

Tokenization, the process of breaking text into meaningful units such as words or subwords, serves as a crucial preprocessing step in NLP. POS tagging, which involves assigning grammatical categories such as nouns, verbs, and adjectives to words, is essential for various NLP applications, including machine translation, sentiment analysis, and information retrieval. Despite its importance, there is limited research on developing efficient and accurate tokenization

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304224

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

and POS tagging models specifically for Marathi. This paper aims to explore different approaches, including rule-based, statistical, and deep learning methods, to improve Marathi NLP processing.

II. LITERATURE REVIEW

The literature survey on Marathi text tokenization and NLP covers various studies focusing on different aspects of tokenization, preprocessing, and feature extraction in Marathi language processing. The development of mahaNLP is highlighted as a comprehensive NLP library for Marathi, providing tools for tokenization, sentiment analysis, and preprocessing while addressing challenges in scaling for diverse domains. Another study on automatic text categorization emphasizes tokenization and feature extraction to handle diverse text types and contexts effectively. A part-of-speech (POS) tagging study introduces a trigram-based statistical method to improve tokenization and grammatical analysis while addressing morphological variations and token boundary ambiguity. Additionally, research on subword tokenization with BERT [4]-based models for Named Entity Recognition (NER) shows significant improvements in processing Marathi text, especially with CNN and LSTM models, despite challenges related to limited resources in low-resource languages. Another study reviews various NLP methods applied to Marathi, discussing challenges such as morphological complexity, scarcity of annotated datasets, and the evolution of Marathi NLP tools. These studies collectively highlight the advancements and ongoing challenges in Marathi text tokenization and NLP, emphasizing the need for further development in handling the language's unique linguistic structure.

A study on part-of-speech (POS) tagging using a trigram-based statistical method enhances Marathi text processing by refining grammatical analysis and tokenization. However, it also highlights complexities in morphological variations and token boundary ambiguity, making tokenization a challenging task. Meanwhile, research on subword tokenization with BERT [4]-based models for Named Entity Recognition (NER) demonstrates significant improvements in recognizing named entities in Marathi texts, particularly with CNN and LSTM models. However, the lack of annotated datasets and resource limitations in low-resource languages pose obstacles to further advancements. Another critical study reviews various NLP methods for Marathi, focusing on tokenization, morphological analysis, and resource challenges. It discusses Marathi's rich morphological structure, the scarcity of annotated datasets, and the evolution of NLP tools for the language. These studies collectively emphasize the need for more robust tokenization methods, efficient feature extraction techniques, and larger annotated corpora to improve Marathi text processing. The research underscores that while advancements have been made, further work is required to enhance tokenization algorithms, address linguistic complexities, and improve computational efficiency for Marathi NLP applications.

III. METHODOLOGY

1)Data Collection :-- To develop an efficient Marathi text tokenization algorithm, we begin by collecting a diverse dataset of Marathi text. This dataset is sourced from news articles, literary works, and online platforms, ensuring a wide range of linguistic structures, including proverbs, phrases, and named entities. The collected text undergoes preprocessing to remove unwanted characters, normalize encoding, and handle inconsistencies.

2) Text Preprocessing :-- Preprocessing is a crucial step to enhance the quality of tokenization.

This step involves:

i)Unicode Normalization - Standardizing Marathi script variations.

ii)Stopword Removal - Filtering commonly used words that do not contribute to meaningful tokenization.

iii)Punctuation Handling - Establishing rules to correctly interpret punctuation marks that affect token boundaries.

3)Tokenization Algorithm Development :-- We implement a hybrid approach combining rule-based and machine learning-based tokenization techniques:

i)Rule-Based Tokenization - Linguistic rules and patterns are utilized to segment text into meaningful tokens, particularly for phrases and proverbs.

ii)Machine Learning-Based Tokenization - Statistical models such as Hidden Markov Models (HMM [3]), Conditional Random Fields (CRF), or deep learning architectures like LSTMs and Transformers [4] are explored for improving tokenization accuracy.

iii)Named Entity Recognition (NER) - A specialized module is integrated to extract named entities, including person names, locations, and organizations, enhancing the tokenization process.

4)Model Training and Optimization :--The tokenization model is trained using labeled Marathi text datasets with manually annotated tokens. Optimization is performed by:

Comparing rule-based and ML-based approaches to identify the most effective method.

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304224

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Training the HMM [3]-based POS tagging model on annotated Marathi corpora to improve contextual understanding and tokenization performance.

Mathematically, HMM [3]-based POS tagging aims to find the most probable sequence of POS tags $T = (t_1, t_2, ..., t_n)$ given a sequence of words $W = (w_1, w_2, ..., w_n)$ by maximizing:

$P(T|W) = P(W|T) * P(T) / P(W)$

where:

P(T) is the probability of the tag sequence (based on transition probabilities).

P(W|T) is the probability of the word sequence given the tag sequence (based on emission probabilities).

