Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 167−178
DOI 10.47065/josh.v8i1.11296
https://ejurnal.seminar-id.com/index.php/josh/

Comparing Machine Learning and BERT for Classifying Grammatical Error Types in Indonesian EFL Essays

Meily Ivane Esther Neman[1], George Morris William Tangka[2,*], Regi Fernando Najoan[3], Enrico Samuel Djimesha[4], Daniel Semuel Warouw[1]

[1]Fakultas Keguruan dan Ilmu Pendidikan, Pendidikan Bahasa Inggris, Universitas Klabat, Minahasa Utara, Indonesia
Jl. Arnold Mononutu, Airmadidi Bawah, Kec. Airmadidi, Kabupaten Minahasa Utara, Sulawesi Utara, Indonesia
[2]Fakultas Ilmu Komputer, Sistem Informasi, Universitas Klabat, Minahasa Utara, Indonesia
Jl. Arnold Mononutu, Airmadidi Bawah, Kec. Airmadidi, Kabupaten Minahasa Utara, Sulawesi Utara, Indonesia
[3]Fakultas Ilmu Komputer, Informatika, Universitas Klabat, Minahasa Utara, Indonesia
Jl. Arnold Mononutu, Airmadidi Bawah, Kec. Airmadidi, Kabupaten Minahasa Utara, Sulawesi Utara, Indonesia
[4]Fakultas Teknik, Arsitektur, Universitas Klabat, Minahasa Utara, Indonesia
Jl. Arnold Mononutu, Airmadidi Bawah, Kec. Airmadidi, Kabupaten Minahasa Utara, Sulawesi Utara, Indonesia
Email: [1]meilyneman@unklab.ac.id, [2,*]gmwtangka@unklab.ac.id, [3]reginajoan@unklab.ac.id, [4]enrico.djimesha@unklab.ac.id, [5]danielwarouw@unklab.ac.id
Correspondence Author Email: gmwtangka@unklab.ac.id
Received: **07/10/2026**, Revised: **18/10/2026**, Accepted: **30/10/2026**, Published**: 31/10/2026**

Abstract−Automated writing feedback tools are usually trained on learner corpora from outside Indonesia, so their accuracy on Indonesian students' grammatical errors is unknown. This study compares Multinomial Naive Bayes, a linear Support Vector Machine (SVM), Random Forest, and a fine-tuned BERT model for classifying nine grammatical error types. The models were trained on 48,011 error edits from the BEA-2019 W&I+LOCNESS and FCE corpora, re-annotated with ERRANT 3, and tested in-domain on 5,129 edits and cross-corpus on 1,900 edits from 66 essays by 33 Indonesian university students in the ICNALE Edited Essays. BERT achieved the highest macro-F1 on both test sets, 0.917 in-domain and 0.849 on the Indonesian essays, and McNemar tests confirmed its advantage over each classical model. All models lost 0.035 to 0.068 macro-F1 under the corpus shift. Noun number errors formed 24% of the Indonesian errors, three times their training share, while word form and pronoun were the hardest classes. This study contributes a cross-corpus benchmark of classical and Transformer-based error-type classifiers on Indonesian students' essays, a reproducible pipeline that labels both corpora with the same ERRANT version, and a class-level reliability profile for classroom feedback. For English teachers in Indonesia, the results support explicit, contrastive practice on articles and plural marking and the use of BERT for immediate feedback on reliably recognised classes, while its word form and pronoun suggestions are confirmed by the teacher.

Keywords: BERT; Error Type Classification; ERRANT; Indonesian EFL Learners; Cross-Corpus Evaluation

1. INTRODUCTION

Writing is widely regarded as the most demanding skill for Indonesian learners of English as a Foreign Language (EFL). A meta-analysis of 21 Indonesian studies published between 2014 and 2024 found that linguistic problems account for 45.16% of the writing challenges reported by students, with grammar as the most prominent of them [1]. Error analyses of Indonesian students' essays and theses document frequent grammatical errors [2]. In descriptive texts written by senior high school students in Manado, misformation (57.94%) and omission (25.23%) were the dominant error types [3], and in English theses article errors were the second most frequent of eleven categories (20%) after unnecessary words [4]. These errors are commonly attributed to first-language interference [5], since Indonesian has no articles and does not mark tense or plural number through inflection. Identifying the errors by hand is slow, which limits how much detailed feedback a lecturer can give in large writing classes.

Automated writing evaluation tools have therefore become common in Indonesian classrooms. Grammarly is the most studied example: a systematic review of 24 studies found that it is used mainly with university students [6], and an Indonesian study that used Grammarly reports to profile 54 students' academic writing again found grammar to be the main problem [7]. Large language models (LLMs) now offer a further option, but their use in education raises open questions of reliability, bias, and transparency [8], and their validity for scoring and correcting learner writing is still being examined [9], [10]. Such tools are nevertheless already part of students' learning: 75.65% of the tenth-grade students surveyed in one Indonesian study perceived ChatGPT as useful for learning English [11]. For teaching, a correction is more useful when the learner is also told what kind of error was made. Knowing that an error concerns noun number rather than word choice allows the learner to connect the feedback to a grammar rule, which supports the metacognitive strategies that Indonesian university students already report using when they learn English [12].

In natural language processing, this problem is studied as grammatical error correction (GEC). Neural GEC systems based on sequence-to-sequence models and pre-trained Transformers now dominate the field [13], [14]. Error types are usually assigned with ERRANT, a rule-based toolkit that aligns an original and a corrected sentence, extracts the edits, and classifies each edit into one of 25 main categories using part-of-speech and morphological information [15]. ERRANT became the standard annotation scheme of the BEA-2019 shared task, which also released the W&I+LOCNESS and FCE learner corpora used in this study [16]. Classifying the type of an error is itself a text classification task. Classical methods such as Naive Bayes, Support Vector Machines

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*.
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 167−178
DOI 10.47065/josh.v8i1.11296
https://ejurnal.seminar-id.com/index.php/josh/

(SVM), and Random Forest work on sparse lexical features, while pre-trained language models such as BERT learn contextual representations that are fine-tuned for the task [17]–[19].

Five recent studies are closest to the present work, and each leaves part of the problem unsolved. First, Antara and Anggreni [20] used DeepSeek AI to analyse 17 opinion essays of Indonesian learners and found 109 grammatical problems, led by article omission (42.2%), subject-verb agreement (26.6%), prepositions (17.4%), and tense (13.8%). The study shows what an AI tool can reveal about Indonesian learners' errors, but its sample is small and the AI judgements were not compared with a reference annotation. Second, Qin [21] proposed a deep learning model for the automatic correction of English grammar errors. The model produces corrected sentences, but it does not tell the learner which type of error was made, and it was not evaluated on English written by Indonesian learners. Third, Musyafa et al. [22] built a Transformer model and a large corpus for correcting grammatical errors in Indonesian-language text; the work targets Indonesian itself rather than the English of Indonesian learners. Fourth, Lin [9] evaluated ChatGPT, Ernie-4, and Llama3 on learner Chinese with the MaxMatch metric, which again measures corrections rather than error types and concerns a different language. Fifth, Sabilla and Dhenabayu [23] compared SVM, a CNN, and BERT for complaint text classification and found BERT best with a macro-F1 of 0.9167, but, like comparable Indonesian studies of SVM, Random Forest, and Naive Bayes [24] and of tree-based regression [25], the models were trained and tested on data from a single source.

