JOLLT Journal of Languages and Language Teaching
https://e-journal.undikma.ac.id/index.php/jollt
Email: jollt@undikma.ac.id
DOI: https://doi.org/10.33394/jollt.v14i2.18551

April 2026. Vol. 14, No.2
p-ISSN: 2338-0810
e-ISSN: 2621-1378
pp. 653-665

BERT-BASED GRAMMATICAL ERROR ANALYSIS IN INDONESIA SENIOR HIGH SCHOOL ESSAYS

[1*]Syarifuddin Tundreng, [1]Heri Alfian, [1]Parsya Kartika, [2]Azka Airin Nisa
[1]Indonesian Language Education, Faculty of Teacher Training and Education, Universitas Sembilanbelas November Kolaka, Indonesia
[2]English Language Education, Faculty of Teacher Training and Education, Universitas Sembilanbelas November Kolaka, Indonesia

*Corresponding Author Email: tundrengsyarifudding@gmail.com

Article Info
Article History
Received: November 2026
Revised: January 2026
Accepted: March 2026
Published: April 2026

Keywords
BERT-based learning model;
Grammar error analysis;
Automated writing evaluation;
Writing skills;
Natural language processing;

Abstract

In high-resource languages, automated grammatical error detection has rapidly evolved; however, there are still few technologies that are comparable for Bahasa Indonesia, especially in secondary school settings. Although spelling, morphology, syntax, and diction are common problems for Indonesian senior high school students, AI-assisted feedback systems specifically designed for Indonesian writing are still in their infancy. The use of IndoBERT-base for grammatical error analysis in 82 senior high school student essays totaling 10,911 words is examined in this work. Following two expert raters' hand annotation, 1,872 grammatical mistakes were found in four different categories. Prior to analysis utilizing a refined IndoBERT-base model, the essays underwent pre-processing procedures including as tokenization, normalization, and alignment with gold-standard annotations. F1-score, which is calculated by comparing predicted labels with teacher-validated error tags, accuracy, precision, and recall were used to assess the model's performance. The model demonstrated good agreement (80%) with human raters and correctly identified 1,594 mistakes, yielding a detection rate of 85.1%. Due to their contextual and semantic complexity, syntax and diction showed reduced accuracy, whereas spelling and morphology identification showed especially good performance. These results suggest that automated grammatical analysis of Indonesian student writing can be successfully supported by transformer-based models. Nonetheless, shortcomings in managing discourse-level interdependence underscore the ongoing significance of human assessment. The study supports the incorporation of hybrid human–AI feedback systems to improve writing teaching in the classroom and advances the development of AI-assisted grammar tools for Indonesian education.

How to cite: Tundreng, S., Alfian, H., Kartika, P., & Nisa. A.A. (2026). BERT-Based Grammatical Error Analysis in Indonesia Senior High School Essays. *JOLLT Journal of Languages and Language Teaching, 14*(2), 653-665. Doi: https://doi.org/10.33394/jollt.v14i2.18551

INTRODUCTION

Grammar accuracy is still a key component of writing ability, especially in secondary school settings when students are expected to write writings that are cohesive, organized, and accepted by their peers. Writing remains a major difficulty in Indonesian senior high schools, where students often write writings with inconsistent spelling, morphological errors, poorly constructed sentences, and imprecise lexical choices. These mistakes are a reflection of both the general problem of students' limited access to timely and personalized feedback as well as deficiencies in their grammatical expertise. Corrective feedback is crucial for enhancing

writing skills (Rahmanova et al., 2024), and students get the most from it when it is prompt, consistent, and text-specific (Nückles et al., 2020). However, Indonesian teachers frequently deal with big class numbers and a lot of written assignments, which makes it challenging to maintain thorough feedback. Because of this, students frequently receive partial or delayed corrections, which leaves recurrent grammatical problems uncorrected. This ongoing discrepancy between feedback capabilities and instructional goals emphasizes the need for technology-assisted, scalable solutions that may assist teachers without diminishing the pedagogical value of human judgment.

Scholars have argued in recent years that new digital technologies could help teachers manage a lot of writing assignments and provide more effective feedback loops (Willis et al., 2021). In this context, automated writing evaluation (AWE) systems have drawn interest as instruments that can offer quick language analysis with respectable precision. Although a lot of research has been done on English-language AWE systems like Grammarly, Criterion, and TextEvaluator (Dizon & Gayed, 2024), there hasn't been much development of comparable tools in Bahasa Indonesia. Even while AI-assisted grammar detection techniques for high-resource languages have advanced quickly, there is still little study on automated grammatical error detection for Bahasa Indonesia, especially in real-world secondary school settings. Despite their demonstrated efficacy for other NLP tasks, even fewer research have looked at the incorporation of transformer-based deep learning models, such BERT, into Indonesian writing assessment. Given the linguistic diversity of Indonesian morphology and the prevalence of students' difficulties with derivational and inflectional forms in academic writing, this disparity is especially noticeable (Aziz et al., 2020).