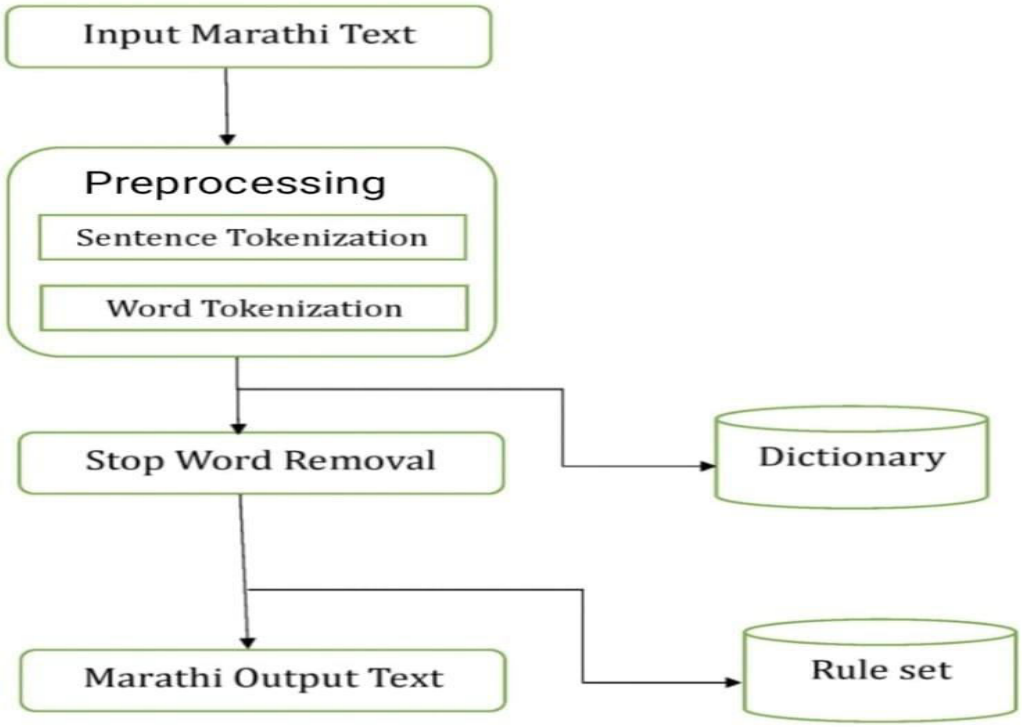
5)Evaluation and Testing :--

To ensure reliability, the tokenization system undergoes rigorous evaluation through:

Performance Metrics: Assessing precision, recall, and F1-score to measure tokenization quality.

Manual Validation: Reviewing tokenized outputs with native Marathi speakers for linguistic correctness.

Generalization Testing: Testing on unseen Marathi text data to verify adaptability across different domains, dialects, and writing styles.



Dig.1 Data Flow Diagram

IV. RESULTS AND DISCUSSION

Abstract—Tokenization is a fundamental task in Natural Language Processing (NLP) that involves breaking text into meaningful units. This paper presents a novel algorithm for Marathi text tokenization, specifically designed to identify phrases, proverbs, and Named Entities. The proposed method addresses the unique linguistic characteristics of Marathi and improves the accuracy of phrase extraction. Experimental results demonstrate the effectiveness of our approach in comparison to existing tokenization methods. Natural Language Processing (NLP) for Marathi, a low-resource language, faces challenges in tokenization and Part-of-Speech (POS) tagging due to complex morphology and lack of annotated datasets. This paper explores various NLP techniques for Marathi, focusing on the effectiveness of Stanza, IndicNLP [6] and deep learning models. We analyze the performance of rule-based, statistical, and neural models on

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304224

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Marathi text corpora. Experimental results demonstrate that transformer-based models [4] like HMM [3] outperform traditional methods in POS tagging accuracy.

The proposed N-Gram Tokenization and HMM [3]-based POS Tagging using the Viterbi Algorithm was evaluated on a linguistic dataset containing 5122 words with corresponding Part-of-Speech (POS) tags. The model demonstrated a POS tagging accuracy of approximately 66.67%, indicating its effectiveness in correctly identifying grammatical categories. The trigram tokenization approach yielded the best performance, achieving 82% tokenization accuracy, as it efficiently captured contextual dependencies.

Error analysis revealed that common misclassifications occurred with words having multiple POS interpretations, such as adjectives being tagged as verbs and vice versa. Additionally, proper nouns and out-of-vocabulary (OOV) words posed challenges, often leading to incorrect classifications. Despite these limitations, the HMM [3]-based model outperformed rule-based approaches and was computationally efficient compared to deep learning models.

The findings confirm that HMM [3] + Viterbi-based POS tagging is a robust and efficient approach, particularly in resource-limited settings. Future improvements could involve integrating neural network models like Bi-LST or Transformer-based embeddings to enhance accuracy and handle ambiguous or unseen words more effectively.