Such single-source evaluations can be misleading, because a model trained on one distribution often performs worse on another, a problem known as domain shift [26], and GEC systems score lower on datasets concentrated on difficult errors than on random samples, so standard benchmarks can overstate real performance [27]. Text produced in different countries can differ systematically even on the same subject; sentiment and topics in Indonesian and US tweets about the same post-pandemic health issues differed markedly [28]. Public GEC corpora were collected mainly from learners outside Indonesia, whereas corpus studies of Indonesian learners have focused on features such as lexical bundles rather than on annotated error types [29]. The International Corpus Network of Asian Learners of English (ICNALE) offers a rare resource here, because its Edited Essays module pairs Indonesian students' essays with versions corrected by professional proofreaders [30].

The comparison of these studies reveals three research gaps. The first is that studies of Indonesian learners' errors rely on manual or AI-assisted analysis of small samples without a reference annotation. The second is that GEC studies produce corrections rather than identifying the type of error, and none of them evaluates English written by Indonesian learners. The third is that comparisons of classical and Transformer-based classifiers train and test on a single data source, so their accuracy under a corpus shift remains unknown. Consequently, no previous study has measured how well error-type classifiers trained on public learner corpora transfer to essays written by Indonesian students. This study addresses three objectives. The first is to describe how the distribution of grammatical error types in Indonesian students' essays differs from that of the training corpora. The second is to determine which of Naive Bayes, SVM, Random Forest, and BERT classifies grammatical error types most accurately, both in-domain and on Indonesian essays. The third is to identify which error types remain difficult after the corpus shift and why. The contributions of this study are threefold. First, it provides a reproducible pipeline that labels the BEA-2019 corpora and the ICNALE essays with the same version of ERRANT, so that training and test labels are directly comparable. Second, it offers, to our knowledge, the first cross-corpus comparison of Naive Bayes, SVM, Random Forest, and BERT on essays written by Indonesian students, reported with bootstrap confidence intervals and significance tests rather than single scores. Third, it provides a class-level profile of which error types can be classified reliably, which shows English teachers in Indonesia where automated error-type feedback can be trusted and where it must be checked.

2. RESEARCH METHODOLOGY

This study used a quantitative experimental design that compared four classifiers under identical data and evaluation conditions. The experiments were run in September and October 2026 on an Apple-silicon laptop, using the Metal Performance Shaders (MPS) backend of PyTorch for BERT. The pipeline used Python 3.11, scikit-learn 1.9.1, PyTorch 2.14.0, Transformers 5.17.0, spaCy 3.8.16, and ERRANT 3.0.2.

2.1 Research Stages

The study began with two data sources processed in parallel. The BEA-2019 corpora, which already contain gold corrections, were re-typed with ERRANT 3 so that every edit received a label from the same version of the tool. The ICNALE Edited Essays, which contain only an original and a proofread version of each essay, were aligned by ERRANT so that the edits and their types could be extracted automatically. In the third stage, the ERRANT categories from both sources were mapped to nine grammatical classes, and edits that concern spelling, punctuation, word choice, or mixed rewrites were removed. The resulting instances were then represented in two ways: as sparse TF-IDF features for the classical models and as marked sentences for BERT. In the fifth stage, Naive Bayes, the linear SVM, and Random Forest were trained with a small grid search, and BERT was fine-tuned. The hyper-parameters and the BERT checkpoint were selected on the FCE validation set only. In the final

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 168
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 167−178
DOI 10.47065/josh.v8i1.11296
https://ejurnal.seminar-id.com/index.php/josh/

stage, every selected model was evaluated on the in-domain test set and on the Indonesian essays with macro-F1, 95% bootstrap confidence intervals, and McNemar tests. Figure 1 depicts the complete sequence of these stages.

2.2 Data

Training and in-domain test data came from the BEA-2019 release of the Cambridge English Write & Improve (W&I) corpus and the First Certificate in English (FCE) corpus [16], [31]. Both corpora are distributed in M2 format, in which each sentence is followed by its gold edits. The W&I training sets for CEFR levels A, B, and C and the FCE training set formed the training data. The FCE development set was used for model selection. Because the W&I test annotations are not public, the W&I development sets and the FCE test set formed the in-domain test set. The native LOCNESS essays were excluded. The cross-corpus test set came from the Edited Essays module of ICNALE version 3.1 [30]. The module contains argumentative essays of 200 to 300 words on two topics, part-time jobs for college students and smoking bans in restaurants, together with versions edited by professional proofreaders. All Indonesian essays for which both the original and the edited version were available were used: 66 essays by 33 students, distributed over the CEFR levels A2 (20 essays), B1.1 (20), B1.2 (20), and B2 (6).

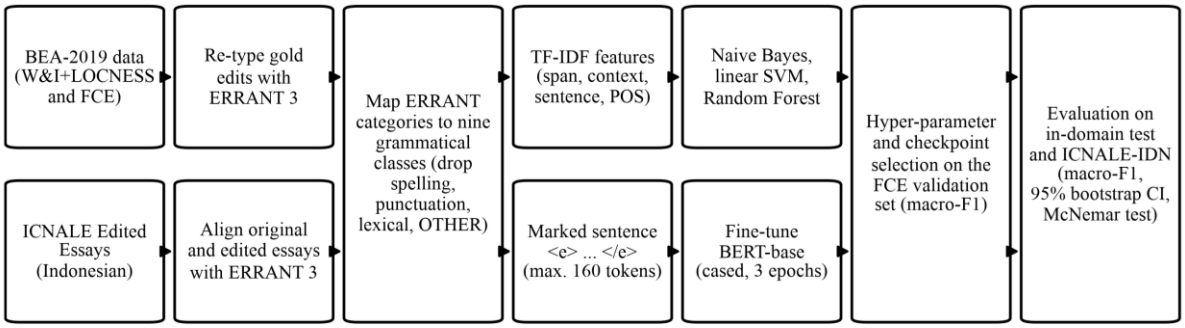


Figure 1. Research stages

Table 1 summarises the source, size, and number of edits of each data split.

Table 1. Dataset statistics

Split	Source	Sentences / essays	All ERRANT edits	Edits used
Training	W&I train (A–C), FCE train	40,452 sentences	105,083	48,011
Validation	FCE dev	1,370 sentences	3,421	1,618
In-domain test	W&I dev (A–C), FCE test	4,104 sentences	11,081	5,129
ICNALE-IDN	ICNALE Edited Essays, Indonesia	66 essays	3,980	1,900

Note: sentences are counted only when they contain at least one edit.

2.3 Error Annotation and Class Definition

All edits were typed with ERRANT 3.0.2 and the spaCy en_core_web_sm 3.8 model [15]. For the BEA-2019 data, each gold edit of the first annotator was re-typed with ERRANT 3. For ICNALE, the original and edited versions of each essay were normalised for whitespace, tokenised, and aligned by ERRANT. Edits that crossed a sentence boundary were discarded. The ERRANT categories were then mapped to nine grammatical classes. Spelling, orthography, and punctuation edits were excluded because they concern mechanics rather than grammar. Replacements of open-class words (NOUN, VERB, ADJ, ADV) were excluded because they usually reflect word choice, and the residual OTHER category was excluded because it has no single grammatical interpretation. Conjunction, particle, contraction, and possessive edits were also excluded because they were too rare to form reliable classes. After this mapping, 48,011 of the 105,083 training edits (45.7%) and 1,900 of the 3,980 ICNALE edits (47.7%) were retained. Every class had more than 100 training examples, so all nine classes were kept. Each instance consists of a sentence, the position of the erroneous span in that sentence, and the class label. For missing-word errors the span is empty and marks the insertion point. Table 2 lists the nine classes, the ERRANT categories mapped to each class, and the frequency of each class in the training data and in ICNALE-IDN.

