iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** XI **Month of publication:** November 2025

DOI: https://doi.org/10.22214/ijraset.2025.75099

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XI Nov 2025- Available at www.ijraset.com

AI-Based Grammatical Error Correction System for Native Language

Gaurav G. Kankuse[1], JayC. Deshmukh[2], Om S.Borse[3], Prof. N.D. Dhamale[4]
Dept. Artificial Intelligence and Data Science MET's Institute of Engineering Nashik, India

Abstract: ***Thisproject focusesonbuilding a smart,easy- to-use Grammatical Error Correction (GEC) system for anativeIndiclanguage,specificallyMarathi.We'reusing powerful, modern AI models called transformer-based models (like IndicBERT and mBART), which we'll fine- tunewithlocallanguagedata.Themainjobofthesystem is to spot and fix grammar mistakes in sentences in real- time.Thefinishedproductwillbeasimplewebtoolwhere users can type their text, see suggested fixes with quick explanations, and choose which changes to accept or reject. This work is essential because it fills a big gap, making writing more accurate and accessible in local languagesthatcurrentlydon'thavegoodgrammartools.***
Keywords: Grammar Correction, Native Language, Marathi,TransformerModels,IndicBERT,mBART,mT5, Natural Language Processing, Rule-Based Correction, Web Application, Text Improvement, Low-Resource Languages.

I. INTRODUCTION

A. Motivation and Context

The increasing reliance on digital tools for collaboration highlights a significant need for better language support, particularlyforregionalandnativelanguagesthatareoften ignored. Most advanced grammar checking software isbuilt and optimized for English, leaving languages like Marathi withouteffective,modernassistance.Nativelanguages,such as Marathi, present unique linguistic complexities like intricatesyntaxandmorphology, making automated correction challenging. The absence of effective tools leads to more errorsinwrittencommunicationand can actually slow down the adoption of digital technologies by native speakers.OurmotivationisdrivenbythepotentialofAIand transformer-based models (like IndicBERT and mBART) to overcome these hurdles in low-resource.

B. Research Problem Statement

Thesignificantchallengeaddressedbythisprojectisthelack of intelligent support for native language users in digital communication. Most existing grammar correction tools are primarilydevelopedandoptimizedfortheEnglish language, offering little tonosupport for nativeorregional languages. Thissituationcreatesamajoraccessibilityandutilitygapfor native speakers. Furthermore, traditional solutions, such as manual proofreading, are inherently time-consuming, prone to errors, and often inconsistent. While traditional rule-based tools exist, they consistently struggle to handle thecomplex sentence structuresandcontext-sensitive correctionsrequired by languages like Marathi. Consequently, there is an urgent and growing need for a smart, user-friendly grammar correction system tailored specifically to address the unique linguistic complexities of native language communication.

C. Contributions

Our proposed system makes several key contributions to the field of low-resource Grammatical Error Correction (GEC). Firstly, we directly address the issue oflimited specialized tools for Marathi by building adedicated GEC engine based on powerful transformer models (IndicBERT/mBART). Secondly, we are tackling thelow- resourcedata challengeby focusing on fine-tuningmodels using regional corpora and, critically, by generating synthetic training databecause large annotated parallel datasetsarenotavailableforMarathi.Ourprimarytechnical contribution is introducing aProposed Hybrid Model, which uniquely combines the high precision of a rule-based checker for deterministic errors (like spelling and simple agreement) with the contextual processing power of a TransformerMTModel(Seq2Seq) for complexerrors(like tense and word order). This hybrid methodology fills the existinglack of hybrid approachesstudied for Marathi GEC. Finally, we contribute a solution to thepractical usability gapbydeploying the system asa real-time, user- friendly web interface, making high-quality grammar correction accessible to native Marathi speakers.

