Jurnal Teknik Informatika (JUTIF) Vol. 6, No. 6, December 2025, Page. 5976-5987
P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5501

Complex Word Identification in Indonesian Children's Texts: An IndoBERT Baseline and Error Analysis

Lisnawita*[1], Juhaida Abu Bakar[2], Ruziana Mohamad Rasli[3], Loneli Costaner[4], Guntoro[5]

[1,4,5]Faculty of Computer Science, Universitas Lancang Kuning, Indonesia
[2]Data Science Research Lab, School of Computing, Universiti Utara Malaysia, Malaysia
[3]School of Multimedia Technology & Communication, Universiti Utara Malaysia, Malaysia

Email: [1]lisnawita@unilak.ac.id

Received : Dec 1, 2025; Revised : Dec 17, 2025; Accepted : Dec 20, 2025; Published : Dec 31, 2025

Abstract

Complex Word Identification (CWI) is a crucial step for building text simplification systems, especially for Indonesian children's reading materials where unfamiliar vocabulary can hinder comprehension. This study formulates token-level CWI for Indonesian children's texts and establishes two baselines: an interpretable rule-based model using linguistic features e.g., length, syllable heuristics, and affix patterns, and an IndoBERT model fine-tuned for token classification. This study construct and annotate a children's text corpus and evaluate both approaches using standard classification metrics. On the test set (22.584 tokens), IndoBERT achieves an F1-score of 0.9972 for the CWI class, substantially outperforming the rule-based baseline (F1 = 0.8607). The IndoBERT system makes only 39 errors (23 false positives and 16 false negatives), indicating near-perfect performance under the evaluated setting. Furthermore, this study provides an error analysis to highlight remaining failure patterns and borderline cases that are difficult even for contextual models. The resulting benchmark and findings contribute to Informatics/Computer Science by providing a strong baseline and analysis for educational NLP in a low-resource language setting, supporting the development of Indonesian child-oriented NLP resources and downstream text simplification tools.

Keywords: *complex word identification, error analysis, IndoBERT, Indonesian children's texts, text simplification, token classification*

1. INTRODUCTION

Vocabulary comprehension is one of the main components of successful reading literacy in children. In practice, many texts consumed by children such as school textbooks, supplementary reading materials, and public information content still contain words that are long, infrequent, or morphologically complex. Such words can become barriers to comprehension, increase cognitive load, and reduce reading motivation, especially for beginning readers or children with learning difficulties.

In Natural Language Processing (NLP), one widely used approach to reduce lexical barriers is text simplification (TS) [1], [2], [3]. TS is commonly viewed as a pipeline of several stages, starting from the identification of difficult parts of the text, followed by the selection of simpler alternatives, and ending with the realization of a new sentence. A crucial first step in this pipeline is Complex Word Identification (CWI), the task of deciding whether a word in each context is complex for a specific target readership. The quality of the CWI module is critical: if too many words are incorrectly flagged as complex, the resulting text will be over-simplified, whereas complex words that are missed will remain obstacles to understanding.

Internationally, CWI research has progressed rapidly in various languages, especially English [4], [5], [6], [7], [8] targeting diverse groups such as second-language (L2) learners [9] and people with specific disabilities. A wide range of approaches has been proposed, from rule-based methods relying on lexical–syntactic features, to classical machine learning models, and more recently to models based on pre-trained language models such as BERT [10]. However, for Indonesian, CWI studies are still very limited, and most work related to text simplification [11] does not formulate CWI as a separate task.

In addition, the broader research community has also explored CWI and lexical complexity prediction across shared tasks and benchmark datasets, including feature-based, kernel-based, and deep learning architectures, as well as transformer-centric solutions that combine contextual embedding and linguistic signals.[12], [13], [14], [15], [16], [17], [18], [19]

At the same time, the emergence of pre-trained language models for Indonesian, such as IndoBERT opens opportunities to build more accurate CWI systems for this language. IndoBERT has been shown to be effective on various tasks such as text classification [20], [21], [22], [23] and part-of-speech tagging [24], [25] [26]Formulating CWI as a token classification task on top of IndoBERT is therefore an attractive approach, especially when it is systematically compared to a linguistically motivated rule-based baseline

Based on previous work, this study identifies two main gaps. First, in terms of language and domain, existing CWI studies are still dominated by English and general-domain texts; for Indonesian, there is almost no work that explicitly formulates CWI on children's texts, even though the lexical and pragmatic characteristics of children's texts differ from those of adult texts. Second, there is no strong IndoBERT-based baseline for CWI on Indonesian children's texts, nor a systematic comparison with a rule-based linguistic approach. As a result, there is no clear benchmark on how much pre-trained models can improve CWI performance for the purpose of children's text simplification.