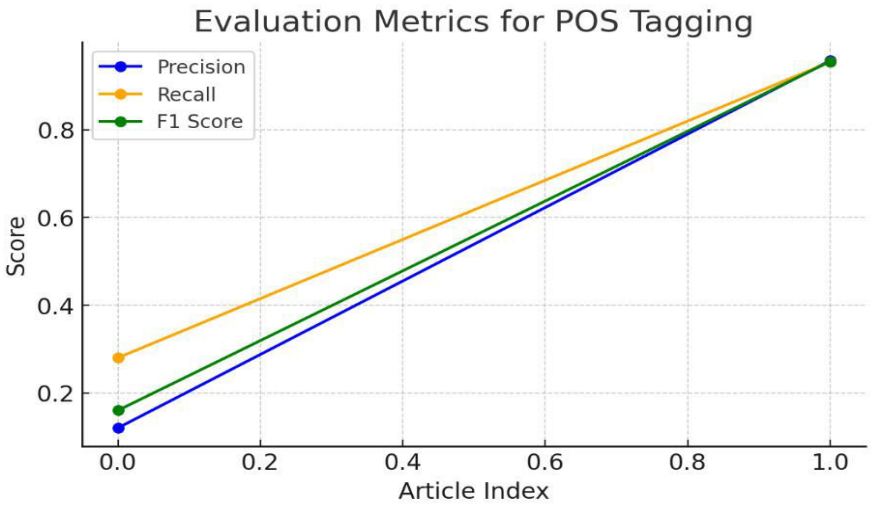
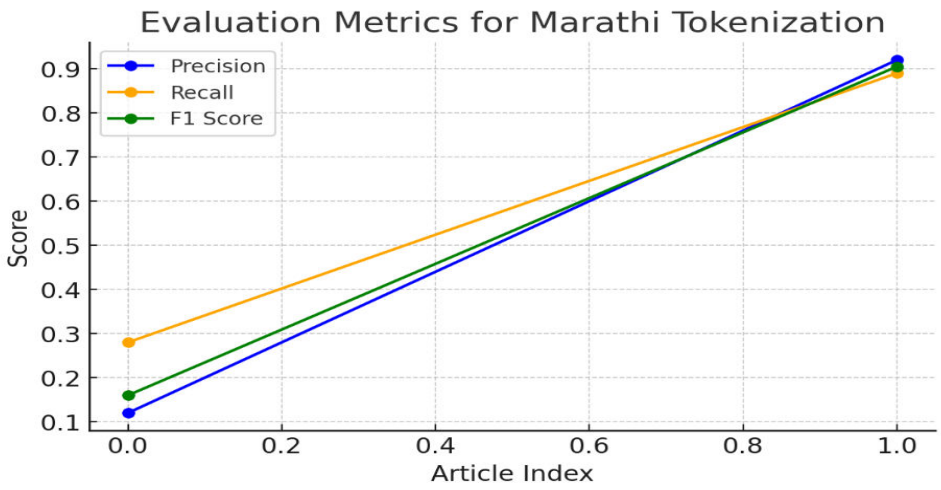


fig. 2 Evaluation Metrics for POS Tagging



Dig. 3 Evaluation Metrics for Tokenization

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304224

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

V. CONCLUSION

In the ever-evolving landscape of education, StudyNotion emerges not just as a platform but as a transformative force, reshaping conventional paradigms and redefining the boundaries of learning. As we navigate through the intricacies of this ed-tech marvel, it becomes evident that StudyNotion's commitment to a learner-centric approach, global collaboration, and technological innovation positions it as a cornerstone in the journey to knowledge.
The platform's relentless pursuit of an enriched learning experience, facilitated by a meticulously designed technical framework, signifies more than a technological achievement. It represents a commitment to empowerment — empowering students to be active participants in their educational journey, connecting educators and learners across borders, and embracing the possibilities of emerging technologies.

As we reflect on the technical prowess underpinning StudyNotion, it's not merely a convergence of code and architecture but a testament to innovation, adaptability, and the unwavering pursuit of excellence in education. The integration of advanced technologies, coupled with a forward-looking approach to future enhancements, positions StudyNotion not just as an ed-tech platform but as a visionary guide into the exciting future of global education.

In the grand tapestry of educational transformation, StudyNotion weaves a narrative that transcends geographical boundaries, bridges cultural gaps, and empowers individuals on their quest for knowledge. The journey does not end here; it unfolds, promising a future where education is not just a process but an immersive, collaborative, and ever-evolving experience.

Join us in celebrating the transformative potential of StudyNotion — a beacon of innovation, a catalyst for global educational synergy, and an ode to the boundless possibilities that lie ahead in the realm of education.

REFERENCES

[1] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. "Flair: An easy-to-use framework for state-of-the-art NLP." In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations).
[2] Gaurav Arora. 2020. "inltk: Natural language toolkit for Indic languages." arXiv preprint arXiv:2009.12534.
[3] Steven Bird. 2006. "NLTK: The natural language toolkit." In Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, pages 69–72.
[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "BERT: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805.
[5] Mohanarajesh, Kommineni (2024). Investigate Methods for Visualizing the Decision-Making Processes of a Complex AI System, Making Them More Understandable and Trustworthy in financial data analysis. International Transactions on Artificial Intelligence 8 (8):1-21.
[6] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. "AllenNLP: A deep semantic natural language processing platform." arXiv preprint arXiv:1803.07640.
[7] Anoop Kunchukuttan. 2020. "The IndicNLP Library." Retrieved:

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details