D. Organizationofpaper

Therestofthisreportisstructured todetailthetechnical foundation and execution of the Grammatical Error Correction (GEC) project. Following this introduction, we outlinethedifferentapproachesinSectionII,RelatedWork and Literature Analysis, including a review of existing methodologiesandasummaryoftheidentifiedresearchgap we aim to fill.Section IIIthen fully details theProposed System, explaining the hybrid architecture that combines rule-based checking with the advanced Transformer model. Subsequently,Section IVoutlines the comprehensive Methodology of Evaluation, covering dataset preparation, the automatic metrics used, and the human evaluation process. Finally,Section Vdiscusses theProject Achievability, risk analysis, and expected outcomes, confirming the viability of the system within the specified time and cost budgets.

II. RELATED WORK AND LITERATURE ANALYSIS

The design of this Grammatical Error Correction (GEC) projectisinformedbyananalysisofrecent,advancedneural- based GEC systems across various global languages. This review highlights the feasibility of our approach while identifying key gaps that our system addresses.

Research focused on GEC technology for Chinese Documents compared rule-based, statistical, and neural methodologies, confirming the superior performance of Transformer-based models over traditional methods. However,italsonotedthatchallengespersist,mainlydueto limited linguistic corpora and complex word segmentation issues. Separately, an evaluation of AI-based Grammar Correction for Portuguese assessed the capabilities of large models like ChatGPT (GPT-3.5/4). While the models achievedhighprecision,withGPT-4yieldingthebestresults,thestudypointedoutlimitationssuchasovercorrectionand themisinterpretationofcontextualnuances.

Finally, the study of GEC for the low-resource language Zarma provided critical practical insights. This research demonstrated that the Machine Translation (MT)-based method performed optimally, achieving a 95.8% detection and 78.9%accuracy.Thisfinding reinforcedthatchallenges related to scarce data and non-standard orthography remain significanthurdlesinlow-resourcesettings.Thesecollective studiessupportour decision to leverageatransformer-based MT approach for Marathi.

A. GrammaticalErrorCorrectionforLow-Resource Languages: The Case of Zarma (Keita et al., 2025)

ThisresearchfocusedonapplyingGECtechniquestoZarma, a low-resource African language. The study investigated rule- based, machine translation (MT), and large language model (LLM) methods for grammatical correction. Results indicated that MT-based approaches achieved a 95.8% detectionrateand78.9%accuracy,outperformingrule-based systems.

Findings:ThepaperdemonstratedthathybridandMT-based GEC systems can effectively address grammatical issues even in languages with limited annotated corpora.RelevancetoCurrentWork:TheMarathiGECsystemdraws inspiration from this study's hybrid model approach to overcomedatascarcitythroughsyntheticdatagenerationand transfer learning.

B. Research and Analysis of Grammatical Error CorrectionTechnologyforChineseDocuments(Jinetal., 2023)

This study, published in*Scientific Research Publishing*, compared rule-based, statistical, and neural methods for Chinese GEC. Transformer-based architectures exhibited superior performance due to their contextual understanding andabilitytogeneralizeacrosserrortypes. Findings:Transformer models achieved higher recall and precisionthantraditionalmethods,althoughlimiteddatasets andsegmentationissuesposedchallenges. Relevance: This study supports the use of IndicBERT and mT5for Marathi, as both can capture grammatical dependencies in morphologically rich languages.

C. EvaluationofAI-BasedGrammarCorrectionfor Portuguese(Juricić&Sarić,2024)

Conducted at the University of Zagreb, this research evaluated GPT-3.5/4 models for Portuguese learner texts. The models demonstrated high precision and fluency but showed tendencies towardover-correctionand occasional semantic shifts.

Findings:GPT-based systems are highly effective for sentence-level grammar but require careful calibration to preserve intended meaning.

Relevance:Highlights the importance of balancing correction aggressiveness in Marathi GEC using rule-based filters before final output.

D. GrammaticalErrorCorrection:ASurveyoftheState of the Art (Bryant et al., 2023)

Published in*Computational Linguistics (MIT Press)*, this comprehensive survey examined the evolution of GEC systems, datasets, and evaluation metrics. It revealed that manyexistingmetrics(likeBLEUandGLEU)areunreliable for non-English languages and emphasized the need for human evaluation.