To address these gaps, this study makes three main contributions: it explicitly formulates the CWI task for Indonesian children's texts, together with a binary token-level annotation scheme; it leverages an annotated CWI corpus in the children's text domain as a basis for evaluation; and it establishes a linguistically motivated rule-based baseline and a strong IndoBERT-based token classification baseline, which are systematically compared on the same test set, thus providing a clear benchmark for the development of CWI systems and children's text simplification pipelines in Indonesian

2. METHOD

Figure 1. illustrates the methodology flowchart of this study

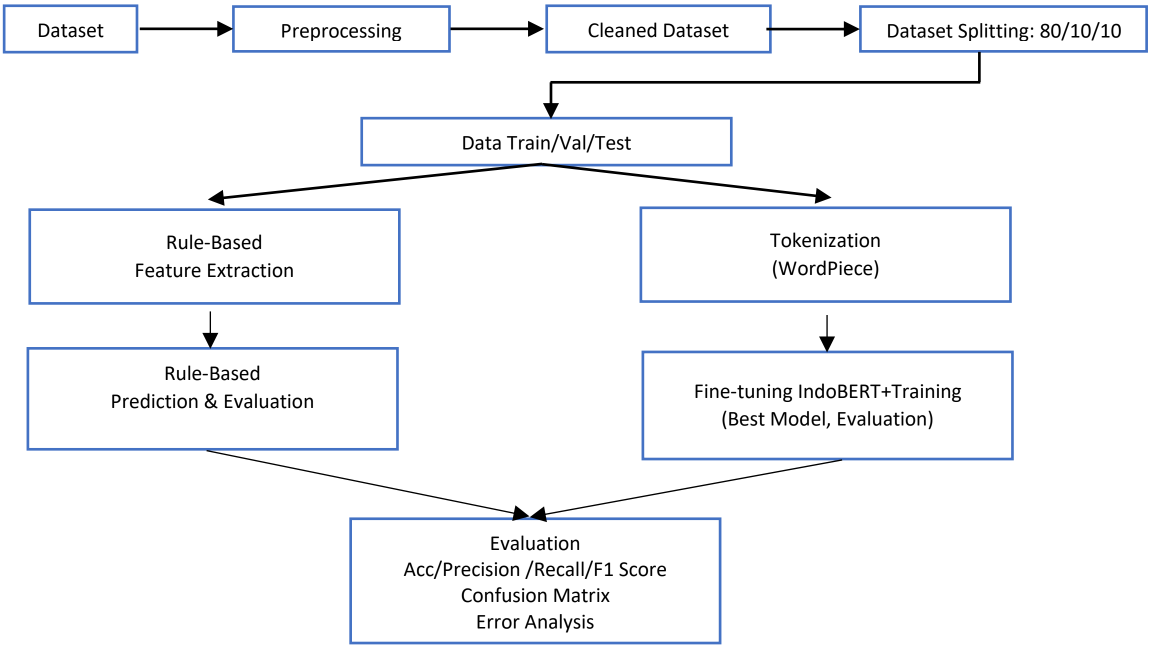

Figure 1. Methodology Flowchart

2.1. Dataset and Annotation Scheme

Dataset for this study were collected from child-oriented Indonesian reading materials. All documents were processed to remove non-linguistic noise while preserving the original lexical content relevant to complexity judgments. All CWI experiments in this study are defined and evaluated on an Indonesian children's text corpus that is manually annotated at the word level. dataset [11] in is not specifically targeted at children; it is used only as a supporting lexical resource, while all complexity annotations and model training are carried out exclusively on children's texts.The CWI dataset is annotated with binary labels at the word (token) level:

1. 0 (NonCWI): the word is considered non-complex for child readers,
2. 1 (CWI): the word is considered complex and potentially difficult for child readers.

Each row in the original corpus contains the following fields: sentence_id, token index within the sentence, the surface form of the word (token), and the complexity label. During preprocessing, tokens with the same sentence_id are grouped back into a single sentence entry, yielding per-sentence representations as a list of tokens and a parallel list of labels.

The dataset is then split into training, validation, and test sets with proportions of approximately 80% : 10% : 10% at the sentence level. This split is designed to preserve the distribution of complex and non-complex words in each subset.

In the test set, there are 22,584 labeled tokens, with the following distribution: 15.617 NonCWI tokens and 6.967 CWI tokens.

Thus, about 30% of the tokens in the test set are categorized as complex. This proportion indicates that, although the texts are aimed at children, there is still a substantial amount of vocabulary that is judged as potentially difficult and therefore relevant for the CWI task.The summary statistics of the Indonesian CWI corpus are shown in Table 1

Table 1. Presents the summary statistics of the Indonesian CWI corpus

Statistic	**Value**
Number of sentences	10.012
Tokens in test set	22.584
NonCWI tokens in test set	15.617
CWI tokens in test set	6.967
Proportion of CWI tokens	30%

2.2. Annotators and Annotation Procedure

The complexity labels in the CWI corpus were produced by a single expert annotator with a background in Indonesian language education and experience in working with children's reading materials. The annotator was familiar with the school curriculum and typical vocabulary exposure for primary school students. Although using a single annotator limits the possibility of measuring inter-annotator agreement, it ensures internal consistency in the application of the annotation guidelines.