Table 2. Mapping of ERRANT categories to the nine classes

Class	ERRANT categories	Train (n)	ICNALE-IDN n (%)
Determiner	DET	12,037	469 (24.7)
Preposition	PREP	11,368	315 (16.6)
Verb tense	VERB:TENSE	6,683	182 (9.6)
Subject-verb agreement	VERB:SVA	2,069	97 (5.1)

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 169
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 167−178
DOI 10.47065/josh.v8i1.11296
https://ejurnal.seminar-id.com/index.php/josh/

Class	ERRANT categories	Train (n)	ICNALE-IDN n (%)
Verb form	VERB:FORM, VERB:INFL	3,949	125 (6.6)
Noun number	NOUN:NUM, NOUN:INFL	3,850	455 (23.9)
Word form	MORPH, ADJ:FORM	2,556	127 (6.7)
Pronoun	PRON	3,663	94 (4.9)
Word order	WO	1,836	36 (1.9)
Total		**48,011**	**1,900 (100)**

2.4 Classification Models

The three classical models used the same sparse feature set, built with TF-IDF weighting and sublinear term frequency. The features combined word unigrams and bigrams of the erroneous span, character 2- to 4-grams of the span, up to three context words on each side encoded with their relative position, word unigrams and bigrams of the sentence within 40 tokens of the error with the span enclosed in <e> and </e> markers (maximum 30,000 features), and 1- to 3-grams of the part-of-speech (POS) tags of the span and its context. Multinomial Naive Bayes was tuned over the smoothing parameter $\alpha \in \{0.01, 0.05, 0.1, 0.5, 1.0\}$. A linear SVM with balanced class weights was tuned over $C \in \{0.05, 0.1, 0.5, 1.0\}$. Random Forest used 300 trees with balanced subsample weights, the 5,000 features with the highest χ^2 scores, and a minimum leaf size of 1 or 2. For each model, the setting with the highest validation macro-F1 was selected: $\alpha = 0.1$, $C = 0.5$, and a minimum leaf size of 1.

For BERT, the bert-base-cased checkpoint was fine-tuned for sequence classification [17], [32]. The markers <e> and </e> were added to the vocabulary as special tokens and inserted around the erroneous span, so that the model received the sentence together with the location of the error, truncated to 160 sub-word tokens. Training used AdamW with a learning rate of 2×10^{-5}, weight decay of 0.01, a batch size of 32, linear warm-up over the first 10% of steps, gradient clipping at 1.0, and a class-weighted cross-entropy loss to counter class imbalance. The model was trained for three epochs with random seed 42, and the epoch with the highest validation macro-F1 was kept. The validation macro-F1 rose from 0.865 after the first epoch to 0.910 and 0.916 after the second and third, so the third epoch was used. Fine-tuning took about 61 minutes, compared with about 0.1 s for Naive Bayes, 2.9 s for the SVM, and 12.1 s for Random Forest with their selected settings.

2.5 Classification Algorithms

The four classifiers differ in how they turn an error instance into a class decision. For the classical models, each instance is represented as a vector of TF-IDF weights. With sublinear term frequency and a smoothed inverse document frequency, the weight of feature t in instance d is computed with Equation (1), where $\mathrm{tf}_{t,d}$ is the count of t in d, N is the number of training instances, and df_t is the number of training instances that contain t. Each feature group is then normalised to unit length [18].

$$w_{t,d} = \left(1 + \ln \mathrm{tf}_{t,d}\right) \cdot \left(\ln \frac{1+N}{1+\mathrm{df}_t} + 1\right) \tag{1}$$

Multinomial Naive Bayes assigns the class with the highest posterior probability under the assumption that features are conditionally independent given the class [18]. Its decision rule and the smoothed feature probabilities are given in Equations (2) and (3), where x_i is the weight of feature i, P(c) is the prior probability of class c, N_{ci} is the total weight of feature i in the training instances of class c, N_c is the total weight of all features in class c, n is the number of features, and α is the smoothing parameter.

$$\hat{y} = \arg\max_c [\ln P(c) + \textstyle\sum_{i=1}^{n} x_i \ln \theta_{ci}] \tag{2}$$

$$\theta_{ci} = \frac{N_{ci}+\alpha}{N_c+\alpha n} \tag{3}$$

The linear SVM learns, for each class, a hyperplane that separates the class from all other classes (one-vs-rest) with a maximum margin. For each binary sub-problem with labels $y_i \in \{-1, +1\}$, the weight vector $\mathbf{w}$ and bias b minimise the regularised squared hinge loss in Equation (4). The term $\frac{1}{2}\|\mathbf{w}\|^2$ maximises the margin, C controls the penalty for margin violations, and s_i is the balanced weight of the original class of instance i, which compensates for class imbalance. A new instance is assigned to the class whose hyperplane gives the highest score.

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C \textstyle\sum_{i=1}^{N} s_i \left[\max\left(0, 1 - y_i(\mathbf{w}\cdot\mathbf{x}_i + b)\right)\right]^2 \tag{4}$$

Random Forest combines B = 300 decision trees, each trained on a bootstrap sample of the training data. Each tree chooses its splits so as to reduce the Gini impurity in Equation (5), where p_c is the proportion of class c in a node, and the forest predicts the class with the highest average probability over all trees, as in Equation (6), where $p_b(c \mid \mathbf{x})$ is the class probability estimated by tree b.

$$G = 1 - \textstyle\sum_{c=1}^{K} p_c^2 \tag{5}$$

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 170
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 167−178
DOI 10.47065/josh.v8i1.11296
https://ejurnal.seminar-id.com/index.php/josh/

$$\hat{y} = \arg\max_{c} \frac{1}{B} \sum_{b=1}^{B} p_b(c|\mathbf{x}) \quad (6)$$

BERT encodes the marked sentence with a stack of Transformer layers and passes the pooled representation **h** of the [CLS] token to a linear output layer [17], [32]. The class probabilities are obtained with the softmax function in Equation (7), where $z_c = \mathbf{W}_c\mathbf{h} + b_c$ is the score of class c. During fine-tuning, all parameters are updated to minimise the class-weighted cross-entropy loss in Equation (8), where $w_c = N/(K \cdot N_c)$ is the balanced weight of class c, N_c is the number of training instances of class c, and K = 9. These weights give rare classes such as word order the same total influence on training as frequent classes such as determiner.

$$p(c|\mathbf{x}) = \frac{\exp(z_c)}{\sum_{j=1}^{K} \exp(z_j)} \quad (7)$$

$$L = -\frac{\sum_{i=1}^{N} w_{y_i} \ln p(y_i|\mathbf{x}_i)}{\sum_{i=1}^{N} w_{y_i}} \quad (8)$$

2.6 Evaluation

Each model was evaluated on the in-domain test set and on ICNALE-IDN with accuracy, macro-averaged precision (Macro-P) and recall (Macro-R), and macro-F1. Macro-F1 was the primary metric because it gives each class the same weight, regardless of its frequency. For a class k with precision P_k and recall R_k, and K = 9 classes, macro-F1 is computed with Equations (9) and (10).

$$F1_k = \frac{2P_kR_k}{P_k+R_k} \quad (9)$$

$$\text{Macro-F1} = \frac{1}{K} \sum_{k=1}^{K} F1_k \quad (10)$$

Uncertainty was estimated with a cluster bootstrap of 1,000 resamples, which resampled whole essays for ICNALE-IDN and whole source sentences for the in-domain set, and reported the 2.5th and 97.5th percentiles as a 95% confidence interval (CI). Differences between BERT and each classical model were tested with the exact McNemar test at $\alpha = 0.05$, which compares the number of instances that only one of the two models classified correctly. Results were also broken down by class, by CEFR level, and in a confusion matrix, and a feature ablation retrained the selected SVM without the POS features.

3. RESULTS AND DISCUSSION

3.1 Implementation of the Classification Procedure

The procedure that addresses the research problem was carried out in the order defined in Section 2.1. In the first step, ERRANT processed 40,452 training sentences that contain at least one edit and produced 105,083 typed edits, and it extracted 3,980 edits from the 66 Indonesian essays. In the second step, the class mapping retained 48,011 training edits and 1,900 Indonesian edits, and each retained edit was stored as one classification instance consisting of the sentence, the position of the error, and its class.

The instance "It have many benefit for them", taken from an Indonesian essay and corrected to has, illustrates the remaining steps. ERRANT typed the edit as VERB:SVA, which the mapping turned into the class subject-verb agreement. For the classical models, the instance was converted into TF-IDF features such as the span word have, the context word it immediately to its left, the context words many, benefit, and for to its right, and the part-of-speech pattern PRP–VBP–JJ around the span. For BERT, the same instance was presented as the marked sentence "It <e> have </e> many benefit for them .". Naive Bayes, Random Forest, and BERT classified the instance correctly as subject-verb agreement, whereas the SVM predicted verb tense, which shows how closely the agreement and tense classes are related.

In the third step, the hyper-parameters of the classical models were tuned on the FCE validation set, and BERT was fine-tuned for three epochs. Table 3 lists the validation macro-F1 and the training time of every setting that was tried, together with the three fine-tuning epochs of BERT.

Table 3. Hyper-parameter search and BERT fine-tuning on the FCE validation set

Model	Setting	Validation macro-F1	Training time (s)
Naive Bayes	$\alpha = 0.01$	0.8287	0.10
	$\alpha = 0.05$	0.8295	0.10
	$\alpha = 0.1$ (selected)	**0.8308**	0.10
	$\alpha = 0.5$	0.8236	0.10
	$\alpha = 1.0$	0.8003	0.10
SVM	C = 0.05	0.8523	1.77
	C = 0.1	0.8569	1.91
	C = 0.5 (selected)	**0.8705**	2.87

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 171
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 167−178
DOI 10.47065/josh.v8i1.11296
https://ejurnal.seminar-id.com/index.php/josh/

Model	Setting	Validation macro-F1	Training time (s)
	C = 1.0	0.8702	3.89
Random Forest	**min. leaf size = 1 (selected)**	**0.8606**	12.13
	min. leaf size = 2	0.8523	5.05
BERT	epoch 1 (training loss 0.897)	0.8646	–
	epoch 2 (training loss 0.252)	0.9095	–
	epoch 3 (training loss 0.144, selected)	**0.9162**	3,640 (total)

Note: Random Forest used 300 trees and the 5,000 features with the highest χ^2 scores in both settings. The BERT training time is the total for all three epochs on the Apple MPS backend.

The classical models were not very sensitive to their hyper-parameters within the tested ranges, as Table 3 shows. Naive Bayes varied between 0.8003 and 0.8308, with the best result at $\alpha = 0.1$, and the SVM improved from 0.8523 at C = 0.05 to 0.8705 at C = 0.5 before levelling off at C = 1.0. Random Forest performed better with a minimum leaf size of 1 (0.8606) than with 2 (0.8523). BERT reached 0.8646 after the first epoch and 0.9162 after the third, while its training loss fell from 0.897 to 0.144, which indicates that the model was still improving on the validation set rather than overfitting it. In the final step, each model was refitted with its selected setting, and its predictions on the in-domain test set and on ICNALE-IDN were evaluated with the metrics, bootstrap intervals, and McNemar tests described in Section 2.6. The following subsections report these results.

3.2 Distribution of Grammatical Error Types

Figure 2 compares the share of each error class in the in-domain test set and in ICNALE-IDN. Determiner and preposition errors were the two largest classes in the training corpora, together accounting for about half of all grammatical edits. Determiner errors were also the most frequent class for the Indonesian students (24.7%). The largest difference concerned noun number. These errors made up 23.9% of the Indonesian students' grammatical errors, compared with 8.0% of the training edits and 7.8% of the in-domain test edits. Typical cases were omitted plural suffixes, as in "smoke in public place" (corrected to places) or "a special room for smoker" (smokers). This pattern is consistent with first-language transfer, because Indonesian nouns are not inflected for number. Preposition (16.6%) and tense errors (9.6%) were less frequent than in the training corpora. Figure 2 presents the proportion of every class in both test sets.

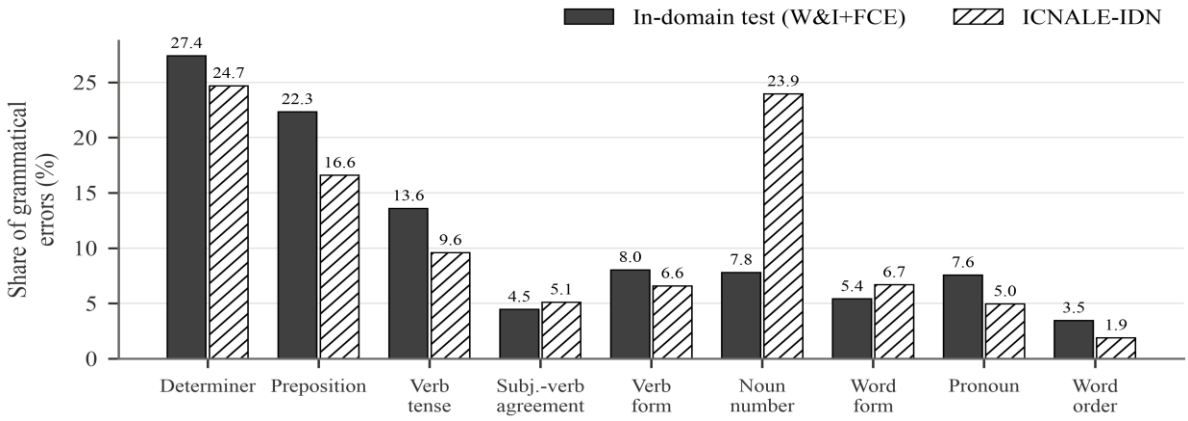


Figure 2. Share of each grammatical error class in the in-domain test set and in ICNALE-IDN

The distribution also varied across proficiency levels. Noun number errors made up 24.0% to 25.2% of the errors at levels A2 to B1.2, but only 17.4% at B2, while determiner errors rose to 38.7% at B2. The B2 group, however, contains only six essays, so this difference should be read as a tendency rather than a firm finding. The edit operations show a similar profile to earlier Indonesian error analyses. Of the 1,900 Indonesian grammatical edits, 66.0% replaced a word, 17.5% inserted a missing word, and 16.5% removed an unnecessary word. These three operations correspond roughly to the misformation, omission, and addition categories of the surface strategy taxonomy, for which Neman and Kamuh [3] reported 57.94%, 25.23%, and 10.28% in the descriptive texts of senior high school students in Manado. In both data sets, wrong forms dominate and omissions come second. The shift in class proportions matters for the next results: a model trained on the BEA-2019 distribution sees comparatively few noun number errors but must handle them as one of the two largest classes in the Indonesian data.

3.3 Overall Classification Performance

Table 4 and Figure 3 present the overall results. BERT achieved the best score on every metric for both test sets, with a macro-F1 of 0.917 in-domain and 0.849 on ICNALE-IDN. Among the classical models, the SVM was best

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 172
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 167−178
DOI 10.47065/josh.v8i1.11296
https://ejurnal.seminar-id.com/index.php/josh/

in-domain (0.862), while the SVM and Random Forest were practically tied on ICNALE-IDN (0.806 and 0.805). Naive Bayes was the weakest model on both test sets. Table 4 lists the accuracy, the macro-averaged precision and recall, and the macro-F1 with its 95% CI for every model.

Table 4. Overall results on the in-domain test set and on ICNALE-IDN

Model	In-domain test (W&I+FCE)				ICNALE-IDN				Δ
	Acc.	P	R	Macro-F1 [95% CI]	Acc.	P	R	Macro-F1 [95% CI]	F1
Naive Bayes	0.856	0.803	0.824	0.808 [0.794, 0.820]	0.823	0.767	0.800	0.773 [0.746, 0.795]	−0.035
SVM	0.899	0.865	0.861	0.862 [0.850, 0.874]	0.860	0.820	0.795	0.806 [0.780, 0.827]	−0.056
Random Forest	0.882	0.854	0.837	0.843 [0.831, 0.856]	0.864	0.834	0.784	0.805 [0.779, 0.824]	−0.038
BERT	**0.937**	**0.912**	**0.925**	**0.917** [0.908, 0.927]	**0.883**	**0.850**	**0.849**	**0.849** [0.825, 0.869]	−0.068

Note: P = macro-averaged precision; R = macro-averaged recall; Δ F1 = ICNALE-IDN minus in-domain macro-F1. Best values per column in bold.

Figure 3 compares the macro-F1 of the four models on the two test sets, with the bootstrap confidence intervals drawn as error bars.

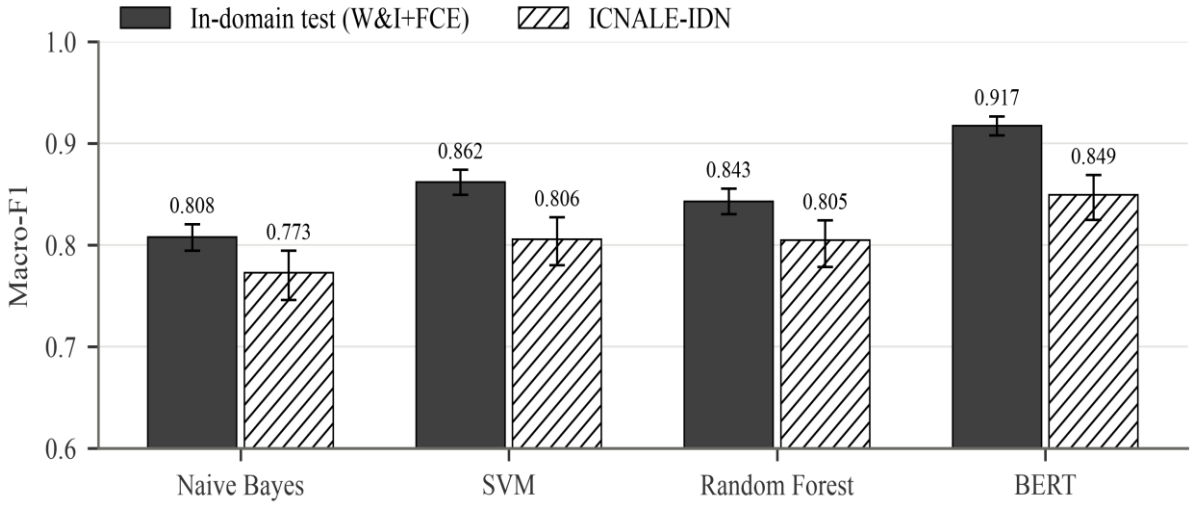


Figure 3. Macro-F1 of each model on both test sets (error bars show 95% bootstrap CI)

Every model performed worse on the Indonesian essays than in-domain. The drop in macro-F1 was 0.035 for Naive Bayes, 0.038 for Random Forest, 0.056 for the SVM, and 0.068 for BERT. BERT therefore lost the most in absolute terms, yet it remained clearly ahead, and the 95% CI of its ICNALE-IDN score (0.825–0.869) is higher than the CIs of all classical models except for a small overlap with the upper end of the SVM interval (0.780–0.827). The exact McNemar tests confirm the difference. On ICNALE-IDN, BERT was correct on 193 instances that Naive Bayes misclassified, against 79 in the opposite direction ($p < 0.001$). The corresponding counts were 123 versus 79 against the SVM ($p = 0.002$) and 133 versus 96 against Random Forest ($p = 0.017$). On the in-domain test set, all three differences were significant at $p < 0.001$.

The feature ablation helps to explain why the classical models degrade. When the SVM was trained on lexical features only, its macro-F1 fell from 0.862 to 0.811 in-domain (−0.051) and from 0.806 to 0.741 on ICNALE-IDN (−0.065). POS features therefore helped more under the corpus shift than in-domain, probably because grammatical categories such as plural nouns or third-person verbs remain stable across corpora, while the specific words that learners use change with topic and background. BERT does not use such hand-crafted features but learns contextual representations during pre-training [17]. Its larger absolute drop shows that these representations do not remove the effect of the corpus shift, although BERT still remained the most accurate model on the Indonesian essays.

3.4 Performance per Error Type

Table 5 gives the F1-score of each model for each class on ICNALE-IDN, and Figure 4 compares BERT's per-class scores across the two test sets. Determiner, preposition, and noun number were classified most reliably, with F1-scores higher than 0.90 for BERT and higher than 0.87 for every model. These three classes cover 65.2% of the Indonesian errors, so the most frequent problems of Indonesian students are also the ones the models recognise best. BERT's advantage was largest for word order (0.914 against at most 0.813) and subject-verb agreement

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 173
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 167−178
DOI 10.47065/josh.v8i1.11296
https://ejurnal.seminar-id.com/index.php/josh/

(0.818 against at most 0.778), although the word order class has only 36 instances. The best classical model differed by class. Random Forest achieved the highest F1 for noun number (0.925), slightly higher than BERT (0.924), which suggests that plural errors are marked by surface cues such as the -s suffix that sparse character n-grams capture well. The SVM was best for determiner (0.924) and tied with BERT for pronoun (0.763). Table 5 reports these per-class F1-scores together with the number of instances of each class.

Table 5. F1-score per error class on ICNALE-IDN

Class	n	Naive Bayes	SVM	Random Forest	BERT
Determiner	469	0.902	**0.924**	0.909	0.923
Preposition	315	0.894	0.928	0.917	**0.933**
Verb tense	182	0.758	0.817	0.836	**0.870**
Subject-verb agreement	97	0.778	0.687	0.701	**0.818**
Verb form	125	0.705	0.736	0.752	**0.807**
Noun number	455	0.875	0.909	**0.925**	0.924
Word form	127	0.594	0.677	0.673	**0.693**
Pronoun	94	0.670	**0.763**	0.750	**0.763**
Word order	36	0.781	0.813	0.781	**0.914**

Note: best value per row in bold.

Figure 4 contrasts the per-class F1-scores of BERT on the in-domain test set with those on ICNALE-IDN, which makes the classes most affected by the corpus shift visible.

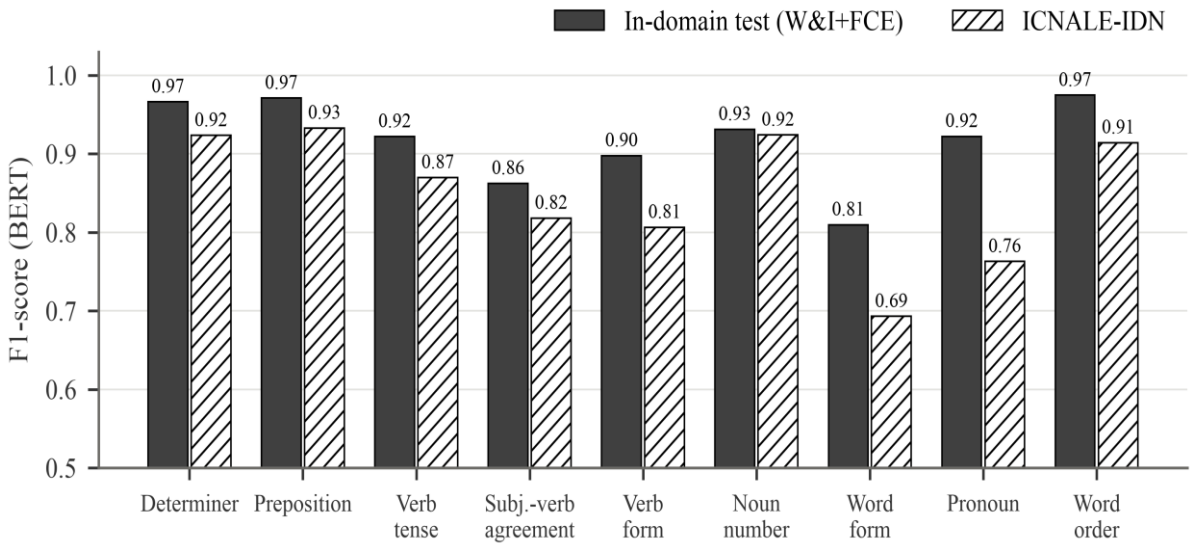


Figure 4. F1-score of BERT for each error class on both test sets

The hardest classes were word form (F1 of 0.693 for BERT) and pronoun (0.763). The pronoun class also showed the largest drop for BERT between the two test sets, from 0.922 to 0.763. Naive Bayes reached a relatively high F1 for subject-verb agreement on ICNALE-IDN (0.778), higher than its in-domain score (0.693), which suggests that the frequent Indonesian agreement errors, such as "It have many benefit", are lexically simple and depend on a small set of verb forms.

3.5 Error Analysis

The confusion matrix in Figure 5 shows where BERT failed on ICNALE-IDN. The three largest confusions were word form predicted as noun number (15.0% of word form instances), pronoun predicted as determiner (14.9%), and subject-verb agreement predicted as tense (12.4%). Many of these cases are linguistically ambiguous rather than clear model errors. ERRANT assigns a type from the POS tags and morphology of the changed tokens, so its label is a convention rather than a judgement about the learner's intention. In "The students which have part time job", ERRANT labels the change of which to that as a determiner error, while BERT predicted pronoun, which is arguably the better description of a relative pronoun. In "We know that lecture start in the morning", the edit to lectures was labelled as subject-verb agreement, while BERT predicted noun number. In "their score are decreasing", corrected to were, the sentence contains both an agreement and a tense problem, and BERT chose agreement instead of the reference label tense. Figure 5 presents the row-normalised confusion matrix of BERT on ICNALE-IDN, in which each row shows how the instances of one true class were distributed over the predicted classes, together with a panel that ranks the six largest confusions by their share of the true class.

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 174
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 167−178
DOI 10.47065/josh.v8i1.11296
https://ejurnal.seminar-id.com/index.php/josh/

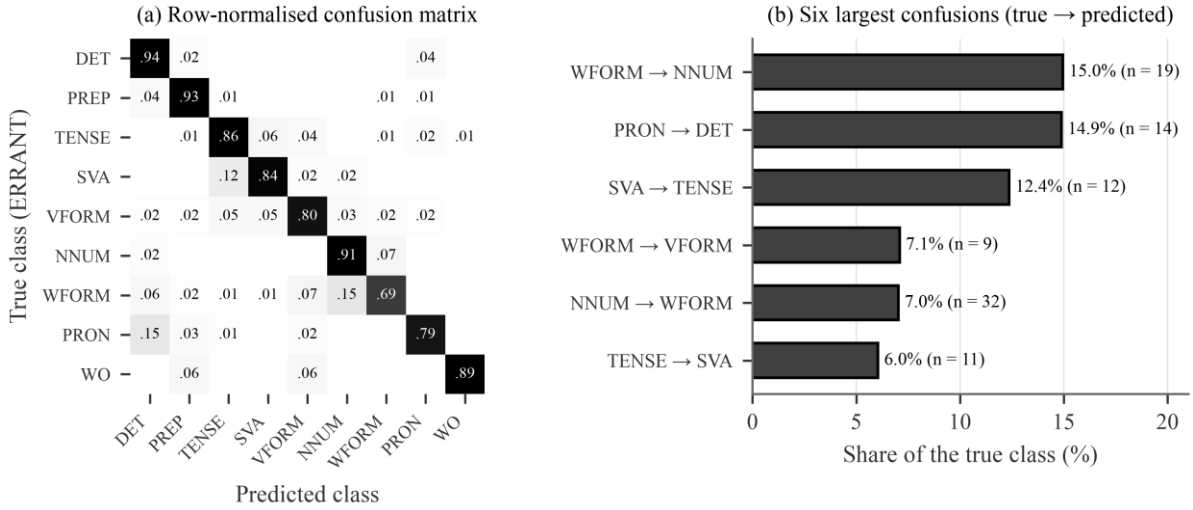


Figure 5. Confusion analysis of BERT on ICNALE-IDN: (a) row-normalised confusion matrix and (b) the six largest confusions

These examples show that part of the measured cross-corpus loss comes from the reference labels rather than from the classifier. The ICNALE essays were corrected by proofreaders rather than annotated edit by edit, and ERRANT had to reconstruct the edits from the two versions. Edits typed as OTHER, which typically arise from rewrites of longer phrases, made up 24.1% of the ICNALE edits compared with 11.2% of the training edits, which indicates that the Indonesian essays were edited less minimally. For classroom feedback, this means that the confusions between noun number and word form, and between agreement and tense, should be presented to learners as related categories rather than as definite diagnoses.

3.6 Performance across Proficiency Levels

Table 6 reports macro-F1 by CEFR level. For BERT, performance was stable from A2 to B1.2 (0.840 to 0.859) and accuracy stayed between 0.867 and 0.891 at all levels. The lower macro-F1 at B2 (0.712) should be interpreted with care, because the B2 group has only 155 error instances from six essays. With so few instances, rare classes such as word order and pronoun contain only a handful of examples, and one wrong prediction changes their F1 considerably. The results therefore do not indicate that the classifiers work worse for more proficient learners, but they show that larger samples are needed at B2. Among the classical models, the SVM was the most stable, with macro-F1 between 0.784 and 0.814 from A2 to B1.2 and the highest score of all models at B2 (0.761). Table 6 lists the number of error instances and the macro-F1 of each model at every CEFR level.

Table 6. Macro-F1 on ICNALE-IDN by CEFR level

Model	A2	B1.1	B1.2	B2
Error instances (n)	601	659	485	155
Naive Bayes	0.791	0.747	0.773	0.734
SVM	0.811	0.784	0.814	**0.761**
Random Forest	0.785	0.792	0.825	0.744
BERT	**0.840**	**0.859**	**0.847**	0.712

3.7 Discussion

The error profile found here agrees with recent Indonesian error analyses in its main lines but differs in the details. Antara and Anggreni [20] reported that article omission made up 42.2% of the grammatical problems in opinion essays, followed by subject-verb agreement (26.6%), prepositions (17.4%), and tense (13.8%). In the ICNALE data, determiner errors were also the largest class (24.7%) and prepositions accounted for a similar share (16.6%), but subject-verb agreement was much less frequent (5.1%) and noun number emerged as a second major class (23.9%). The difference is partly one of method: the earlier study grouped AI-detected problems in 17 essays into four broad categories, whereas this study derived nine classes from all proofreader edits in 66 essays, which separates plural errors from agreement errors. The dominance of article and plural errors is consistent with first-language interference, which Puspita and Suswanto [5] identified as the main source of article errors in Indonesian students' writing.

The model comparison matches the pattern reported in recent informatics studies. The SVM was the strongest classical model, as in the sentiment study of Elisa and Isnain [24], and BERT reached an in-domain macro-F1 of 0.917, close to the 0.9167 that Sabilla and Dhenabayu [23] reported for complaint classification. The ranking among the classical models also changed with the data, since Random Forest caught up with the SVM on ICNALE-IDN, which agrees with the observation of Mokodaser et al. [25] that model rankings depend on the task

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 175
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 167−178
DOI 10.47065/josh.v8i1.11296
https://ejurnal.seminar-id.com/index.php/josh/

and on careful tuning. The present study adds a dimension that those comparisons lacked: all models lost accuracy when they were applied to essays from a different population. This loss is consistent with the general domain-shift problem [26], with the finding that text from Indonesia and the US can differ systematically on the same topic [28], and with the evidence that GEC benchmarks overstate performance on harder data [27]. For Indonesian institutions, the practical implication is that a score reported on a public benchmark should not be taken as the expected accuracy on local essays.

These findings also bear on the feedback tools that Indonesian students already use. Grammarly has become the most studied tool in L2 writing research [6], and most students in one Indonesian study perceived ChatGPT as useful for learning English [11], but the judgements of such tools are rarely evaluated on Indonesian learner data, and concerns about the validity and transparency of LLM feedback remain [8], [10]. A dedicated classifier with a known, class-specific error rate offers a more transparent alternative. Its predictions for determiner, preposition, and noun number, which together cover about two thirds of the Indonesian errors, are reliable enough to support feedback, while predictions for word form and pronoun should be shown as suggestions for the teacher to confirm.

Four limitations should be noted. First, the ICNALE-IDN test set is small (66 essays by 33 students on two topics), so the per-class and per-level results have wide uncertainty. Second, the reference labels on ICNALE were derived automatically by ERRANT. A random sample of 200 instances has been prepared for manual validation by two annotators, but this validation had not been completed at the time of writing, so agreement between ERRANT and human judgement could not yet be reported. Third, BERT was fine-tuned with a single random seed, so the bootstrap CIs capture sampling uncertainty in the test data but not the variation between training runs. Fourth, the task assumes that the location of each error is known, and edits typed as OTHER, which made up 24.1% of all ICNALE edits, were excluded. A complete feedback system would also have to detect errors and handle these mixed edits.

4. CONCLUSION

This study set out to measure how well grammatical error-type classifiers trained on public learner corpora transfer to essays written by Indonesian students, and to compare classical and Transformer-based models under this corpus shift. Three findings answer the research objectives. First, the error profile of Indonesian students differs clearly from that of the training corpora: noun number errors made up 23.9% of their grammatical errors, about three times their share in the BEA-2019 training data, and determiner errors remained the most frequent class, which is consistent with first-language interference from Indonesian. Second, BERT was the most accurate model on both test sets, with a macro-F1 of 0.917 in-domain and 0.849 on the Indonesian essays, and its advantage over Naive Bayes, the SVM, and Random Forest was statistically significant in exact McNemar tests. All models lost between 0.035 and 0.068 macro-F1 under the corpus shift, and the SVM depended strongly on part-of-speech features to generalise. Third, determiner, preposition, and noun number errors, which together cover about two thirds of the Indonesian errors, were classified with F1-scores higher than 0.90 by BERT, whereas word form and pronoun remained the weakest classes, partly because ERRANT labels are ambiguous for relative pronouns and plural forms. A BERT classifier trained on public corpora can therefore support error-type feedback for Indonesian EFL writing, provided that predictions for the weaker classes are checked by a teacher. The study thereby contributes a reproducible ERRANT-based labelling pipeline for public and Indonesian learner corpora, a cross-corpus benchmark of classical and Transformer-based error-type classifiers supported by significance tests, and a class-level reliability profile that helps teachers decide which automated error-type feedback can be used directly in Indonesian EFL writing classes. The conclusions are limited by the small Indonesian sample of 66 essays on two topics, by reference labels that were derived automatically and not yet validated by human annotators, by a single BERT training run, and by the assumption that the location of each error is known. Future work should complete the manual validation of the ICNALE labels and report agreement between annotators and ERRANT, train BERT with several random seeds, extend the Indonesian test set with essays from other topics, institutions, and proficiency levels, add error detection and the treatment of mixed edits, and compare the classifier with LLM-based feedback tools in writing courses with lecturers and students.

5. DECLARATIONS

5.1 Author Contributions

Conceptualization: Meily Ivane Esther Neman and George Morris William Tangka;
Methodology: Meily Ivane Esther Neman, George Morris William Tangka, and Regi Fernando Najoan;
Software: George Morris William Tangka and Regi Fernando Najoan;
Validation: Meily Ivane Esther Neman, Regi Fernando Najoan, and Daniel Semuel Warouw;
Formal Analysis: George Morris William Tangka and Regi Fernando Najoan;
Investigation: Meily Ivane Esther Neman, George Morris William Tangka, and Daniel Semuel Warouw;
Resources: Meily Ivane Esther Neman and Daniel Semuel Warouw;

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 176
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 167−178
DOI 10.47065/josh.v8i1.11296
https://ejurnal.seminar-id.com/index.php/josh/

Data Curation: George Morris William Tangka and Regi Fernando Najoan;
Writing Original Draft Preparation: Meily Ivane Esther Neman and George Morris William Tangka;
Writing Review and Editing: Meily Ivane Esther Neman, George Morris William Tangka, Regi Fernando Najoan, Enrico Samuel Djimesha, and Daniel Semuel Warouw;
Visualization: George Morris William Tangka and Enrico Samuel Djimesha;
All authors have read and agreed to the published version of the manuscript.

5.2 Data Availability

The W&I+LOCNESS and FCE data are publicly available from the BEA-2019 shared task, and the ICNALE Edited Essays are available from the ICNALE project after registration. The code and the derived data used in this study are available on request from the corresponding author.

5.3 Funding

This research received no external funding. The publication and manuscript processing fees were paid by the authors.

5.4 Institutional Review Board Statement

Not applicable. The study used anonymised learner corpora that are distributed for research purposes.

5.5 Informed Consent Statement

Not applicable. No new data were collected from human participants.

5.6 Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

5.7 Generative AI and AI-assisted Technologies in The Writing Process

The authors used Generative AI to assist in drafting and editing the manuscript text, checking the bibliographic details of the references, and preparing the figures from the experimental results. The authors reviewed and edited all AI-assisted content and take full responsibility for the final publication.

REFERENCES

[1] M. H. Hamdani and N. Abid, “Indonesian EFL students' writing challenges: A qualitative meta-analysis”, Getsempena English Education Journal, vol. 12, no. 1, pp. 14–36, 2025. https://doi.org/10.46244/geej.v12i1.3091
[2] Fitrawati and D. Safitri, “Students' grammatical errors in essay writing: A pedagogical grammar reflection”, International Journal of Language Education, vol. 5, no. 2, pp. 74–88, 2021. https://doi.org/10.26858/ijole.v5i2.16128
[3] M. I. E. Neman and V. S. Kamuh, “Errors in English descriptive text produced by high school students in Manado”, ELS Journal on Interdisciplinary Studies in Humanities, vol. 7, no. 1, pp. 17–21, 2024. https://doi.org/10.34050/elsjish.v7i1.32782
[4] Y. Mubarok and T. Budiono, “An error analysis on EFL students' writing”, Englisia: Journal of Language, Education, and Humanities, vol. 9, no. 2, pp. 187–203, 2022. https://doi.org/10.22373/ej.v9i2.11386
[5] Puspita and Suswanto, “An analysis of English article errors and pedagogical implications in students' writing skill”, Buletin Poltanesa, vol. 24, no. 1, pp. 60–67, 2023. https://doi.org/10.51967/tanesa.v24i1.1972
[6] G. Dizon and J. M. Gayed, “A systematic review of Grammarly in L2 English writing contexts”, Cogent Education, vol. 11, no. 1, Art. no. 2397882, 2024. https://doi.org/10.1080/2331186X.2024.2397882
[7] A. R. S. Tambunan, W. Andayani, W. S. Sari, and F. K. Lubis, “Investigating EFL students' linguistic problems using Grammarly as automated writing evaluation feedback”, Indonesian Journal of Applied Linguistics, vol. 12, no. 1, pp. 16–27, 2022. https://doi.org/10.17509/ijal.v12i1.46428
[8] E. Kasneci et al., “ChatGPT for good? On opportunities and challenges of large language models for education”, Learning and Individual Differences, vol. 103, Art. no. 102274, 2023. https://doi.org/10.1016/j.lindif.2023.102274
[9] S. Lin, “Evaluating LLMs' grammatical error correction performance in learner Chinese”, PLoS ONE, vol. 19, no. 10, Art. no. e0312881, 2024. https://doi.org/10.1371/journal.pone.0312881
[10] A. Pack, A. Barrett, and J. Escalante, “Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability”, Computers and Education: Artificial Intelligence, vol. 6, Art. no. 100234, 2024. https://doi.org/10.1016/j.caeai.2024.100234
[11] D. F. E. Meluwu and M. I. E. Neman, “Students' perception towards the use of ChatGPT as a tool for learning English”, Global Education Journal, vol. 3, no. 2, pp. 206–216, 2025. https://doi.org/10.59525/gej.v3i2.712
[12] D. S. Warouw and M. I. E. Neman, “The use of English language learning strategies in learning as foreign language”, ELS Journal on Interdisciplinary Studies in Humanities, vol. 7, no. 1, pp. 32–42, 2024. https://doi.org/10.34050/elsjish.v7i1.32847
[13] C. Bryant, Z. Yuan, M. R. Qorib, H. Cao, H. T. Ng, and T. Briscoe, “Grammatical error correction: A survey of the state of the art”, Computational Linguistics, vol. 49, no. 3, pp. 643–701, 2023. https://doi.org/10.1162/coli_a_00478
[14] Y. Wang, Y. Wang, K. Dang, J. Liu, and Z. Liu, “A comprehensive survey of grammatical error correction”, ACM Transactions on Intelligent Systems and Technology, vol. 12, no. 5, Art. no. 65, 2021. https://doi.org/10.1145/3474840

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 177
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 167−178
DOI 10.47065/josh.v8i1.11296
https://ejurnal.seminar-id.com/index.php/josh/

[15] C. Bryant, M. Felice, and T. Briscoe, "Automatic annotation and evaluation of error types for grammatical error correction", in Proc. 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, Canada, 2017, pp. 793–805. https://doi.org/10.18653/v1/P17-1074
[16] C. Bryant, M. Felice, Ø. E. Andersen, and T. Briscoe, "The BEA-2019 shared task on grammatical error correction", in Proc. Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, Florence, Italy, 2019, pp. 52–75. https://doi.org/10.18653/v1/W19-4406
[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding", in Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, 2019, pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423
[18] Q. Li et al., "A survey on text classification: From traditional to deep learning", ACM Transactions on Intelligent Systems and Technology, vol. 13, no. 2, Art. no. 31, 2022. https://doi.org/10.1145/3495162
[19] B. Min et al., "Recent advances in natural language processing via large pre-trained language models: A survey", ACM Computing Surveys, vol. 56, no. 2, Art. no. 30, 2023. https://doi.org/10.1145/3605943
[20] I. M. A. R. Antara and N. P. Y. Anggreni, "Analyzing grammatical errors in EFL students' opinion essays using DeepSeek AI: Patterns and pedagogical implications", Stilistika: Jurnal Pendidikan Bahasa dan Seni, vol. 13, no. 2, pp. 210–218, 2025. https://doi.org/10.59672/stilistika.v13i2.4560
[21] M. Qin, "A study on automatic correction of English grammar errors based on deep learning", Journal of Intelligent Systems, vol. 31, no. 1, pp. 672–680, 2022. https://doi.org/10.1515/jisys-2022-0052
[22] A. Musyafa, Y. Gao, A. Solyman, C. Wu, and S. Khan, "Automatic correction of Indonesian grammatical errors based on transformer", Applied Sciences, vol. 12, no. 20, Art. no. 10380, 2022. https://doi.org/10.3390/app122010380
[23] P. Sabilla and R. Dhenabayu, "Comparative complaint text classification using SVM, CNN, and BERT with priority scoring for NYC 311 service requests", Jurnal Riset Informatika, vol. 8, no. 4, pp. 576–590, 2026. https://doi.org/10.34288/jri.v8i4.589
[24] P. Elisa and A. R. Isnain, "Comparison of random forest, support vector machine and naive bayes algorithms to analyze sentiment towards mental health stigma", Jurnal Teknik Informatika (JUTIF), vol. 5, no. 1, pp. 321–329, 2024. https://doi.org/10.52436/1.jutif.2024.5.1.1817
[25] W. G. Mokodaser, T. I. Soewignyo, G. M. W. Tangka, and F. Soewignyo, "Forecasting health insurance payer income: A comparative analysis of decision tree and SVR algorithms", Jurnal Riset Informatika, vol. 7, no. 3, pp. 210–218, 2025. https://doi.org/10.34288/jri.v7i3.369
[26] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 4, pp. 4396–4415, 2023. https://doi.org/10.1109/TPAMI.2022.3195549
[27] V. Starchenko et al., "Fighting evaluation inflation: Concentrated datasets for grammatical error correction", Journal of Language and Education, vol. 10, no. 4, pp. 112–129, 2024. https://doi.org/10.17323/jle.2024.22272
[28] G. M. W. Tangka, I. R. Chrisanti, J. Waworundeng, R. C. Maringka, and G. A. Sandag, "Sentiment analysis and topic detection on post-pandemic healthcare challenges: A comparative study of Twitter data in the US and Indonesia", CogITo Smart Journal, vol. 10, no. 2, pp. 561–579, 2024. https://doi.org/10.31154/cogito.v10i2.819.561-579
[29] I. N. Oktavianti and I. Prayogi, "Discourse functions of lexical bundles in Indonesian EFL learners' argumentative essays: A corpus study", Studies in English Language and Education, vol. 9, no. 2, pp. 761–783, 2022. https://doi.org/10.24815/siele.v9i2.23995
[30] S. Ishikawa, The ICNALE Guide: An Introduction to a Learner Corpus Study on Asian Learners' L2 English. London: Routledge, 2023.
[31] H. Yannakoudakis, T. Briscoe, and B. Medlock, "A new dataset and method for automatically grading ESOL texts", in Proc. 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, 2011, pp. 180–189.
[32] L. Tunstall, L. von Werra, and T. Wolf, Natural Language Processing with Transformers: Building Language Applications with Hugging Face, Revised ed. Sebastopol, CA: O'Reilly Media, 2022.

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 178
This work is licensed under a Creative Commons Attribution 4.0 International License.