Findings:The paper served as an authoritative guide on standardpracticesandevaluationchallenges. Relevance:Informs the evaluation methodology of the currentproject,combiningautomaticmetrics($F_{0.5}$,BLEU) with native speaker validation for reliability.

E. RevisitingMeta-EvaluationforGrammaticalError Correction(Kobayashietal.,2024)

Publishedinthe*TACL,ACLAnthology*,thisstudyintroduced SEEDA,ahuman-ratedmeta-evaluationdatasetdesignedto analyze the reliability of automatic metrics. The authors showedthatstandardmetricsoftenmisrepresentnon-English grammar improvements.

Findings: Human-annotated meta-evaluation produced more consistent results than automated metrics.

F. OrganicData-DrivenApproachforTurkishGEC and LLMs(Ersoy&Yildız,2024)

This paper explored grammar correction for the Turkish language using synthetic data combined with LLM fine- tuning. It showed that even low-resource languages can achieve high accuracy using synthetic parallel corpora. Findings: Demonstrated the effectiveness of data augmentationandhybridLLMapproaches.Relevance: The proposed Marathi GEC system adopts a similar data-driven methodology by generating synthetic error datasets to train the transformer mode**ls.**

G. Research Gap Summary

Despite global advancements, this project addresses several criticalgapsspecifictoGrammaticalErrorCorrection(GEC) for Marathi and similar low-resource Indian languages. A major issue is thelimited availability of specialized tools forMarathi,asexistingsoftwareoftenfocusesonlyonbasic spelling checks, failing to provide comprehensive, full- grammar correction. This problem is compounded by the low-resource challenge, meaning large, high-quality annotated parallel datasets required for training modern models are simply not available for Marathi.

Furthermore,atechnicalgapexistsduetothecurrentlackof hybrid approachesfor Marathi GEC; most existing work relies entirely on either limited rule-based systems or error- proneTransformermodels.Ourprojectisdesignedtofillthis voidbystrategicallycombiningboth.Finally,thereisaclear evaluation gapandpractical usability gap. Most new MarathiGECeffortslackrigorousevaluationusingstandard metricslikeBLEU,andveryfewmodelsevertransitionfrom research papers into real-time, accessible web tools for the community.

III. SYSTEM ARCHITECTURE

1) The proposed system employs a powerful hybrid architecture designed to ensure comprehensive and accurate Grammatical Error Correction (GEC) for Marathi. The flow is organized into sequential and parallel processing stages.

2) The process begins at the Input Module, where the user submits text via the web interface. This input immediately moves to Preprocessing, which is a crucial first step that includes Tokenization, Unicode Normalization (to correctly handle Devanagari script issues), and Part-of-Speech (POS) tagging for foundational grammatical analysis.

3) Thetextthenentersadual-correctionpathway.The Rule-BasedGrammarCheckerdetectsandcorrects basic, deterministic errors like spelling mistakes, subject–verb agreement, and gender/number mismatches. Simultaneously, the core intelligence lies in the Transformer-Based Correction (MT Model). This module uses a Seq2Seq Transformer (e.g., mT5-small or MarianMT) which treats GEC asatranslationtask(incorrectsentence$\rightarrow$ correct sentence). It is specialized to handle complex, context-dependent errors such as tense issues, missing or extra words, and incorrect word order.

4) Finally,theSuggestionEngineandPost-processing phasecombinesthecorrectionsfromboththeRule- Based and MT Transformer modules. It highlights errors in the original text, merges the suggested changes, and prepares the final corrected sentence fortheUserInterface,allowingtheusertheultimate choice to accept or reject the suggestions.