Annotation was carried out in several iterations. In the initial phase, a small subset of sentences was annotated and used to refine the guidelines for what should be considered complex for child readers. The main criteria included: words that are long or morphologically complex (e.g., multiple affixes or derivational morphology), infrequent or specialized terms that are unlikely to appear in early-grade textbooks, loanwords and foreign names that may be unfamiliar to children, and words whose meaning is abstract or conceptually demanding relative to typical primary school content. In subsequent phases, the annotator applied these criteria consistently across the corpus, revisiting ambiguous cases when necessary.

2.3. Preprocessing

Before being used for model training, the annotated corpus went through a series of preprocessing steps. First, the texts were cleaned from irrelevant characters, such as HTML tags, emoticons, and non-alphanumeric symbols that do not provide meaningful linguistic information. Normalization was then applied to several forms of writing, including unifying punctuation variants, removing extra spaces, and handling numbers and abbreviations according to the annotation guidelines.

Next, the texts were segmented into sentences and word tokens using sentence splitters and tokenizers that are consistent with the modeling needs. At this stage, alignment between the preprocessed tokens and the CWI annotation labels was performed so that each token in the corpus has a valid label. The final output of preprocessing is a cleaned dataset in a structured format, consisting of (token, label) pairs for every sentence, which is ready to be used for data splitting and for training both the rule-based and IndoBERT models.

2.4. Data Splitting

After preprocessing, the cleaned dataset was divided into three subsets: training (train)**,** validation (val)**,** and test data. The splitting was carried out using a proportion of 80%:10%:10%**.** The training set is used to fit the models, the validation set is used for hyperparameter tuning and monitoring overfitting, while the test set is used only once at the end to objectively measure the final model performance.

The splitting process was conducted at the document or source-text level to prevent data leakage, ensuring that sentences from the same original document do not appear in more than one subset. Furthermore, the splitting procedure was designed so that the distribution of CWI and NonCWI labels in the three subsets remains comparable to the label distribution in the original corpus. To ensure reproducibility, sampling was performed randomly using a fixed random seed. In this way, the resulting train/val/test partitions are consistent across runs and can be reused for both the rule-based baseline and the IndoBERT-based model experiments.

2.5. Rule-based Baseline

As an interpretable baseline, this study develops a simple rule-based model that classifies each token as complex or non-complex based on several linguistic features:

1. word length (number of characters),
2. number of syllables (estimated using a vowel-based heuristic),
3. presence of common Indonesian affixes (prefixes such as *meng-*, *men-*, *mem-*, *me-*, *peng-*, *pen-*, *pem-*, *pe-*, *ber-*, *di-* and suffixes such as *-kan*, *-an*),
4. proper names (capitalization patterns and specific orthographic cues),
5. non-ASCII characters (e.g., accented or foreign characters).

Each feature contributes to a simple scoring function whose output is compared against a threshold to decide whether a token is labeled as CWI or NonCWI. The feature thresholds and weights are tuned on the validation set using a small grid search to maximize the F1-score for the CWI class. This rule-based model serves as a linguistically motivated and fully interpretable baseline for comparison with the IndoBERT-based approach.

2.6. IndoBERT for Token Classification

The main approach in this study uses IndoBERT as a token classification model for CWI. The modeling details are as follows:

1. Base model: indobenchmark/indobert-base-p1,
2. Tokenization: WordPiece with AutoTokenizer. Inputs are lists of tokens per sentence, passed with the argument is_split_into_words=True,
3. Maximum sequence length: 256 subwords,
4. Labeling strategy: only the first subword of each word receives the original label (0/1), while subsequent subwords are assigned the label –100 so that they are ignored during training.

The output layer is a token classification head with two classes (NonCWI and CWI) on top of IndoBERT. The model is trained via full fine-tuning using the following TrainingArguments configuration, The hyperparameters are shown in Table 2.

The compute_metrics function calculates precision, recall, F1, and accuracy for the positive class (CWI), ignoring positions with label –100. The model is evaluated on the validation set at the end of each epoch, and the checkpoint with the best F1-score is used for testing

Table 2. Training hyperparameters of IndoBERT for the CWI task

Hiperparameter	**Value**
Learning rate	3×10^{-5}
Batch size (train)	8
Batch size (eval)	16
Maximum number of epochs	3
Evaluation strategy	"epoch"
Save strategy	"epoch"
Best model selection	load_best_model_at_end = True, main metric F1
Seed	42
Early stopping	patience = 2 (training stops if validation F1 does not improve for two epochs)

2.7. Evaluation

Evaluation is carried out on the test set under two main scenarios:

1. Basic IndoBERT evaluation:
2. IndoBERT predictions are obtained using the default probability threshold of 0.5 for the CWI class.
3. Calibrated evaluation and direct comparison with the rule-based baseline:
    1. The predicted CWI probabilities from IndoBERT are swept over several threshold values.
    2. An operational threshold τ ≈ 0.21 is selected to provide a good balance of F1 for the CWI class on the development set.
    3. Confusion matrices and classification reports are produced for:
        1. IndoBERT, and
        2. the rule-based baseline, on the same set of 22,584 test tokens.

Across all scenarios, the evaluation metrics include accuracy, precision, recall, and F1-score for the CWI class, as well as macro F1 over both classes

2.8. Fine-tuning IndoBERT for CWI

Let a sentence $S = [w_1, \ldots, w_n]$with word-level labels $Y = [y_1, \ldots, y_n]$, where $y_i \in \{1,0\}$marks whether word $w_i$is complex (1) or non-complex (0). This study formulate the CWI task as a sequence labeling problem at the word level.

This study use IndoBERT [24] adopted from [27] as a pre-trained Transformer encoder and add a token classification head on top of it. Before entering the model, the sentence is split into words and then tokenized into subwords using WordPiece. Since one word can be split into multiple subwords, this study use word_ids to track the mapping from subwords back to their original words: only the first subword of each word receives the gold label, while all subsequent subwords are masked (assigned the label –100) so that they are ignored by the loss function.

For each token position, IndoBERT produces an encoded representation h_i. A linear head then produces two logits

$$z_i = W_o h_i + b_o, \tag{1}$$

and the probability of the “complex” class is obtained via softmax:

$$p_i = \text{softmax}(z_i)_{[CWI]}. \tag{2}$$

Training minimizes a masked cross-entropy loss over labeled positions only:

$$\mathcal{L} = -\sum_i m_i \left(y_i \log p_i + (1 - y_i) \log(1 - p_i)\right), \tag{3}$$

where $m_i = 1$for the first subword of each word and $m_i = 0$for all other subwords. Special tokens such as [CLS] and [SEP] are automatically added by the tokenizer but are not used as labeled positions.

At inference time, the model outputs a score $p_i$for each word. A binary prediction is then obtained with a threshold τ:

$$\hat{y}_i = \begin{cases} 1, & \text{if } p_i \geq \tau, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

In this study, a single operational threshold is used $\tau \approx 0.21$chosen on the development set to control the precision–recall trade-off for the CWI class

2.9. Algorithm: Inference Pipeline for IndoBERT-CWI

The following pseudo-code summarizes the inference pipeline for the IndoBERT-based CWI model:

Algorithm 1. CWI with IndoBERT (token classification)
Input: sentence S; threshold τ(chosen on the development set)
Output: ranked list $L = [(w, p_{\text{CWI}})]$of complex words

```
1.  words ← tokenize_to_words($S$)
2.  enc ← WordPieceTokenizer(words,is_split_into_words = True)
3.  logits ← IndoBERT_TokenClassifier(enc)   →IndoBERT fine-tuned for CWI
4.  probs_sub ← softmax(logits)[:,1]   →per-subword probability of the complex class
5.  For each word $i$:
        $p_{\text{CWI}}[i]$ ← probs_sub[first_subword_of_word$_i$] →via word_ids mapping
6.  $y[i]$ ← 1if $p_{\text{CWI}}[i] \geq \tau$, else 0
7.  $L$ ← sort_desc({(words[$i$], $p_{\text{CWI}}[i]$) | $y[i] = 1$})
8.  return $L$
```

3. RESULTS

This section presents the experimental results and analysis of the proposed complex word identification models on the Indonesian children's text corpus. Two modeling strategies are evaluated, namely a linguistically motivated rule-based baseline and a fine-tuned IndoBERT[20], [23] token classification model. Both models are trained and tested on the same train/validation/test split, and their performance is reported in terms of accuracy, precision, recall, and F1 for the CWI class, complemented by confusion matrices and an error analysis to better understand the patterns of misclassification.

This Section Reports Results In The Same Order As The Methodology: Rule-Based Baseline, Indobert Token Classification, And Comparative/Error Analysis. Overall, Indobert Achieves F1 = 0.9972 On The Cwi Class With 39 Total Errors On The Test Set, While The Rule-Based Baseline Reaches F1 = 0.8607, Indicating A Substantial Performance Gap Between Contextual And Feature-Driven Approaches.

3.1. Rule-based Model Results

With the best configuration obtained on the validation set, the rule-based baseline achieves the following performance on the test set. Table 3 shows the rule-based baseline results on the test set.

Table 3. Experiment results of the rule-based baseline on the test set

Metric	**Value**
Accuracy	0.9133
Precision	0.8529
Recall	0.8687
F1 Score	0.8607

The confusion matrix is shown in Figure 2

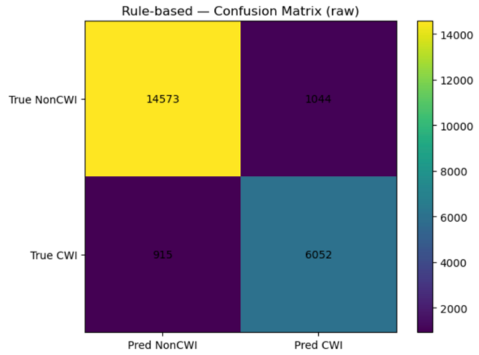

Figure 2. Confusion matrix (rule-based baseline)

Out of 22.584 tokens, there are 1.044 *false positives* (non-complex words incorrectly flagged as complex) and 915 *false negatives* (complex words that are not detected). The relatively high recall 0.8687 indicates that the baseline tends to be “aggressive” in marking complex words, but at the cost of a substantial number of false positives.

From a linguistic perspective, the rule-based model tends to label as complex:

1. long words,
2. words with many syllables,
3. words with derivational affixes,

without taking into account frequency or context. This explains why several high-frequency words that are relatively easy for children are still classified as complex (over-flagging)

3.2. IndoBERT Model Results

When evaluated with the default probability threshold of 0.5, IndoBERT already achieves very high performance on both the validation and test sets F1 for the CWI class around 0.993. After calibrating the decision threshold and aligning evaluation with the rule-based setting, the final IndoBERT results on the same 22.584 test tokens are as follows. The IndoBERT results on the test set are shown in Table 4.

Table 4. IndoBERT results on the test set

Metric	**Value**
Accuracy	0.9983
Precision	0.9967
Recall	0.9977
F1 Score	0.9972

The confusion matrix is shown in Figure 3

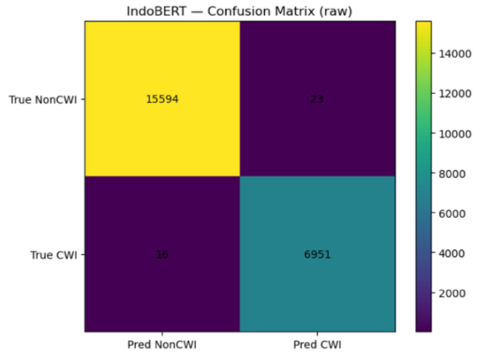

Figure 3. Confusion matrix (IndoBERT)

On the same 22,584 test tokens, IndoBERT produces only 23 false positives and 16 false negatives. In practical terms, this means that IndoBERT almost always agrees with the human annotation in deciding whether a word is complex for children.

3.3. Comparative Analysis IndoBERT vs Rule-Based

Table 3 and Table 4 show that the IndoBERT model consistently outperforms the rule-based baseline on all evaluation metrics. To highlight these differences, Table 5 summarizes their performance side by side on the same test set of 22,584 tokens.

Table 5. Comparison IndoBERT vs Rule-based

Metric	IndoBERT	Rule-based
Accuracy	0.9983	0.9133
Precision	0.9967	0.8529
Recall	0.9977	0.8687
F1 Score	0.9972	0.8607

The Comparison Experiment Result (Rule-based - IndoBERT) is shown in Figure 4.

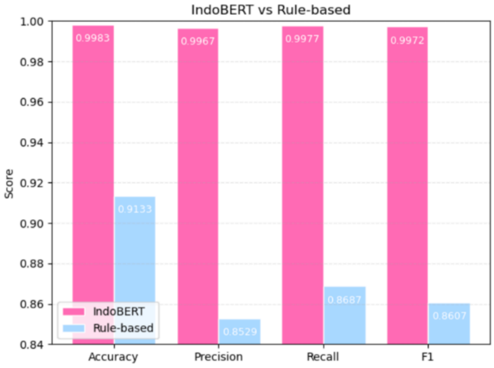

Figure 4. Comparison Experiment Result (Rule-based - IndoBERT)

The absolute difference in F1 for the CWI class is approximately 0.1365, which is substantial given that both models are evaluated on the same gold-standard annotations. In addition, the total number of misclassified tokens drops from 1,959 (1,044 false positives and 915 false negatives) for the rule-based model to only 39 (23 false positives and 16 false negatives) for IndoBERT.

These results indicate that IndoBERT is able to capture contextual cues that are inaccessible to the rule-based model. For example, the rule-based baseline relies heavily on surface form such as word length and the presence of affixes so it cannot distinguish between long but familiar words and genuinely rare or conceptually difficult terms. IndoBERT, on the other hand, can use surrounding context to infer whether a word is used as part of a technical expression, a proper name, or a relatively simple description, and therefore makes more accurate complexity judgments. This strong performance supports the use of pre-trained language models as a new baseline for CWI in Indonesian.

3.4. Error Analysis

To better understand the remaining errors, this study constructs an error table for IndoBERT's on the test set. Each token is categorized into one of the following classes:

1. FN (false negative): gold label = CWI, prediction = NonCWI,
2. FP (false positive): gold label = NonCWI, prediction = CWI,
3. OK: prediction matches the gold label.

Since the total number of errors is very small (39 cases out of 22,584 tokens), the patterns that emerge are mostly associated with borderline or ambiguous cases:

1. Some *false negatives* occur on words that look relatively common in form and frequency but are considered complex in the annotation because of their high conceptual load (technical terms or abstract concepts). The model appears to be "fooled" by their similarity in surface form to simpler words.
2. Some *false positives* occur on proper names or terms that appear frequently in the corpus, where the annotator judged them as non-complex while the model labeled them as complex, possibly because these words tend to co-occur with heavier or more technical contexts (e.g., scientific or formal news topics).

Although the overall reliability of IndoBERT is very high, these patterns suggest that lexical complexity is not determined solely by surface form and contextual distribution, but also by factors beyond the text itself, such as age of acquisition, curricular exposure, and children's real reading experiences. Integrating additional resources such as age-graded lexicons or curriculum-based word lists—could help align CWI decisions more closely with the profiles of real child readers

4. DISCUSSIONS

The experimental results show that IndoBERT can achieve almost perfect performance on the CWI task in the Indonesian children's text corpus, with an F1 of 0.9972 for the CWI class and only 39 misclassified tokens out of 22,584. In comparative terms, the F1 difference of about 0.13 points relative to the rule-based baseline indicates that the contextual and distributional information modeled by IndoBERT is far more informative than simple surface features such as word length, number of syllables, and affixes. This finding is consistent with trends in CWI research for other languages [5], [7], [28], [29], [30], [31], [32], [33]where models based on pre-trained language models typically outperform traditional lexical–syntactic approaches.

However, such high scores must also be interpreted critically. First, the corpus used in this study is still relatively homogeneous in terms of domain and genre, so its vocabulary and sentence patterns may not fully represent the variety of children's texts found in classrooms and other media (e.g., storybooks, popular science books, or exam materials). Second, lexical complexity annotation is carried out by a single annotator following practical guidelines that focus on word length, morphology, and perceived novelty of loanwords. While this brings internal consistency, it limits our ability to measure agreement across multiple child readers. In other words, the current IndoBERT model primarily learns to reproduce the decisions of a single annotator on a specific domain, and its external validity to other populations and domains remains to be tested.

The error analysis also indicates that, even with a small number of errors, FP and FN patterns are not trivial. The model tends to over-flag long words that are in fact familiar to children, and under-flag proper names and loanwords whose frequency is increasing in the corpus. This suggests that lexical complexity depends not only on surface form and contextual distribution, but also on extra-textual factors such as age of acquisition, curricular exposure, and socio-cultural context. In future work, integrating additional resources such as age-graded lexicons or school curriculum word lists may help calibrate CWI decisions to better match real child readers.

From an application perspective, the results indicate that the IndoBERT-based CWI module is mature enough to be used as a detector of complex words in children's text simplification pipelines. Nevertheless, given the current limitations in domain coverage and annotation, practical deployment should include a human-in-the-loop mechanism, for example by using CWI outputs as a ranked list of candidate words for teachers or editors to review, rather than as fully automatic final decisions. Moreover, using CWI to highlight words that need glossaries, illustrations, or graded substitutions appears to be safer and more pedagogically sound than mechanically replacing all complex words. With such a design, a strong CWI model can serve as an assistive tool within the broader ecosystem of children's literacy, rather than completely replacing human judgment

From an Informatics/Computer Science perspective, this work contributes to educational NLP and low-resource language technology by establishing a strong IndoBERT baseline and an interpretable comparison point for Indonesian children's CWI. The benchmark enables systematic component evaluation for Indonesian text simplification pipelines and can support real-world applications such as adaptive reading tools, digital learning platforms, and human-in-the-loop simplification systems that prioritize child comprehension and literacy outcomes.

5. CONCLUSION

This study investigates the task of Complex Word Identification (CWI) in Indonesian children's texts by leveraging an annotated corpus and comparing two main approaches: a linguistically motivated rule-based baseline and a pre-trained IndoBERT model for token classification. The corpus consists of 10,012 children's sentences with binary token-level annotations, yielding 22,584 labeled tokens in the test set. This corpus represents a practically relevant domain for efforts to simplify children's reading materials.

The rule-based baseline relies on simple linguistic features such as word length, number of syllables, and the presence of common Indonesian affixes. After parameter tuning on the validation set, this baseline achieves an accuracy of 0.9133 and an F1 of 0.8607 for the CWI class on the test set. These results show that a rule-based approach can provide a reasonable starting point while offering high interpretability regarding which lexical patterns are considered complex.

The second approach uses IndoBERT fine-tuned as a token classifier. Evaluated on the same test tokens, the model attains an accuracy of 0.9983 and an F1 of 0.9972 for the CWI class, with only 39 errors (combined false positives and false negatives). This comparison demonstrates that IndoBERT significantly outperforms the rule-based baseline across all evaluation metrics and is therefore suitable as a strong baseline for CWI in Indonesian children's texts.

Overall, the main contributions of this work are: an explicit formulation of the CWI task for Indonesian children's texts, together with a binary token-level annotation scheme, the use of an annotated CWI corpus as a benchmark for evaluating CWI models in this domain; and the establishment of both a linguistic rule-based baseline and a strong IndoBERT-based baseline, compared systematically on the same dataset. Importantly, this research advances Informatics/Computer Science by strengthening educational NLP resources for a low-resource language and enabling more reliable automatic support for children's reading comprehension.

In future work, the IndoBERT-based CWI module can be integrated into end-to-end text simplification pipelines for children, for example by adding modules to select simpler synonyms or to generate interactive glossaries. The corpus can also be enriched with additional information, such as grade level and reader profiles, enabling readability analyses to be conducted in a more comprehensive and fine-grained manner. Furthermore, exploring other pre-trained models (e.g., larger IndoBERT variants or multilingual models) and applying knowledge distillation techniques may yield lighter CWI models that are easier to integrate into real-world educational applications and digital learning platforms

CONFLICT OF INTEREST

The author declares that the entire research, analysis, and manuscript preparation process was conducted without any conflict of interest that could affect the academic and scientific integrity of this article.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the financial support for this research provided by Universitas Lancang Kuning.

REFERENCES

[1] S. A. Bahrainian, J. Dou, and C. Eickhoff, "Text Simplification via Adaptive Teaching," 2024. https://doi.org/ 10.18653/v1/2024.findings-acl.392

[2] M. Anschütz, J. Oehms, T. Wimmer, B. Jezierski, and G. Groh, "Language Models for German Text Simplification: Overcoming Parallel Data Scarcity through Style-specific Pre-training," *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 1147–1158, 2023, https://doi.org/10.18653/v1/2023.findings-acl.74.

[3] S. Alissa and M. Wald, "Text Simplification Using Transformer and BERT," Computers, Materials and Continua,vol.75,no.2, pp. 3479–3495, 2023, https://doi.org/10.32604/cmc.2023.033647.

[4] M. Shardlow, "Predicting lexical complexity in English texts: the Complex 2.0 dataset," *Lang Resour Eval*, vol. 56, no. 4, pp. 1153–1194, 2022, https://doi.org/10.1007/s10579-022-09588-2.

[5] A. Aziz, "CSECU-DSG at SemEval-2021 Task 1: Fusion of Transformer Models for Lexical Complexity Prediction," 2021. https://doi.org/10.18653/v1/2021.semeval-1.80

[6] G. E. Zaharia, "Domain Adaptation in Multilingual and Multi-Domain Monolingual Settings for Complex Word Identification," 2022. https://doi.org/10.18653/v1/2022.acl-long.6

[7] J. Ortiz-Zambrano, "SINAI at SemEval-2021 Task 1: Complex word identification using Word-level features," 2021. https://doi.org/ 10.18653/v1/2021.semeval-1.11

[8] E. Zotova, "Vicomtech at alexs 2020: Unsupervised complex word identification based on domain frequency," 2020. https://ceur-ws.org/Vol-2664/alesx_paper1.pdf

[9] J. R. Irina Rets, "To simplify or not? Facilitating English L2 users' comprehension and processing of open educational resources in English using text simplification," 2020, *Wiley*. https://doi.org/ 10.1111/jcal.12517/v1/decision1.

[10] J. Qiang, Y. Li, Y. Zhu, Y. Yuan, Y. Shi, and X. Wu, "LSBert: Lexical Simplification Based on BERT," *IEEE/ACM Trans Audio Speech Lang Process*, vol. 29, pp. 3064–3076, 2021, https://doi.org/ 10.1109/TASLP.2021.3111589.

[11] M. S. Wibowo, "Lexical and Syntactic Simplification for Indonesian Text," 2019. https://doi.org/10.1109/ISRITI48646.2019.9034582.

[12] P. Śliwiak and S. A. A. Shah, "Text-to-text generative approach for enhanced complex word identification," *Neurocomputing*, vol. 610, Dec. 2024, https://doi.org/10.1016/j.neucom.2024.128501

[13] R. Azvan-Alexandru, S. Adu, D.-G. Ion, D.-C. Cercel, F. Pop, and M.-C. Cercel, "Investigating Large Language Models for Complex Word Identification in Multilingual and Multidomain Setups," 2014. https://doi.org/10.18653/v1/2024.emnlp-main.933

[14] A. Kelious, M. Constant, and C. Coeur, "Complex Word Identification: a Comparative Study Between ChatGPT and a Dedicated Model for this Task," 2024. https://aclanthology.org/2024.lrec-main.323.pdf

[15] S. Gooding and E. Kochmar, "Complex Word Identification as a Sequence Labelling Task," Association for Computational Linguistics, 2019. https://doi.org/10.18653/v1/P19-1109

[16] K. North, M. Zampieri, and M. Shardlow, "Lexical Complexity Prediction: An Overview," Sep. 30, 2023, *Association for Computing Machinery*. https://doi.org/10.1145/3557885

[17] Z. Yuan, G. Tyen, and D. Strohmaier, "Cambridge at SemEval-2021 Task 1: An Ensemble of Feature-Based and Neural Models for Lexical Complexity Prediction," 2021. https://doi.org/10.18653/v1/2021.semeval-1.74

[18] T. Han, X. Zhang, Y. Bi, M. Mulvenna, and D. Yang, "From Complex Word Identification to Substitution: Instruction-Tuned Language Models for Lexical Simplification," 2025. https://doi.org/10.18653/v1/2025.starsem-1.4

[19] A. Tucker, "An investigation of complex word identification (CWI) systems for English," 2023. https://home.cltl.labs.vu.nl/static/data

[20] Anderies, R. Rahutomo, and B. Pardamean, "Finetunning IndoBERT to Understand Indonesian Stock Trader Slang Language," *Proceedings of 2021 1st International Conference on Computer Science and Artificial Intelligence, ICCSAI 2021*, no. October, pp. 42–46, 2021, https://doi.org/10.1109/ICCSAI53272.2021.9609746.

[21] S. M. Isa, "Indobert For Indonesian Fake News Detection," *ICIC Express Letters*, vol. 16, no. 3, pp. 289–297, 2022, https://doi.org/10.24507/icicel.16.03.289.

[22] E. Fernandez, "Improving IndoBERT for Sentiment Analysis on Indonesian Stock Trader Slang Language," 2022. https://doi.org/10.1109/IoTaIS56727.2022.9975975.

[23] F. Baharuddin and M. F. Naufal, "Fine-Tuning IndoBERT for Indonesian Exam Question Classification Based on Bloom's Taxonomy," *Journal of Information Systems Engineering and Business Intelligence*, vol. 9, no. 2, pp. 253–263, Oct. 2023, https://doi.org/10.20473/jisebi.9.2.253-263.

[24] F. Koto, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," 2020. https://doi.org/10.18653/v1/2020.coling-main.66

[25] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," pp. 843–857, 2020, http://arxiv.org/abs/2009.05387

[26] S. Cahyawijaya *et al.*, "IndoNLG: Benchmark and Resources for Evaluating Indonesian Natural Language Generation, 2021" https://doi.org/10.18653/v1/2021.emnlp-main.

[27] W. Zhou, T. Ge, K. Xu, F. Wei, and M. Zhou, "BERT-based Lexical Substitution," 2019, *Association for Computational Linguistics*. https://doi.org/10.18653/v1/p19-1328.

[28] N. El Mamoun, A. El Mahdaouy, A. El Mekki, K. Essefar, and I. Berrada, "CS-UM6P at SemEval-2021 Task 1: A Deep Learning Model-based Pre-trained Transformer Encoder for Lexical Complexity," *SemEval 2021 - 15th International Workshop on Semantic Evaluation, Proceedings of the Workshop*, pp. 585–589, 2021, https://doi.org/10.18653/v1/2021.semeval-1.73.

[29] J. A. Ortiz-Zambrano, C. Espin-Riofrio, and A. Montejo-Ráez, "Combining Transformer Embeddings with Linguistic Features for Complex Word Identification," *Electronics (Switzerland)*, vol. 12, no. 1, pp. 1–10, 2023, https://doi.org/10.3390/electronics12010120.

[30] G. E. Zaharia, D. C. Cercel, and M. Dascalu, "Cross-Lingual Transfer Learning for Complex Word Identification," *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, vol. 2020-Novem, pp. 384–390, 2020, https://doi.org/10.1109/ICTAI50040.2020.00067.

[31] R. Flynn and M. Shardlow, "Manchester Metropolitan at SemEval-2021 Task 1: Convolutional Networks for Complex Word Identification," *SemEval 2021 - 15th International Workshop on Semantic Evaluation, Proceedings of the Workshop*, pp. 603–608, 2021, https://doi.org/10.18653/v1/2021.semeval-1.76.

[32] A. M. Butnaru, "UnibucKernel: A kernel-based learning method for complex word identification," 2018. https://doi.org/10.18653/v1/W18-0519

[33] D. de Hertog, "Deep learning architecture for Complex Word Identification," 2018. https://doi.org/ 10.18653/v1/W18-0539