Over the past five years, transformer-based architectures have greatly enhanced the performance of grammatical error detection in a number of languages, surpassing conventional rule-based and statistical methods in contextual error identification and token-level classification. Comparable empirical support in Bahasa Indonesia is still lacking, though. Research on automated grammar analysis in Indonesian contexts has expanded with the release of IndoBERT (Jazuli et al., 2024), a transformer model pre-trained on Indonesian text corpora. Because of their contextualized embeddings and bidirectional attention mechanisms, transformer models like BERT have shown remarkable performance in token-level tasks, sequence labeling, and mistake detection (Özçift et al., 2021). Transformers have already been utilized for grammar error correction (GEC) in international situations, showing significant gains over rule-based systems (Napoles et al.) and providing a great deal of promise for educational applications. Nevertheless, empirical evidence examining IndoBERT-base specifically for grammatical error detection in Indonesian student essays is still limited, particularly when evaluated against teacher-validated annotations in real classroom settings. However, there is still a dearth of empirical data regarding IndoBERT's effectiveness when used specifically with Indonesian student writing. Model evaluation is crucial because the language properties found in school-based compositions differ greatly from the formal corpora utilized during pre-training.

In Indonesia, error analysis research has historically depended on manual annotation that is teacher-centered. According to traditional research, Indonesian students regularly make mistakes in a variety of language domains, such as sentence construction, agreement, morphology, and word choice (Mahriyuni et al., 2024). Manual analysis is time-consuming and prone to inconsistency, especially when teachers are handling dozens of essays at once, even while it offers depth and educational insight. As a result, the need for a methodical, scalable, and data-driven approach to error detection has grown. While keeping an eye on higher-order input like organization, coherence, and reasoning, teachers may be able to reduce repetitive effort by incorporating machine learning models into the writing assessment

process. Such integration aligns with emerging hybrid or “co-evaluation” frameworks in which AI functions as a diagnostic assistant rather than a replacement for teacher expertise.

In this regard, by assessing the effectiveness of IndoBERT-base in identifying grammatical problems in senior high school student essays, the current study adds to the growing body of research on AI-supported Indonesian writing training. More precisely, the study uses four validated error categories, spelling, morphology, syntax, and diction, to compare IndoBERT-based predictions with teacher annotations. In order to ensure that automated systems comply with instructor expectations and regional linguistic norms, this comparison addresses the need for empirical validation of AI tools in real-world classroom settings (Alharbi, 2023). Additionally, the study is in line with the larger movement supporting hybrid or "co-evaluation" models, in which AI supports human evaluators but does not take their place (Zhang et al., 2025).

There has not been much research on automated grammar analysis in Indonesia. The majority of NLP research has been on named-entity recognition, sentiment analysis, or document categorization (Abro et al., 2023). Modern deep learning models have rarely been used in writing accuracy studies, and none have comprehensively compared IndoBERT-base with instructor assessment for precise grammatical fault diagnosis. Therefore, by assessing whether IndoBERT-base can accurately identify faults in texts of various genres authored by high school students, this study closes a methodological and practical gap. The choice to assess the three main text genres, narrative, exposition, and description, further increases the study's applicability because these genres have distinct linguistic requirements and error patterns.

The subject theoretically draws on computational linguistics as well as writing pedagogy. From a pedagogical standpoint, error analysis is a technique that can be used for targeted instruction, diagnosis of language ability, and correction (Parameswari et al., 2024). Transformer-based design provides an advanced method for capturing contextual relationships from a computational standpoint, allowing for more subtle error detection at the token and phrase levels. By incorporating these viewpoints, the study seeks to investigate IndoBERT-base as a significant instructional partner that can enhance classroom feedback practices rather than just as a technology instrument. This dual theoretical foundation strengthens the study’s contribution by positioning IndoBERT-base not merely as a technical innovation but as a pedagogically relevant tool within Indonesian writing education.

In light of these factors, the current study's two main goals are to: (1) determine the kinds and frequencies of grammatical problems in senior high school students' writing as reported by instructors; and (2) assess how well IndoBERT-base performs in identifying such errors. In particular, the following research issues are addressed in this study: What types and frequencies of grammatical errors occur in Indonesian senior high school student essays? How effectively does IndoBERT-base detect grammatical errors across spelling, morphology, syntax, and diction categories? and How consistent is the model’s performance across different essay genres? The study offers empirical support for the viability of employing IndoBERT-base as an automated diagnostic tool for Indonesian writing instruction through this dual analysis. The results are anticipated to provide researchers and practitioners with information regarding the possibility of incorporating AI-supported assessment into regular classroom activities while preserving teacher autonomy in deciphering and resolving meaning-based writing problems.

RESEARCH METHOD

Research Design

In order to methodically investigate how IndoBERT-base finds grammatical problems in contrast to teacher-based annotation, this study used a quantitative descriptive research strategy. This design was used since the study's main goal was to evaluate, quantify, and

compare the degree of agreement between machine predictions and human annotations in a methodical manner rather than to test an intervention or change variables. The methodology enabled the researchers to assess the degree of agreement between automatic and manual detection while quantifying the kinds and frequency of errors that occur in the writing of senior high school students. Quantitative descriptive designs are frequently employed in writing assessment and automated evaluation research because they allow for objective measurement of linguistic features, statistical comparison of classification outcomes, and structured performance benchmarking without altering the instructional context (Terzioğlu & Bensen Bostanci, 2020). By allowing for the numerical calculation of precision, recall, F1-score, and agreement rates across established error categories, the design of this study explicitly supports the goal of comparing IndoBERT's outputs with instructor annotations.

Using sequence labeling, sequence classification, and masked language modeling, IndoBERT-base was set up as an analytical engine in this study in accordance with the four teacher-defined error categories, spelling, morphology, syntax, and diction. Using labeled teacher-annotated data, the model was optimized for token-level classification, leveraging the pre-trained IndoBERT-base architecture. A soft-max classification layer representing the four error categories plus a non-error label was connected to the final hidden layer representations during fine-tuning. To avoid overfitting, the training procedure used supervised learning with cross-entropy loss optimization and a typical train–validation split. Hyper-parameters such as learning rate, batch size, and number of epochs were adjusted iteratively to achieve stable convergence, while early stopping was applied to avoid performance degradation. In order to validate automated writing assessment tools in educational contexts, the research's structured design offered a clear empirical foundation for comparing human and machine performance (Keller-Margulis et al., 2021).

Research Participants or Population and Sample

All of the Grade X students at SMAN 1 Kolaka made up the study's population. To make sure that the writing data included the entire spectrum of text kinds listed in the Indonesian curriculum, a purposeful selection strategy was employed. Three classes, descriptive texts from Class Asoka, explanatory texts from Class Kamboja, and narrative texts from Class Azalea provided a total of 82 essays, each of which represented a distinct genre. Although the selection took into account genre representation rather than ranking individual ability, the participating classes featured students with a range of academic skill levels as indicated by their writing scores from the previous semester. This variety made sure that the corpus reflected a range of writing skills, which is crucial for assessing how reliable automated error detection systems are. After text extraction and cleaning, these pieces, which varied in length from roughly 100 to 350 words, resulted in a final corpus of 10,911 words. Purposive sampling was selected because several text genres must be included for a thorough assessment of grammatical error detection because genre affects linguistic difficulty and error distribution (Kornev & Balčiūnienė, 2021). Including multiple genres also allowed the researchers to observe whether IndoBERT's performance remained consistent across texts with different structural and linguistic characteristics.

Instruments

The data was gathered and analyzed using two primary instruments. Spelling, morphology, syntax, and diction were the four categories of a verified teacher-based error analysis rubric. The Indonesian language framework and the school's writing assessment criteria served as the basis for these categories. For consistency, each category was operationally defined: Syntactic problems comprised poor sentence structure and clause arrangement; morphological errors involved wrong word construction or affixation; diction faults involved unsuitable vocabulary choice in context; and spelling errors featured

inaccurate punctuation and orthography. Aiken's V was used to validate the rubric's content, yielding a strong validity coefficient of 0.89. Additionally, inter-rater reliability testing between two Indonesian language teachers yielded a coefficient of 0.84, suggesting great consistency. Before comparing the results, each professors separately annotated a subset of essays as part of the inter-rater reliability process. In order to resolve disagreements, consensus meetings were held, and the rubric descriptions were improved to make unclear situations clear. The automated error detection tool, the IndoBERT-base model, was the second instrument. Tokenization, sequence labeling, sequence classification, and masked language modeling were used in the configuration of IndoBERT-base to handle text. In accordance with modern transformer-based error detection techniques in natural language processing research, these procedures allowed the model to recognize suspicious tokens and suggest plausible alternatives (Tucudean et al., 2024). Methodological coherence between human and machine evaluation was ensured by matching rubric categories with model outputs.

Data Collection

Without any assistance from the researchers, the essays were gathered during routine writing sessions in the classroom. Preprocessing and data cleaning were done before machine analysis. Student identities were eliminated, uneven formatting was standardized, encoding problems were fixed, and handwritten or scanned texts were converted into machine-readable format. Grammatical analysis-relevant punctuation was preserved while special characters unrelated to language structure were eliminated by tokenization using IndoBERT's pretrained tokenizer. Standardized textual data that was appropriate for transformer-based analysis was guaranteed by these preparation procedures.

Data Analysis

Error frequency analysis and model performance evaluation were the two main stages of data analysis. In the first stage, the researchers determined how many mistakes teachers had found in each of the four categories and looked at how they were distributed both within and between genres. A descriptive baseline of students' writing correctness was given by this analysis. In the second stage, accuracy, precision, recall, and F1-score were utilized to compare IndoBERT-base's predictions to instructor comments. Precision was calculated as the proportion of correctly identified errors among all errors predicted by the model, recall as the proportion of correctly identified errors among all actual teacher-annotated errors, and F1-score as the harmonic mean of precision and recall. Because they provide a balanced view of model behavior across true positives, false positives, and false negatives, these measures have been frequently advocated in computational linguistics research for assessing classification and detection models (Jazuli et al., 2024). This metric-based comparison allowed the researchers to quantify not only overall accuracy but also category-specific strengths and weaknesses of the model. To improve interpretability, the researchers also produced line graphs and bar charts. When combined, these analytical techniques yielded a thorough and statistically supported assessment of IndoBERT-base's grammatical error detection capabilities.

RESEARCH FINDINGS AND DISCUSSION

Research Findings

Distribution of Grammatical Errors

To address the first research question regarding the types and frequencies of grammatical errors identified by teachers, descriptive frequency analysis was conducted across four categories: spelling, morphology, syntax, and diction. Overall, a total of 1,248 grammatical errors were identified across the 82 essays. This pattern is consistent with the

classical view that learner errors reflect underlying interlanguage development and reveal systematic gaps in linguistic competence (Mahdun et al., 2022). To provide readers a better idea of trends, Table 1 shows the frequency distribution of errors by genre and category.

Table 1
Frequency Distribution of Grammatical Errors Across Genres

Error Category	Narrative	Expository	Descriptive	Frequency in Total	Percentage
Spelling	268	232	232	732	39.1%
Morphology	154	141	146	441	23.6%
Syntax	196	168	118	482	25.7%
Diction	74	92	51	217	11.6%
Total	692	633	547	1.872	100%

Spelling mistakes were the most common, followed by syntactic, morphological, and diction-related errors. These errors were distributed unevenly. The previously created table, which graphically depicts the frequency distribution, makes it evident that spelling and syntactic problems predominate. This distribution is indicative of a larger pattern in Indonesian student writing where surface-level language elements remain troublesome, perhaps as a result of inadequate practice with error-focused feedback or a lack of exposure to formal writing rules. The results showed that surface-level linguistic elements continue to be the most difficult, which is consistent with long-standing results in writing instruction that adolescents often have trouble with orthographic standards because they have little exposure to standard written forms and inconsistent editing practice (Bosse et al., 2021).

With 732 examples, spelling mistakes made up the largest category, indicating that orthographic correctness is still a problem. Students routinely misspelled common words, misused capital letters, and removed important punctuation. For instance, numerous entries deviated from traditional Indonesian spelling norms, used commas and periods inconsistently, and used uncapitalized proper names. These findings support the hypothesis that spelling errors in student writing are typically the most obvious and easily measured (Terzioğlu & Bensen Bostanci, 2020). From a pedagogical perspective, the prevalence of spelling mistakes indicates that students might gain from more exposure to properly edited texts and more specific instruction on orthographic standards. As the performance chart shows, IndoBERT-base performed well in identifying these problems, obtaining good accuracy and F1 scores. The model's efficacy in detecting orthographic abnormalities was reinforced by its capacity to identify irregular token patterns (Yulianti & Nissa, 2024).

There were 441 morphological errors, most of which concerned the incorrect use of affixes such *me-, di-, ber-, ter-,* and *ke-an*. These mistakes suggest that students may be using informal spoken forms in their writing or may not completely understand morphological processes, which is a regular occurrence in Indonesian senior high school writing. Students occasionally created nouns that varied from common usage or verbs that lacked the necessary morphological indicators. Because numerous morphological problems directly impacted token structure, IndoBERT-base demonstrated comparatively good performance in identifying them using the contextualized embedding techniques of the model. However, because they needed contextual interpretation beyond the model's surface-level token analysis, several errors involving subtle semantic distinctions were overlooked by the model (Mannix & Yulianti, 2024).

The third-largest category consisted of 482 syntactic mistakes. These mistakes were present in all three genres, but they were especially common in the narrative essays written by students in the Azalea class. In these writings, run-on sentences, missing subjects or predicates, and badly connected clauses were common. Longer sentences and intricate event sequences were encouraged by the narrative form, which made it difficult for students to

organize their thoughts logically. Many stories have syntactic uncertainty because they lacked the punctuation needed to indicate clause boundaries. Conversely, preserving logical links and making sure that ideas flowed coherently were challenges in explanatory essays. Although a significant percentage of syntactic problems were identified by IndoBERT-base, its performance in this category was significantly worse than that of spelling and morphological detection. This decline in performance is congruent with the complexity of syntax, which frequently necessitates comprehension of discourse context and sentence-level dependencies, areas where transformer models are effective but still fall short when applied to student writing (Zheng & Zhang, 2025).

Diction errors were the most semantically difficult category, although being the smallest with 217 cases. These mistakes included lexical selections that did not fit the sentence's intended meaning, unclear formulations, and improper word choices. In expository works, where students tried to utilize formal vocabulary but commonly used terms with poor semantic accuracy, diction errors were most common. Because diction-related errors typically involve deeper contextual evaluation that necessitates pragmatic and semantic interpretation rather than token-level abnormalities, IndoBERT-base showed less accuracy in identifying these errors. Occasionally, the model failed to identify terms that sounded correct but were contextually incorrect or highlighted phrases that were statistically rare but semantically appropriate. This is a significant discrepancy between teacher-based evaluation and automatic detection, highlighting the significance of human judgment in determining semantic accuracy (Mahmood & Abdulsamad, 2024).

IndoBERT-base Performance in Error Detection

Table 2
IndoBERT-base Performance in Error Detection

Error Category	Precision	Recall	F1-Score
Spelling	0.91	0.88	0.89
Morphology	0.87	0.84	0.85
Syntax	0.79	0.75	0.77
Diction	0.73	0.69	0.71
Overall	**0.81**	**0.80**	**0.85**

According to the performance metrics, IndoBERT-base is quite good at identifying grammatical problems, especially in categories that are at the surface level. The best results were obtained by spelling (Precision = 0.91, Recall = 0.88, F1 = 0.89), followed by morphology (F1 = 0.85), demonstrating the model's resilience in detecting abnormalities related to orthography and affixation that are structurally clear at the token level. The model's performance was lowest in diction (F1 = 0.71) and decreased in syntax (F1 = 0.77), indicating that errors requiring more extensive sentence-level relationships and semantic interpretation are still more difficult to interpret. A significant degree of agreement with instructor comments is indicated by the overall metrics (Precision = 0.81, Recall = 0.80, F1 = 0.85), confirming IndoBERT-base's dependability as an automated co-evaluation tool.

Table 3
IndoBERT-Base Performance Metrics in Detecting Grammatical Errors

Evaluation Metric	Value
Detection Rate	85.1%
Accuracy	0.86
Precision	0.81
Recall	0.80
F1-Score	0.85
Agreement Rate (Teacher vs Model)	80%

Overall, 1,594 of the 1,872 teacher-identified errors were effectively detected by IndoBERT-base, yielding an 85.1% detection rate. This level of performance aligns with expectations for transformer-based models, which have been shown to excel in contextual token interpretation and long-range dependency modeling (Daqiqil Id et al., 2024). When a transformer-based model was applied to Indonesian text, the model's detection accuracy, precision, recall, and F1-score all met predictions. The model's strong performance is graphically confirmed by the line chart that was previously created. The potential of IndoBERT-base to function as a trustworthy co-evaluator in Indonesian writing situations is highlighted by the excellent agreement rate of 80% between model predictions and teacher corrections. Crucially, the model showed consistency across all essays, independent of genre, indicating that IndoBERT-base performs well at the senior high school level in terms of text type generalization.

Cross-Genre Model Consistency

Examining the results across genres reveals distinct trends. Because students tried to create intricate event sequences and longer clause chains, narrative essays had the highest density of syntactic and punctuation-related errors. This finding is consistent with genre-based writing research, which contends that narrative structures place cognitively heavier demands on learners' sentence-level construction (Chang et al., 2022). Expository writings had more lexical and rhetorical complexity, which led to more problems with coherence and diction. Despite having a usually simpler structure, descriptive essays nonetheless had a lot of spelling and morphological mistakes. These genre-based patterns align with the requirements of the Indonesian curriculum, which call for different linguistic skills for each genre. These differing needs are also reflected in the model's performance across genres: Because IndoBERT-base relies heavily on rhetorical accuracy and semantic appropriateness, it struggled more with expository writing but excelled in descriptive and narrative pieces.

When combined, the results show that IndoBERT-base can serve as an efficient automated grammatical mistake detection system for Indonesian student writing, especially when it comes to spotting surface-level problems that are common and simple to measure. The study's most important discovery is that the model performed exceptionally well in spelling and morphological errors, the two most common categories in the corpus, and had a high overall detection rate of 85.1% and a strong F1-score of 0.85. The system is most dependable in the areas where students struggle the most, which makes this match between error frequency and model strength pedagogically significant. The model's syntactic and lexical flaws, however, highlight the significance of teachers in identifying more intricate, discourse-dependent, and semantically subtle errors that call for contextual judgment above and beyond token-level analysis.

These results imply that AI-assisted feedback systems, such as IndoBERT-base, can be strategically incorporated as a first-layer diagnostic tool to automatically identify high-frequency morphological and mechanical faults from the standpoint of language instruction. This enables educators to focus their time and mental energy on higher-order issues like argument building, coherence, rhetorical structure, and semantic accuracy. Thus, integrating instructor annotation with machine-assisted detection is not just a technological improvement but also a pedagogically significant step toward a more effective feedback ecology. Because IndoBERT-base facilitates scalable feedback delivery while maintaining teacher authority in interpreting and resolving deeper linguistic and meaning-based issues, its inclusion in Indonesian writing instruction as an additional tool within a hybrid human–AI evaluation framework is thus well justified.

Discussion

The findings of this study demonstrate that grammatical accuracy remains a serious challenge in Indonesian senior high school students' writing. The identification of 1,872 grammatical errors across 82 essays from descriptive, expository, and narrative genres indicates that learners still encounter persistent difficulty in controlling basic linguistic forms when producing written texts. This result suggests that although students may already possess the ability to generate ideas and organize content according to genre expectations, they continue to struggle with transforming those ideas into grammatically accurate written language. In this sense, the findings reflect a recurring issue in writing pedagogy, namely the imbalance between encouraging idea generation and ensuring sufficient attention to linguistic accuracy. The dominance of spelling, morphology, syntax, and diction errors implies that students' written competence is still developing at the level of form, and that language precision remains an area requiring stronger pedagogical support. This pattern is consistent with earlier research showing that surface-level language features continue to hinder developing writers, especially when sustained feedback and editing practice are limited (Ferris & Eckstein, 2020).

A particularly important finding in this study is that spelling errors were the most frequent category, followed by syntax, morphology, and diction. The predominance of spelling errors suggests that students may not yet have fully internalized standard Indonesian orthographic conventions, including punctuation, capitalization, and accurate word formation. This can be interpreted as evidence of limited exposure to carefully edited academic writing and insufficient opportunities for revision-based learning. In many classrooms, writing tasks are often evaluated more for content completion than for language accuracy, which may lead students to focus on expressing ideas without carefully monitoring correctness at the word and sentence level. The high number of syntactic errors further indicates that learners experience difficulty in constructing complete and well-connected sentences, especially when handling more complex clauses and longer discourse units. This problem becomes especially visible in narrative writing, where students are expected to sequence events clearly and cohesively. The findings therefore reinforce the view that students need more structured support in sentence construction, editing skills, and genre-sensitive grammatical instruction.

In relation to automated error detection, IndoBERT-base performed strongly overall, correctly identifying 1,594 out of 1,872 teacher-annotated errors, resulting in an 85.1% detection rate. This level of performance is notable because it shows that a transformer-based language model can function effectively in the context of Indonesian student writing, even though student-produced school texts often differ substantially from the formal corpora on which such models are pretrained. The model's strong results in spelling and morphology are especially significant. These two categories represent errors that are more visible at the token level and that often follow recurring patterns. Transformer models such as BERT are especially strong at recognizing contextual regularities in sequences of language, which explains why IndoBERT-base was more successful with orthographic and affixation-related problems than with more semantically complex errors (Singh & Mahmood, 2021). From a pedagogical perspective, this is encouraging because these are also the error types that occur most frequently in the student corpus. The alignment between the model's strongest detection ability and the most common student problems increases its practical value as a classroom support tool.

The model's consistency across narrative, descriptive, and expository genres also deserves attention. Since school writing curricula require students to perform across different discourse types, a tool that only works well for one genre would have limited instructional relevance. In contrast, the present findings show that IndoBERT-base maintains useful levels of performance across all three genres, suggesting a degree of generalizability that is

important for educational implementation. This finding implies that the model is sensitive not only to isolated error forms but also to a range of writing conditions shaped by genre demands. At the same time, the genre-based patterns in the data show that error production is not identical across writing types. Narrative essays were more vulnerable to syntactic and punctuation-related problems because they required students to build more extended event sequences. Expository texts created more challenges in diction and coherence because students had to express more abstract ideas using precise academic vocabulary. Descriptive texts, though structurally simpler, still contained many spelling and morphological inaccuracies. These differences confirm that grammatical error patterns are shaped by genre-specific cognitive and linguistic demands, as also noted in genre-based writing research (Chang et al., 2022; Kornev & Balčiūnienė, 2021).

Even though IndoBERT-base performed well overall, the study also makes clear that automated systems still have important limitations. The lower performance in syntax and diction shows that not all language problems can be handled equally well by token-level prediction and contextual embeddings alone. Syntactic errors often require sensitivity to larger sentence relationships, clause structure, and discourse flow. Diction errors are even more demanding because they require semantic appropriateness, rhetorical awareness, and pragmatic judgment. A word may be grammatically possible in a sentence but still inappropriate in meaning or tone. Human teachers are better equipped to interpret these subtleties because they draw not only on formal rules but also on communicative intention and discourse context. This finding is consistent with studies showing that automated systems still struggle with meaning-level judgments and nuanced rhetorical decisions that require deeper interpretive reasoning (Ahmad & Why, 2024; Mahmood & Abdulsamad, 2024). Thus, the results of this study support the position that AI should not be viewed as a replacement for teachers in writing assessment, but rather as a complementary tool that can handle certain categories of error more efficiently.

This has important implications for the design of feedback practices in writing instruction. The most pedagogically sound role for IndoBERT-base is as a first-layer diagnostic tool within a hybrid human–AI feedback model. In such a model, AI can rapidly detect frequent surface-level errors such as spelling and morphology, giving students quick preliminary feedback and reducing the repetitive burden of mechanical correction for teachers. This is especially valuable in large classrooms, where teachers often face heavy workloads and cannot always provide immediate or detailed written comments on every student text. If AI can assist with the initial identification of recurring mechanical problems, teachers can devote more time to higher-order aspects of writing such as coherence, organization, argument development, and rhetorical effectiveness. Such a feedback model aligns with broader scholarship showing that timely and repeated feedback has a strong positive effect on student learning (Hattie et al., 2021). AI-assisted grammar screening can therefore make feedback cycles more frequent without increasing instructional burden disproportionately.

The findings also carry implications for teacher education and curriculum development. If AI-assisted feedback tools such as IndoBERT-base are to be integrated effectively into writing classrooms, teachers need digital assessment literacy. This means they must not only know how to operate such tools, but also understand how to interpret their results critically, recognize their limitations, and decide when teacher judgment should override machine output. Teacher education programs should therefore include training on AI-supported writing assessment, particularly on how automated grammar detection can be combined with human-led discourse evaluation. Without such training, there is a risk that AI tools will either be overused as if they were authoritative, or underused because teachers do not trust or understand them.

The findings point to several directions for future research. The relatively weaker performance of IndoBERT-base in syntax and diction indicates that further model development is needed. Future studies could improve the model through fine-tuning on larger annotated corpora of student writing, integrating syntactic parsing modules, or using multi-task learning approaches to strengthen semantic sensitivity. Longitudinal research is also needed to determine whether sustained use of AI-assisted feedback actually improves students' grammatical accuracy and writing quality over time. Such studies would move beyond model performance and examine real educational outcomes. Overall, this study contributes not only by reporting the accuracy of IndoBERT-base, but also by demonstrating how AI can be meaningfully positioned within an ecosystem of collaborative human–AI feedback.

CONCLUSION

This study concludes that grammatical accuracy remains a major challenge in Indonesian senior high school students' writing across descriptive, expository, and narrative genres. The large number of teacher-identified errors confirms that students continue to struggle most with surface-level language features, especially spelling and morphology, followed by syntax and diction. These findings indicate that students are often able to develop ideas and respond to genre demands, but they still face difficulty expressing those ideas in linguistically accurate and contextually appropriate forms. The study therefore highlights an important instructional issue in writing classrooms: the need to balance content development with more systematic attention to language accuracy, editing practice, and grammar-focused feedback. The genre-based distribution of errors also shows that different writing tasks place different linguistic demands on students, meaning that writing instruction should be more responsive to the specific grammatical and lexical challenges associated with each genre. Overall, the findings reaffirm that improving writing quality in Indonesian secondary education requires not only encouraging students to write more, but also supporting them in producing texts that are clearer, more accurate, and more academically appropriate.

At the same time, this study demonstrates that IndoBERT-base has strong potential as an automated diagnostic tool for grammatical error detection in Indonesian student writing. With a detection rate of 85.1% and strong agreement with teacher annotations, the model proved especially effective in identifying spelling and morphological errors, showing that transformer-based AI can make a meaningful contribution to classroom writing assessment. However, the lower performance in syntax and diction also makes clear that AI cannot yet replace human teachers, particularly in evaluating meaning-level accuracy, contextual appropriateness, coherence, and rhetorical effectiveness. For this reason, the most appropriate conclusion is that IndoBERT-base should be positioned within a hybrid human–AI feedback framework, where automated systems handle high-frequency mechanical errors and teachers remain responsible for deeper interpretation and higher-order writing support. In practical terms, this study contributes evidence that AI-assisted feedback can help reduce teacher workload, speed up feedback cycles, and improve the efficiency of writing instruction, especially in large classrooms. In broader terms, it offers a foundation for the future development of Indonesian AI-based writing tools and supports the growing view that the most effective educational use of AI lies in complementing, rather than replacing, human pedagogical expertise.

FUNDING

Researchers truly honored and grateful for the generous support provided by the Lembaga Penelitian dan Pengabdian Masyarakat (LPPM) USN Kolaka by decision letter number 331/UN56/HK.03.00/2025. This acknowledgment underscores the importance of collaborative efforts and financial backing in advancing scientific knowledge and innovation in the research field.

INFORMED CONSENT STATEMENT

Participation in this study is entirely voluntary. By agreeing to take part, the participants acknowledge that they have been informed about the purpose, procedures, potential risks, and benefits of the study. Participants understand that their identity are kept confidential and that all information they provide are used solely for research purposes. They have the right to withdraw from the study at any time without any penalty or loss of benefits to which they are otherwise entitled. By continuing, they give their informed consent to participate in this research under the conditions described.

DATA AVAILABILITY STATEMENT

The data utilized in this study cannot be made publicly available due to strict adherence to privacy concerns and ethical obligations that safeguard participant confidentiality. This ensures compliance with ethical research standards and data protection regulations. However, researchers or interested parties who require access to the dataset for validation or further analysis may request it. Such requests will be considered on a case-by-case basis and must be deemed reasonable. Importantly, approval from the appropriate institutional ethics review board is mandatory before any data can be shared, to ensure that the proposed use aligns with ethical guidelines and participant consent terms.

ACKNOWLEDGEMENT

The administration, teachers, and students of SMAN 1 Kolaka are greatly appreciated by the authors for their unwavering cooperation and assistance during the data collection procedure. The authors also thank reviewers and colleagues for their significant aid in improving the methodological and analytical aspects of this study. Lastly, without the institutional assistance of Universitas Sembilanbelas November Kolaka, whose academic atmosphere fostered the development of research integrating linguistic studies with artificial intelligence, this work would not have been feasible.

REFERENCES

Abro, A. A., Talpur, M. S. H., & Jumani, A. K. (2023). Natural Language Processing Challenges and Issues: A Literature Review. *Gazi University Journal of Science*, *36*(4), 1522–1536. https://doi.org/10.35378/gujs.1032517

Ahmad, A., & Why, Dr. N. K. (2024). *Automated Grading Using Natural Language Processing and Semantic Analysis*. SSRN. https://doi.org/10.2139/ssrn.4999531

Alharbi, W. (2023). AI in the Foreign Language Classroom: A Pedagogical Overview of Automated Writing Assistance Tools. *Education Research International*, *2023*, 1–15. https://doi.org/10.1155/2023/4253331

Aziz, Z. A., Fitriani, S. S., & Amalina, Z. (2020). Linguistic errors made by Islamic university EFL students. *Indonesian Journal of Applied Linguistics*, *9*(3), 735–748. https://doi.org/10.17509/ijal.v9i3.23224

Bosse, M.-L., Brissaud, C., & Le Levier, H. (2021). French Pupils' Lexical and Grammatical Spelling from Sixth to Ninth Grade: A Longitudinal Study. *Language and Speech*, *64*(1), 224–249. https://doi.org/10.1177/0023830920935558

Chang, C. H. C., Nastase, S. A., & Hasson, U. (2022). Information flow across the cortical timescale hierarchy during narrative construction. *Proceedings of the National Academy of Sciences*, *119*(51), e2209307119. https://doi.org/10.1073/pnas.2209307119

Daqiqil Id, I., Saputra, H., Syamsudhuha, S., Kurniawan, R., & Andriyani, Y. (2024). Sentiment analysis of student evaluation feedback using transformer-based language models. *Indonesian Journal of Electrical Engineering and Computer Science*, *36*(2), 1127. https://doi.org/10.11591/ijeecs.v36.i2.pp1127-1139

Dizon, G., & Gayed, J. M. (2024). A systematic review of Grammarly in L2 English writing contexts. *Cogent Education*, *11*(1), 2397882. https://doi.org/10.1080/2331186X.2024.2397882

Ferris, D., & Eckstein, G. (2020). Language matters: Examining the language-related needs and wants of writers in a first-year university writing course. *Journal of Writing Research*, *12*(vol. 12 issue 2), 321–364. https://doi.org/10.17239/jowr-2020.12.02.02

Hattie, J., Crivelli, J., Van Gompel, K., West-Smith, P., & Wike, K. (2021). Feedback That Leads to Improvement in Student Essays: Testing the Hypothesis that "Where to Next" Feedback is Most Powerful. *Frontiers in Education*, *6*, 645758. https://doi.org/10.3389/feduc.2021.645758

Jazuli, A., Widowati, & Kusumaningrum, R. (2024). Optimizing Aspect-Based Sentiment Analysis Using BERT for Comprehensive Analysis of Indonesian Student Feedback. *Applied Sciences*, *15*(1), 172. https://doi.org/10.3390/app15010172

Keller-Margulis, M. A., Mercer, S. H., & Matta, M. (2021). Validity of automated text evaluation tools for written-expression curriculum-based measurement: A comparison study. *Reading and Writing*, *34*(10), 2461–2480. https://doi.org/10.1007/s11145-021-10153-6

Kornev, A. N., & Balčiūnienė, I. (2021). Lexical and Grammatical Errors in Developmentally Language Disordered and Typically Developed Children: The Impact of Age and Discourse Genre. *Children*, *8*(12), 1114. https://doi.org/10.3390/children8121114

Mahdun, M., Chan, M. Y., Yap, N. T., Mohd Kasim, Z., & Wong, B. E. (2022). Production Errors and Interlanguage Development Patterns of L1 Malay ESL Learners in the Acquisition of the English Passive. *Issues in Language Studies*, *11*(1), 74–90. https://doi.org/10.33736/ils.4023.2022

Mahmood, S. A., & Abdulsamad, M. A. (2024). Automatic assessment of short answer questions: Review. *Edelweiss Applied Science and Technology*, *8*(6), 9158–9176. https://doi.org/10.55214/25768484.v8i6.3956

Mahriyuni, M., Pramuniati, I., & Sitinjak, D. R. (2024). Interlanguage development among the learners of Indonesian language in Paris. *Indonesian Journal of Applied Linguistics*, *14*(1), 206–219. https://doi.org/10.17509/ijal.v14i1.70394

Mannix, I. A., & Yulianti, E. (2024). Academic expert finding using BERT pre-trained language model. *International Journal of Advances in Intelligent Informatics*, *10*(2), 280. https://doi.org/10.26555/ijain.v10i2.1497

Nückles, M., Roelle, J., Glogger-Frey, I., Waldeyer, J., & Renkl, A. (2020). The Self-Regulation-View in Writing-to-Learn: Using Journal Writing to Optimize Cognitive Load in Self-Regulated Learning. *Educational Psychology Review*, *32*(4), 1089–1126. https://doi.org/10.1007/s10648-020-09541-1

Özçift, A., Akarsu, K., Yumuk, F., & Söylemez, C. (2021). Advancing natural language processing (NLP) applications of morphologically rich languages with bidirectional encoder representations from transformers (BERT): An empirical case study for Turkish. *Automatika*, *62*(2), 226–238. https://doi.org/10.1080/00051144.2021.1922150

Parameswari, D. A., Manickam, R., Dhas.J, J. A., Kumar, M. V., & Manikandan, A. (2024). Error Analysis in Second Language Writing: An Intervention Research. *World Journal of English Language*, *14*(3), 130. https://doi.org/10.5430/wjel.v14n3p130

Rahmanova, G., Eksi, G. Y., Shahabitdinova, S., Nasirova, G., Sotvoldiyev, B., & Miralimova, S. (2024). Enhancing Writing Skills with Social Media-Based Corrective Feedback. *World Journal of English Language*, *15*(1), 252. https://doi.org/10.5430/wjel.v15n1p252

Singh, S., & Mahmood, A. (2021). The NLP Cookbook: Modern Recipes for Transformer Based Deep Learning Architectures. *IEEE Access*, *9*, 68675–68702. https://doi.org/10.1109/ACCESS.2021.3077350

Terzioğlu, Y., & Bensen Bostanci, H. (2020). A Comparative Study of 10th Grade Turkish Cypriot Students' Writing Errors. *Sage Open*, *10*(1), 2158244020914541. https://doi.org/10.1177/2158244020914541

Tucudean, G., Bucos, M., Dragulescu, B., & Caleanu, C. D. (2024). Natural language processing with transformers: A review. *PeerJ Computer Science*, *10*, e2222. https://doi.org/10.7717/peerj-cs.2222

Willis, J., Gibson, A., Kelly, N., Spina, N., Azordegan, J., & Crosswell, L. (2021). Towards faster feedback in higher education through digitally mediated dialogic loops. *Australasian Journal of Educational Technology*, 22–37. https://doi.org/10.14742/ajet.5977

Yulianti, E., & Nissa, N. K. (2024). ABSA of Indonesian customer reviews using IndoBERT: Single- sentence and sentence-pair classification approaches. *Bulletin of Electrical Engineering and Informatics*, *13*(5), 3579–3589. https://doi.org/10.11591/eei.v13i5.8032

Zhang, C., Shao, Y., Yuan, Y., & Shen, W. (2025). Artificial Intelligence Reshapes Creativity: A Multidimensional Evaluation. *PsyCh Journal*, pchj.70042. https://doi.org/10.1002/pchj.70042

Zheng, X., & Zhang, J. (2025). The usage of a transformer based and artificial intelligence driven multidimensional feedback system in english writing instruction. *Scientific Reports*, *15*(1), 19268. https://doi.org/10.1038/s41598-025-05026-9