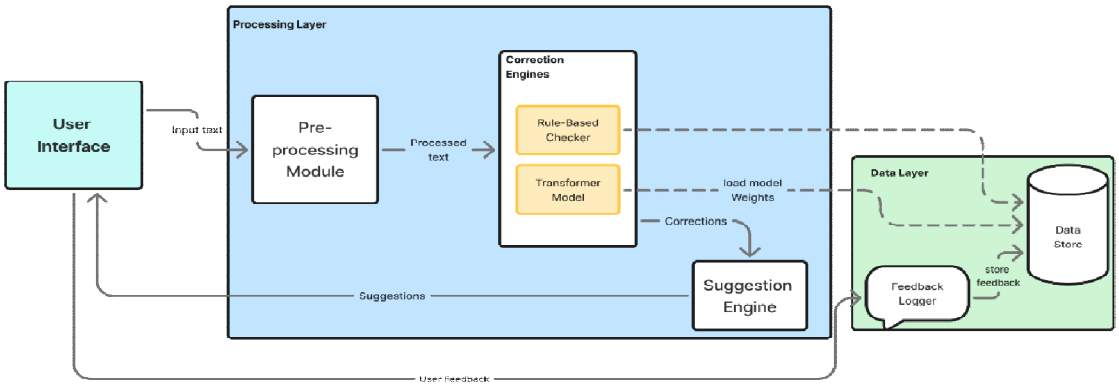


Figure1:GrammaticalErrorCorrectionArchitecture

IV. EVALUATION METHODOLOGY

A. *Evaluationprocess*

1) The first stage isDataset Preparation. This involves creating a synthetic error dataset that deliberately mimics common Marathi grammar mistakes, such as errors in spelling, tense, gender, and word order. This data is paired with agold standarddatasetcomposedofsentencesthathave been manually corrected. The full dataset will be split into 80%for training, 10% for validation, and 10% for testing.

2) The second stage involves Automatic Evaluation Metrics.Wewillmeasurethetechnicalproficiency usingPrecision(percentage of correct corrections suggested) ,Recall(percentage of actual errors successfully corrected) , and theF_{0.5} Score(a weighted measure prioritizing precision). Standard NLPmetricslikeBLEU,GLEU,andROUGEwill also be used to measure the similarity between the system's output and the reference sentences.

3) The final stage isHuman Evaluation. Native Marathi speakers will rate the corrected text based onthreecrucialcriteria:Grammaticality(isthetext error-free?),Fluency(is the sentence smooth and natural?), andMeaning Preservation(does the correction keep the original intent?). This process will conclude with a Comparative Analysisof the Rule-BasedModel,theTransformerMTModel,and our Proposed Hybrid Model across different error types.

B. *Practical Implications*

1) The successful deployment of the Marathi GEC system is expected to deliver substantial practical benefits across various user demographics. The primaryoutcomeisenablingusers,whethertheyare students,contentcreators,orprofessionals,towrite grammatically correct Marathi withsignificantly increased confidence.

2) By providing an accessible, real-time tool, the system helps indrafting formal documentsand high-quality written material. Furthermore, the system’s design, which includes suggestions and explainable corrections, serves a pedagogical purpose by activelyencouraging better learning and understanding of the native language. Ultimately, this tool will support and promote greaterdigitaladoptionandcommunicationquality among native Marathi speakers.

C. *Limitations and Challenges*

ThedevelopmentanddeploymentofthisGrammaticalError Correction (GEC) system face several inherent limitations and anticipated challenges, primarily related to data, technology, and operations. A major technical limitation is the reliance onLarge LanguageModels(LLMs),whichbringsriskslikepotential overcorrectionbythegrammarmodel,resultinginincorrect suggestions or changing the user's intended meaning. We must also overcome the inherentlinguistic complexity of Marathi, which includes handling compound words and specificcasemarkers.Acriticaldatachallengeisthelimited availability of high-quality annotated GEC datafor Marathi, which necessitates time-consuming manual annotationandthegenerationofsyntheticdata,introducing the risk of human bias or a failure to reflect all real-world errors.

Operationally, theprojectisexposed toriskssuchaspotential integration issuesbetween the AI models on the backend and the frontend web user interface. Furthermore, dependency on cloud GPU resourcesfor training and inference is a cost factor that needs careful management. Finally, user-related challenges include the risk that users may reject suggestions if the accuracy is perceived as low, or there may beresistance to adoptingAI-based writing support among native speakers. different sources, clean and check all incoming data, and constantly watch out for new ways attacks might happen.

V. FUTURE WORK

The development of the Marathi GEC system sets the stage forseveralcrucialshort-termextensionsandambitiouslong- term research directions aimed at improving its capabilities and reach.

1) Short-Term:In the immediate future, our focus is on refining the core functionality and user experience. This includes fully integrating the Causal Language Model (CLM), MahaGPT, to provide robust and accurate real-time next-word suggestions while the user types. We also need to dedicate effort to creating standardized benchmark datasetsfortestingMarathiGEC.Thiswillallowfor systematic evaluation using metrics like BLEU, GLEU, and F_{0.5} to accurately measure performanceandcompareourhybridmodelagainst other approaches..

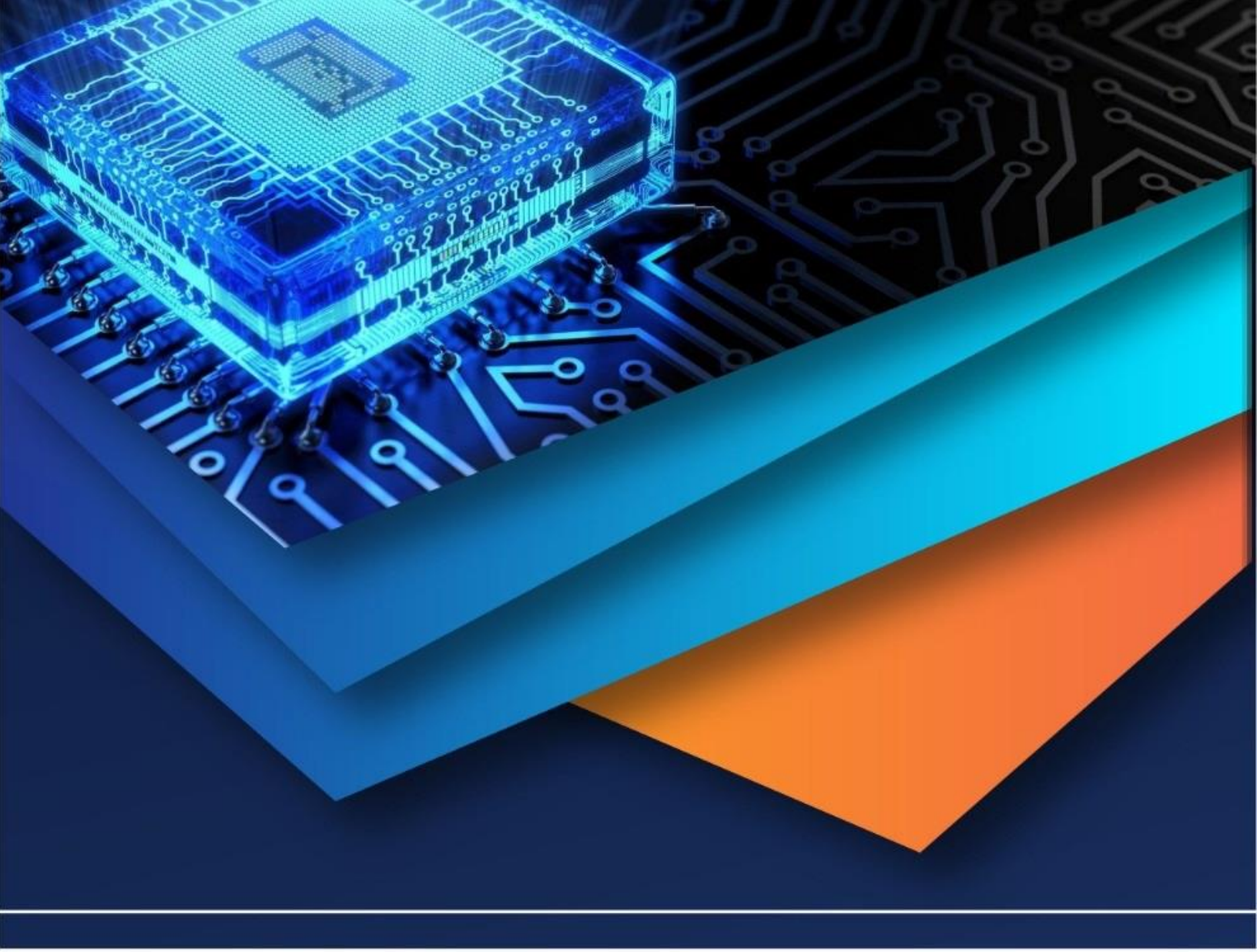
2) Long-TermVision:For the long term, we envision twomajorareasofgrowth.First,weplantoenhance thesystem'sintelligencebyintroducingExplainable AI (XAI) capabilities. This means moving beyond simplecorrectionstoofferbrief,clearexplanations for suggested edits, which will help educate users andincreasetrustinthesystem.Second,weaimfor expansion to other Indic languages. By leveraging the existing multilingual capacity of the core Transformermodels(likemT5andIndicBERT),we can apply the successful hybrid architecture and fine-tuning techniques developed for Marathi to support other low-resource regional languages, vastly increasing

VI. CONCLUSION

TheAI-Based Grammatical Error Correction (GEC) systemfor Marathi represents a significant step forward in addressing the technological gap for low-resource native languages. The project successfully established the viability ofahybridarchitecturethatstrategicallycombinesahigh- precision rule-based checker with the contextual processing power of aTransformer Seq2Seq MT model(e.g., mT5/MarianMT). This approach effectively handles both simple and complex, context-dependent errors. By framing GECasatranslation task and utilizing techniqueslikenoise injectionfor syntheticdatageneration,thesystem overcomes the challenge of data scarcity specific to Marathi. The deployment as areal-time web toolsolves the practical usability gap, providing native speakers with an accessible means to improve writing accuracy and confidence. Ultimately, this project delivers a functional, tested, and scalablesolutionthatcanserveasatemplateforGECefforts in other underserved Indic languages

REFERENCES

[1] I.Keita,A.W.Maiga,A.Sounaye,etal.(2025),"GECforLow-ResourceLanguages: Case of Zarma."
[2] Y. Jin, B. Zhang, and Y. He (2023), "Research and Analysis of GECTechnology for Chinese Documents."
[3] M. Juričić and F. Sarić (2024), "Evaluation of AI-based GrammarCorrection for Portuguese."
[4] C. Bryant, M. Felice, and T. Briscoe (2023), "Grammatical ErrorCorrection: A Survey of the State of the Art," Computational Linguistics,MITPress.
[5] S.Kobayashi,S.Flachs,andM.Rei(2024),"RevisitingMeta-evaluationfor Grammatical Error Correction," Transactions of the Association forComputational Linguistics (TACL).
[6] A. Ersoy and E. Yıldız (2024), "Organic Data-Driven Approach forTurkish GEC and LLMs," Workshop Proceedings / arXiv.
[7] J. Latouche, et al. (2024), "Zero-shot Cross-Lingual Transfer forSynthetic Data in Grammatical Error Correction," EMNLP / arXiv.
[8] [ACL/LRECPapers](2024), "GEC forCode-switched andMultilingualContexts," Proceedings of ACL & LREC.
[9] S. Chollampatt, D. T. Hoang, and H. T. Ng (2016), "AdaptingGrammaticalErrorCorrectionBasedontheNativeLanguageofWriterswithNeural Network Joint Models," in Proc. EMNLP, pp. 1901–1911.
[10] J. Park (2019), "An AI-based English Grammar Checker vs. HumanRaters in Evaluating EFL Learners' Writing," Multimedia-AssistedLanguage Learning, vol. 22, no. 1, pp. 112–131, doi:10.15702/mall.2019.22.1.112.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY