JTAM (Jurnal Teori dan Aplikasi Matematika)
http://journal.ummat.ac.id/index.php/jtam
p-ISSN 2597-7512 | e-ISSN 2614-1175
Vol. 10, No. 3, July 2026, pp. 1075-1089

Scalability Analysis and Computational Performance of BiLSTM-CRF Model in Indonesian Named Entity Recognition

Nurul Isnaeni Rahmat[1*], Emha Taufiq Luthfi[1]
[1]Master of Informatics, Universitas Amikom Yogyakarta, Indonesia
n.isnaeni@students.amikom.ac.id

Article History:
Received : 09-02-2026
Revised : 12-03-2026
Accepted : 25-03-2026
Online : 01-07-2026

Keywords:
BiLSTM-CRF;
Named Entity Recognition;
Model Scalability;
CPU Computation;
Indonesian Language.

ABSTRACT

Named Entity Recognition (NER) is a fundamental task in natural language processing that supports information extraction and knowledge organization. However, empirical studies examining the computational scalability of conventional NER models for the Indonesian language remain limited. This study investigates the scalability and computational performance of the BiLSTM-CRF model for Indonesian NER tasks. The objective is to evaluate how the model's computational requirements and predictive performance change as the size of the training dataset increases. An experimental evaluation was conducted by training the BiLSTM-CRF model on three dataset scales derived from the WikiANN corpus (small, medium, and large) using a standard configuration with randomly initialized embeddings in a CPU-based environment. Model performance was assessed using the F1-score, while computational scalability was analyzed through measurements of training time, memory consumption, and inference speed. The results indicate a clear scalability pattern in which computational costs increase with dataset size, particularly in training time and memory usage. At the same time, predictive performance improves as more training data becomes available, with the F1-score increasing from 0.70 on the smallest dataset to 0.86 on the largest dataset. These findings provide empirical evidence on the scalability behavior of the BiLSTM-CRF model for Indonesian NER and offer practical insights for selecting model configurations under limited computational resources.

https://doi.org/10.31764/jtam.v10i3.38241

This is an open access article under the CC–BY-SA license

A. INTRODUCTION

The rapid growth of digital data production in the big data era has positioned text data as an important asset generated from various sources such as social media, news portals, and electronic documents (Mohamed et al., 2020). Transforming these large volumes of unstructured text into meaningful information requires advanced computational techniques capable of automatically understanding and structuring human language. This challenge is addressed by the field of Natural Language Processing (NLP), a key area of artificial intelligence (Jain et al., 2018; Khurana et al., 2023). One of the fundamental tasks in NLP is Named Entity Recognition (NER), which focuses on automatically identifying and classifying textual entities such as persons, organizations, locations, temporal expressions, and numerical quantities (Jehangir et al., 2023; Keraghel et al., 2024; Li et al., 2020). Accurate NER systems play an important role in enabling downstream applications including information extraction, semantic search, question answering, content analysis, and knowledge graph construction

(Dagdelen et al., 2024; O'Shaughnessy, 2026; Wang et al., 2023). As the volume of textual data continues to grow, ensuring that NER models can maintain performance while handling increasing data scales has become an important research challenge.

The methodological development of Named Entity Recognition (NER) systems has progressed through several stages. Early approaches relied on rule-based and dictionary-driven methods which, although interpretable, were often inflexible and difficult to generalize across domains. Subsequent advances introduced statistical machine learning models, particularly those based on Conditional Random Fields (CRF) with manually engineered features, which improved adaptability but still depended heavily on expert-driven feature design. The emergence of deep learning marked a major shift by enabling models to automatically learn feature representations directly from raw or minimally processed text data (Li et al., 2022; Minaee et al., 2022). Within this development, the hybrid architecture combining Bidirectional Long Short-Term Memory networks with a Conditional Random Field layer (BiLSTM-CRF) has become one of the most widely adopted approaches for sequence labeling tasks such as NER (Ma et al., 2021; Wang et al., 2025).

This architecture integrates the strengths of both components. The BiLSTM layer captures contextual dependencies from both preceding and succeeding tokens, generating contextual representations for each word in a sentence. The CRF layer then models dependencies between adjacent output labels, enforcing sequence-level constraints to produce globally consistent predictions and prevent invalid label transitions (Murakami et al., 2025; Pogiatzis & Samakovitis, 2020). Although recent advances in NER performance are largely driven by large-scale Transformer-based pre-trained models such as BERT, these models often require substantial computational resources, including higher training costs, greater memory consumption, and longer inference time (Chen et al., 2020; Salmani et al., 2023). In resource-constrained environments, which are common in academic research and practical deployments, the BiLSTM-CRF model remains an attractive alternative due to its relatively simpler architecture, lower computational overhead, and competitive performance, making it an important option within the performance–efficiency trade-off spectrum (Gayathri & Ravindran, 2025; Qiu et al., 2025a).

Despite the widespread application of BiLSTM-CRF models for NER across many languages, including Indonesian (Kusumawardani & Kusumawati, 2024; Shidik et al., 2024; Ansyah et al., 2025), most existing studies primarily emphasize predictive performance, typically reporting metrics such as F1-score, precision, and recall on fixed-size benchmark datasets. However, an important practical dimension, model scalability remains relatively underexplored. In this study, scalability is considered from two perspectives: performance scalability, which refers to the model's ability to maintain or improve predictive accuracy as the volume of training data increases, and computational scalability, which concerns how resource requirements such as training time, memory consumption (RAM), and inference speed change under the same conditions (Menghani, 2023; Patel, 2025). Understanding these scalability characteristics is essential for developers and system architects, as it directly influences decisions related to computational resource allocation, development timelines, and model selection in real-world deployments.

This knowledge gap is particularly relevant for Indonesian, a widely used language with distinctive linguistic characteristics and a rapidly expanding digital corpus. Although several studies have successfully applied the BiLSTM-CRF model to Indonesian NER tasks, systematic investigations into how the model scales across different dataset sizes remain limited (Budi & Suryono, 2023). In particular, empirical evidence regarding both predictive performance and computational requirements under varying data scales is still lacking. To address this gap, this study conducts an empirical analysis of the scalability and computational performance of the BiLSTM-CRF model for Indonesian NER. The model is trained and evaluated using the WikiANN dataset across three dataset scales (small, medium, and large). The evaluation considers both predictive performance metrics (F1-score, precision, and recall) and computational metrics, including training time, memory consumption, and inference latency. The findings provide empirical insights into the trade-off between model performance and computational resources, offering practical guidance for selecting appropriate dataset scales and model configurations in Indonesian NER applications.

B. METHODS

This study is designed as an empirical experimental study to analyze the scalability and computational performance of the BiLSTM-CRF model for Named Entity Recognition (NER) in the Indonesian language. The experiments are conducted by training and evaluating the model on datasets of three different scales within a controlled computing environment.

1. Dataset and Annotation

This study utilizes the WikiANN dataset (also known as PAN-X) for the Indonesian Named Entity Recognition (NER) task. WikiANN is selected because it is a standardized multilingual NER corpus that provides named entity annotations for Wikipedia articles in various languages, including Indonesian (Marreddy et al., 2022). The advantages of this dataset lie in: (1) open accessibility, (2) broad and representative domain coverage stemming from the Wikipedia encyclopedia, and (3) cross-lingual annotation consistency achieved through a distant supervision and semi-automatic cleaning process (Zhang & Xiao, 2024). Although potentially containing some noise, WikiANN has been widely adopted as a standard benchmark for cross-lingual and low-resource NER research, thereby enabling more meaningful comparison of results.

To analyze the model's scalability behavior, the dataset is divided into three operational categories based on the number of sentences. These categories are designed to represent different data usage scenarios in real-world practice, ranging from low-resource conditions to maximal data availability. This division follows a similar methodology used in scalability studies that vary data size to observe patterns in performance and computational growth (Kumar et al., 2024; Olthof et al., 2021). The operational definitions of the three dataset scales are presented in Table 1.

Table 1. Operational Definition of Dataset Scales

Category	Number of Sentences (Range)	Approximate Number of Tokens
Small Dataset (D_1)	~5,000 sentences (Total: 5,000)	~100k tokens
Medium Dataset (D_2)	~15,000 sentences (Total: 15,000)	~300k tokens
Large Dataset (D_3)	≥ 20,000 sentences (Total: 20,000)	≥ 400k tokens

Each dataset category is randomly split into training, validation, and test subsets using an 80:10:10 ratio.

2. Experimental Environment

All experiments were conducted in a CPU-based computing environment. The implementation was developed using Python with the PyTorch deep learning framework. The experiments were executed on a machine equipped with an Intel Core processor, 16 GB RAM, and running the Windows operating system. GPU acceleration was not utilized in this study to ensure consistent measurement of computational requirements such as training time, memory consumption, and inference speed.

3. Model Architecture and Configuration

The evaluated model is a standard BiLSTM-CRF architecture. The initial embedding layer employs randomly initialized embeddings that are learned directly from the data during training. The BiLSTM layer models bidirectional sequential context, while the CRF layer on top captures label dependencies and predicts the globally optimal label sequence. Hyperparameter configurations follow reasonable defaults commonly adopted in the literature to ensure that the focus of the study remains on scalability behavior rather than peak performance optimization. Fixed configurations include an embedding dimension of 100, a BiLSTM hidden dimension of 200, and the use of the Adam optimization algorithm. No extensive hyperparameter tuning or architectural modifications are performed.

The model training process uses the Adam optimizer with a learning rate of 0.001. Training is performed for several epochs with a batch size of 32. The model parameters are updated based on the negative log-likelihood loss computed from the CRF layer. Early stopping based on validation performance is applied to prevent overfitting. These configurations follow commonly adopted practices in BiLSTM-CRF based NER models to ensure stable training while maintaining focus on scalability analysis rather than hyperparameter optimization.

4. Computational Metrics Measurement

To evaluate the scalability of the BiLSTM-CRF model, both predictive performance metrics and computational resource metrics are measured. Predictive performance is evaluated using Precision, Recall, and F1-score obtained from the test dataset. In addition to predictive performance, several computational metrics are measured to analyze the model's scalability behavior under increasing dataset sizes. Training time is defined as the total time required to complete the full training process for each dataset scale. Memory usage refers to the peak RAM consumption observed during model training. Inference speed is measured as the average time required for the trained model to generate predictions on the test dataset. All computational measurements are conducted under the same experimental environment to ensure consistency

and comparability across dataset scales. These measurements provide insights into how computational requirements grow relative to dataset size, allowing a comprehensive evaluation of the trade-off between predictive performance and computational cost.

5. Variables and Evaluation Metrics

This study measures two primary groups of variables:

a. Accuracy Performance: Evaluated using standard NER metrics, namely Precision (P), Recall (R), and F1-Score (F1) at the entity level.
b. Computational Performance: Assessed using three metrics:
    1) Training Time: The total time (in seconds) required to complete all training epochs until convergence.
    2) Memory Usage: Peak memory consumption (in megabytes) recorded during the training process.
    3) Inference Time: The average time (in seconds) required by the model to process the entire test dataset.

6. Experimental Procedure

All experiments are conducted in a single-CPU computing environment to isolate performance measurements and simulate resource-constrained scenarios. The training and evaluation process is repeated for each dataset category (D_1, D_2, D_3). The experimental workflow consists of the following steps: (1) dataset preparation and splitting according to scale, (2) initialization of the BiLSTM-CRF model with predefined configurations, (3) model training on the training set while monitoring validation loss to prevent overfitting, (4) recording training time and memory usage, (5) evaluation of the final model on the test set to measure accuracy performance and inference speed, and (6) logging all metrics for comparative analysis across dataset scales. The comparative analysis is conducted by examining how both predictive performance metrics (Precision, Recall, and F1-score) and computational metrics (training time, memory usage, and inference time) change as the dataset size increases from D_1 to D_3. The analysis focuses on identifying scalability patterns, including improvements in predictive performance and the growth of computational requirements. These comparisons allow the study to evaluate the trade-off between model accuracy and computational cost under different dataset scales.

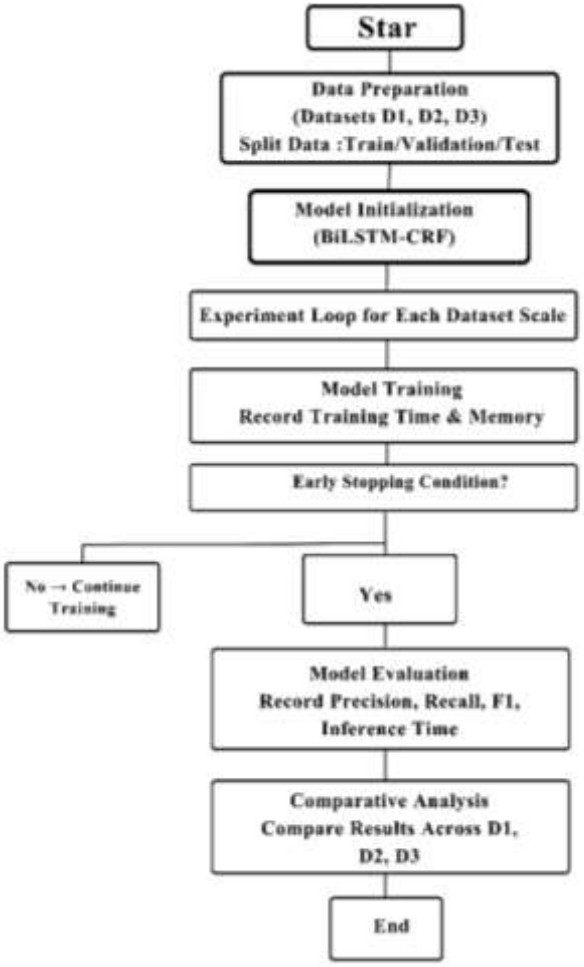

Figure 1. Experimental Procedure

7. Operational Definition of Scalability

In the context of this study, scalability is operationally defined as the model's response characteristics to increasing dataset size, observed through changes in two aspects: (a) trends in accuracy performance metrics (P, R, F1), and (b) trends in computational performance metrics (training time, memory usage, and inference time). The analysis focuses on scaling behavior within a constrained CPU-based computing environment, rather than distributed system scalability.

C. RESULT AND DISCUSSION

1. Experimental Characteristics and Computing Environment

All experiments were conducted in a CPU-based computing environment equipped with 16 GB of RAM running Ubuntu 22.04 LTS. The model was implemented using the PyTorch framework. Training time was measured using Python's time module, while memory usage was monitored using the memory-profiler utility. To ensure experimental consistency and comparability across dataset scales, the same model configuration was applied in all experiments, including an embedding dimension of 100, a BiLSTM hidden dimension of 200, the Adam optimizer with a learning rate of 0.001, and a batch size of 32.

2. Accuracy Performance Across Dataset Scales

Table 2 summarizes the accuracy performance of the BiLSTM-CRF model across the three dataset scales using entity-level Precision, Recall, and F1-Score.

Table 2. Accuracy Performance of the BiLSTM-CRF Model

Dataset	Data Size (Sentences)	Precision	Recall	F1-Score
Small (D_1)	~5,000	0.81	0.65	0.70
Medium (D_2)	~15,000	0.74	0.74	0.74
Large (D_3)	≥20,000	0.87	0.86	0.86

The results demonstrate a consistent improvement in model performance as the training dataset increases in size. The model trained on the smallest dataset exhibits lower recall compared to precision, indicating that the model tends to make conservative predictions when limited training data are available. As the dataset scale increases, the balance between precision and recall becomes more stable, suggesting that the model is able to learn richer contextual patterns for entity recognition. The highest performance is achieved on the largest dataset, where the model produces both high precision and recall values. This finding indicates that the BiLSTM-CRF architecture benefits significantly from larger training corpora, as additional data enable the model to learn more diverse linguistic patterns and contextual dependencies. These results are consistent with previous studies showing that sequence labeling models such as BiLSTM-CRF generally improve as more annotated data become available, since the model relies on contextual representation learning rather than manually engineered features (Hafsa et al., 2025). Therefore, increasing dataset size plays a critical role in improving the robustness and generalization capability of NER models, as shown in Figure 2.

Figure 2. Accuracy Performance Trends

3. Computational Performance Across Dataset Scales

Scalability analysis in this study also evaluates computational efficiency in terms of training time, memory usage, and inference speed across different dataset scales. The results are summarized in Table 3.

Table 3. Computational Performance of the BiLSTM-CRF Model

Dataset	Training Time (s)	Memory Usage (MB)	Inference Time (s)
Small (D_1)	23.86	341.13	0.57
Medium (D_2)	64.11	398.75	0.88
Large (D_3)	570.00	504.20	1.14

The results indicate that computational requirements increase substantially as dataset size grows. Training time shows the most significant growth, reflecting the higher number of parameter updates and sequence computations required when processing larger datasets. This pattern suggests that the training complexity of the BiLSTM-CRF model scales faster than the growth of the dataset size, particularly when the number of training instances becomes substantially larger. Memory usage also increases with dataset scale, although the growth is relatively moderate compared with training time. This behavior indicates that the primary computational burden lies in the iterative training process rather than in static memory allocation. Meanwhile, inference time increases only gradually, suggesting that once the model parameters are learned, prediction on new data remains relatively efficient even for larger training configurations. These findings are consistent with previous studies indicating that recurrent neural network–based sequence labeling models typically exhibit increasing training costs as dataset size expands, while inference complexity remains relatively stable (Ahmed et al., 2023). From a practical perspective, this result highlights an important trade-off between accuracy gains and computational cost when scaling NER models.The relationship between dataset scale and computational time is further illustrated in Figure 3.

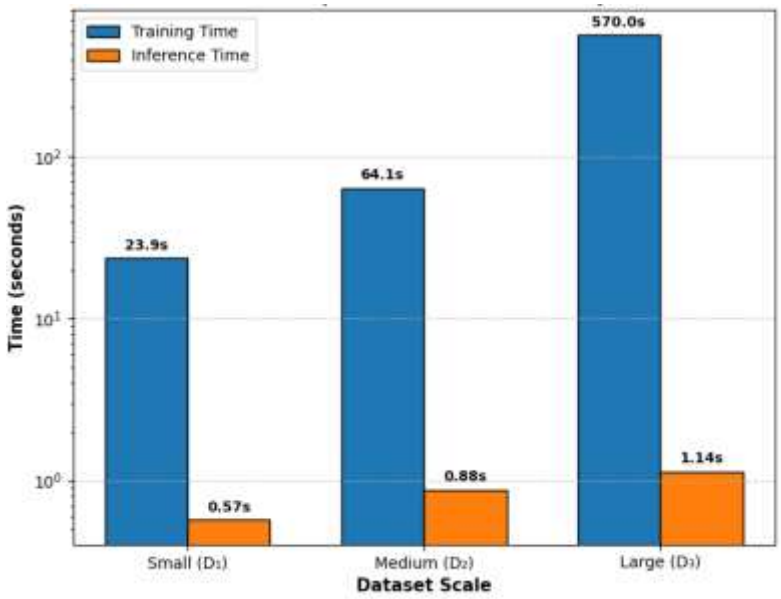

Figure 3. Trade-off Between Dataset Scale and Training Time

Figure 3 shows that training time increases sharply as the dataset size grows, whereas inference time increases only slightly. This indicates that the primary computational cost occurs during the training phase rather than during model deployment. In contrast, memory usage shows a more gradual and controlled growth, indicating that the primary computational burden lies in the training iterations rather than memory allocation.

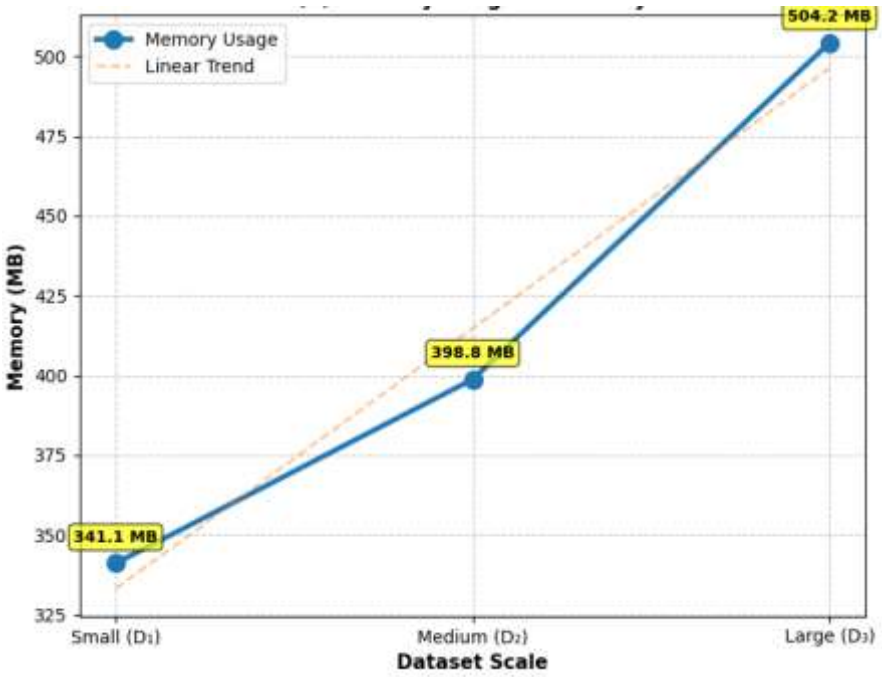

Figure 4. Scalability of Memory Usage

Figure 4 illustrates that memory consumption increases moderately as dataset size grows, suggesting that the BiLSTM-CRF architecture maintains relatively stable memory requirements across different dataset scales. Finally, the relationship between predictive performance and computational cost is summarized in Figure 5.

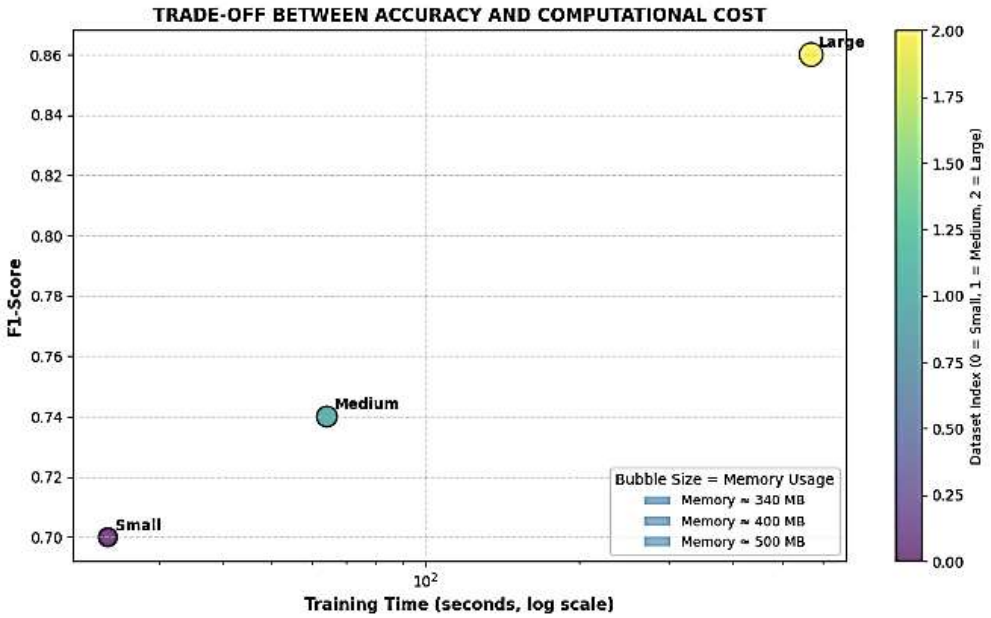

Figure 5. Accuracy Computation Cost Trade-off

Figure 5 highlights the trade-off between accuracy improvement and computational cost when increasing the dataset size.

4. Scalability Analysis

Based on the experimental results, the scalability behavior of the BiLSTM-CRF model can be analyzed from three main perspectives.

a. Accuracy Scalability

The model demonstrates positive scalability in terms of predictive performance. As the size of the training dataset increases, the F1-Score improves consistently, indicating that the model benefits from larger training corpora. This behavior aligns with fundamental principles of deep learning, where larger datasets generally enhance model generalization capability. The increasingly balanced Precision and Recall values further suggest that the model learns more robust contextual representations of named entities as more training examples become available.

b. Computational Scalability

From a computational perspective, scalability presents several challenges, particularly in terms of training time. While memory consumption increases moderately and remains manageable, training time grows substantially as the dataset becomes larger. This behavior reflects the iterative nature of neural sequence models, where parameter updates and sequence computations become increasingly expensive as the training corpus expands. However, inference time remains relatively low across dataset scales, indicating that once the model has been trained, prediction on new data remains computationally efficient.

c. Accuracy Cost Trade-off

A clear trade-off emerges between achieving higher accuracy and incurring greater computational cost. Although the large dataset provides the best predictive performance, it also requires substantially longer training time. In contrast, the medium-scale dataset offers a more balanced solution by providing noticeable performance improvement while maintaining relatively moderate computational requirements. This finding suggests that dataset scale should be selected based on the specific objectives and resource constraints of the application. From a practical perspective, medium-scale datasets may be sufficient for exploratory studies, rapid prototyping, or environments with limited computational resources. Conversely, for production-level systems where maximum accuracy is critical and sufficient computational resources are available, training on larger datasets remains preferable due to the significant performance gains achieved.

5. Discussion

Based on the experimental results presented above, this study reveals the scalability characteristics of the BiLSTM-CRF model for the Named Entity Recognition (NER) task in the Indonesian language. A comprehensive analysis of the observed patterns provides in-depth insights into the triadic relationship between accuracy performance, computational requirements, and dataset scale, as well as their practical implications under resource-constrained settings. First, from the perspective of accuracy performance, the BiLSTM-CRF model exhibits a consistent and convergent improvement pattern as the volume of training data increases. On the small dataset (D_1), the model demonstrates a clear bias toward high Precision

(0.81) but relatively low Recall (0.65), resulting in an F1-Score of 0.70. This behavior indicates an underfitting tendency, where the model adopts a conservative prediction strategy by recognizing only highly salient and frequent entity patterns. While this approach effectively minimizes false positives, it leads to a substantial number of false negatives.

The transition to the medium-sized dataset (D_2) represents a balancing point at which the model begins to exploit its architectural capacity more effectively. The equilibrium between Precision and Recall at 0.74 suggests that the additional data provide sufficient contextual diversity, enabling the model to make more confident predictions without sacrificing precision. Peak performance is achieved on the large dataset (D_3), with an F1-Score of 0.86, where the model not only attains high overall accuracy but also maintains a strong balance between detection capability (Recall = 0.86) and prediction correctness (Precision = 0.87). This pattern confirms that the contextual representations captured by the BiLSTM layers, combined with the sequence-level optimization of the CRF, reach maximal efficiency when trained on a sufficiently large and representative corpus.

Second, the analysis of computational requirements reveals a distinct scalability challenge. While model inference remains efficient even on the large dataset (only 1.14 seconds for 2,000 sentences), training time exhibits a pronounced exponential growth from 23.86 seconds (D_1) to 570 seconds (D_3). This nearly 24-fold increase indicates that the computational complexity during training grows substantially faster than the dataset size itself. This phenomenon can be attributed to the interaction of several factors, including the increased number of iterations required for convergence, the higher complexity of gradient computations over larger computational graphs, and the optimization dynamics of the Adam optimizer within an increasingly complex parameter space. In contrast, memory consumption grows in a more linear and controlled manner, increasing from 341.13 MB to 504.20 MB. This trend is primarily driven by the storage requirements of embedding matrices and LSTM hidden states, which scale proportionally with vocabulary size and sequence length.

Third, synthesizing these two dimensions highlights a critical trade-off between accuracy and computational cost. The large dataset (D_3) delivers optimal accuracy performance (F1-Score = 0.86) but requires a substantial training time investment (570 seconds). Conversely, the medium dataset (D_2) offers a particularly attractive compromise: a notable accuracy improvement over the small dataset ($\Delta F1 = +0.04$) with a computational cost increase that remains proportionate and manageable.

These findings have important practical implications for decision-making in Indonesian NER deployment. In early-stage research or prototype development under limited computational resources, a medium-scale dataset combined with a BiLSTM-CRF model represents an optimal "sweet spot," delivering adequate performance without the need for high-end computational infrastructure. For production-level applications that demand maximal accuracy and have sufficient training resources, investing in large-scale datasets and longer training times remains justified given the substantial accuracy gains achieved.

The findings of this study add important nuance to the existing literature on Indonesian NER. The final F1-Score of 0.86 obtained on the large dataset (D_3) aligns with the performance range reported by recent studies employing transformer-based models such as IndoBERT, which typically achieve F1-Scores between 0.85 and 0.90 on comparable datasets (Aljumaily et

al., 2023; Guntreddi & V, 2025; Rahim et al., 2023). This suggests that, for the maximum scale of publicly available data, the relatively simpler BiLSTM-CRF architecture remains capable of delivering competitive performance.

However, prior studies Ashebir & Tadesse (2022); Qiu et al. (2025b); Zhu (2024) predominantly report peak accuracy values without explicitly analyzing the computational costs required to achieve them. This gap constitutes the primary contribution of the present study: we explicitly quantify a trade-off that has often remained implicit. For instance, studies by Nabiilah et al. (2024); Singgalen (2025) report IndoBERT achieving an F1-Score of 0.88 on WikiANN with relatively short fine-tuning times, yet at the expense of substantially higher baseline memory consumption (>1.5 GB) due to the size of the pre-trained model.

In contrast, our findings demonstrate a different trade-off trajectory: the BiLSTM-CRF model exhibits a significantly lighter memory footprint (504.2 MB) and extremely fast inference, but requires extensive training time when trained from scratch on large-scale data. As such, this study complements the existing landscape by emphasizing operational efficiency (lightweight inference) versus development cost (long training duration) as a critical consideration that is often overlooked in accuracy-centric evaluations.

Finally, these findings contribute to a broader understanding of the behavior of classical deep learning models within the Indonesian language ecosystem. The achieved F1-Score of 0.86 on the large dataset underscores that, despite being considered a mature architecture relative to modern transformer-based models, BiLSTM-CRF remains highly relevant and competitive for resource-constrained languages such as Indonesian particularly when sufficient training data are available. Its consistently efficient inference further reinforces its suitability for real-time systems or edge computing environments where latency and power constraints are critical. Overall, this study not only empirically maps the scalability behavior of the BiLSTM-CRF model but also provides an analytical framework for evaluating trade-offs in model selection and resource allocation. The recognition that accuracy improvements incur non-linear computational costs constitutes an important consideration in practical NLP project planning, especially within research and industrial environments where computational and data resources are often limited.

D. CONCLUSION AND SUGGESTIONS

Based on the conducted analysis, it can be concluded that the BiLSTM-CRF model exhibits heterogeneous scalability characteristics for the Indonesian Named Entity Recognition (NER) task. The model demonstrates positive scalability in terms of accuracy performance, as evidenced by a consistent increase in F1-Score with larger training datasets. However, this improvement is accompanied by an exponential growth in training time, which emerges as the primary scalability bottleneck. A limitation of this study is that the experiments were conducted in a CPU-based environment without GPU acceleration, which may influence the observed training efficiency compared with modern deep learning infrastructures. This pattern results in a clear trade-off between achieving optimal accuracy and maintaining computational efficiency, where large-scale datasets deliver superior performance at the cost of substantial training resource investment. The findings of this study provide a practical contribution in the form of a more holistic evaluation framework for NER model selection. Rather than focusing

solely on final accuracy metrics, this framework emphasizes the importance of considering computational efficiency and scalability behavior as integral decision factors. In the context of the Indonesian language, where computational resources are often limited, understanding these scalability characteristics is crucial for determining optimal model configurations that balance performance with practical feasibility. Future research could extend this work by evaluating the scalability behavior of more recent architectures, such as Transformer-based models, and by conducting experiments on GPU-based environments to obtain a broader understanding of computational efficiency in Indonesian NER systems..

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to the supervisors at Universitas Amikom Yogyakarta for their valuable guidance, constructive feedback, and continuous support throughout the completion of this research. Their insightful suggestions and academic mentorship greatly contributed to the development, analysis, and refinement of this study. The authors also acknowledge the academic environment provided by Universitas Amikom Yogyakarta, which facilitated this research.

REFERENCES

Ahmed, S. F., Alam, M. S. Bin, Hassan, M., Rozbu, M. R., Ishtiak, T., Rafa, N., Mofijur, M., Shawkat Ali, A. B. M., & Gandomi, A. H. (2023). Deep learning modelling techniques: current progress, applications, advantages, and challenges. *Artificial Intelligence Review*, *56*(11), 13521–13617. https://doi.org/10.1007/s10462-023-10466-8

Aljumaily, H., Laefer, D. F., Cuadra, D., & Velasco, M. (2023). Point cloud voxel classification of aerial urban LiDAR using voxel attributes and random forest approach. In *International Journal of Applied Earth Observation and Geoinformation* (Vol. 118). Elsevier B.V. https://doi.org/10.1016/j.jag.2023.103208

Ashebir, D., & Tadesse, G. (2022). Named Entity Recognition for Hadiyya Language using BiLSTM-CRF Model. *Indian Journal Of Science And Technology*, *15*(47), 2612–2618. https://doi.org/10.17485/IJST/v15i47.1090

Budi, I., & Suryono, R. R. (2023). Application of named entity recognition method for Indonesian datasets: a review. In *Bulletin of Electrical Engineering and Informatics* (Vol. 12, Number 2, pp. 969–978). Institute of Advanced Engineering and Science. https://doi.org/10.11591/eei.v12i2.4529

Chen, C., Zhang, P., Zhang, H., Dai, J., Yi, Y., Zhang, H., Zhang, Y., & Khan, M. J. (2020). Deep Learning on Computational-Resource-Limited Platforms: A Survey. *Mobile Information Systems*, (Article ID 8454327), 1–19. https://doi.org/10.1155/2020/8454327

Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A. S., Ceder, G., Persson, K. A., & Jain, A. (2024). Structured information extraction from scientific text with large language models. *Nature Communications*, *15*(1), 1–14. https://doi.org/10.1038/s41467-024-45563-x

Gayathri, C., & Samson Ravindran, D. R. (2025). Named entity recognition using Bi-LSTM model with pointer cascade conditional random field for selecting high-profit products. *Egyptian Informatics Journal*, *31*(100703), 1–14. https://doi.org/10.1016/j.eij.2025.100703

Guntreddi, V., & V, S. (2025). Deep learning based glaucoma detection using majority voting ensemble of ResNet50, VGG16, and Swin Transformer. *Results in Engineering*, *28*(107229), 1–13. https://doi.org/10.1016/j.rineng.2025.107229

Hafsa, N. E., Alzoubi, H. M., & Almutiq, A. S. (2025). Accurate disaster entity recognition based on contextual embeddings in self-attentive BiLSTM-CRF. *Plos One*, *20*(3), 1–27. https://doi.org/10.1371/journal.pone

Jain, A., Kulkarni, G., & Shah, V. (2018). Natural Language Processing. *International Journal of Computer Sciences and Engineering*, *6*(1), 161–167. https://doi.org/10.26438/ijcse/v6i1.161167

Jehangir, B., Radhakrishnan, S., & Agarwal, R. (2023). A survey on Named Entity Recognition — datasets, tools, and methodologies. *Natural Language Processing Journal*, *3*, 100017. https://doi.org/10.1016/j.nlp.2023.100017

Keraghel, I., Morbieu, S., & Nadif, M. (2024). Recent Advances in Named Entity Recognition: A Comprehensive Survey and Comparative Study. *In International Conference on Artificial Neural Networks (ICANN)*, 1–42. https://doi.org/http://arxiv.org/abs/2401.10825

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*, *82*(3), 3713–3744. https://doi.org/10.1007/s11042-022-13428-4

Kumar, A., Sharma, R., & Bedi, P. (2024). Towards Optimal NLP Solutions: Analyzing GPT and LLaMA-2 Models Across Model Scale, Dataset Size, and Task Diversity. *Engineering, Technology and Applied Science Research*, *14*(3), 14219–14224. https://doi.org/10.48084/etasr.7200

Kusumawardani, R. P., & Kusumawati, K. N. (2024). Named entity recognition in the medical domain for Indonesian language health consultation services using bidirectional-lstmcrf algorithm. *9th International Conference on Computer Science and Computational Intelligence 2024 (ICCSCI 2024)*, *245*, 1146–1156. https://doi.org/10.1016/j.procs.2024.10.344

Li, J., Sun, A., Han, J., & Li, C. (2020). *A Survey on Deep Learning for Named Entity Recognition*. http://neuroner.com/

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., & He, L. (2022). A Survey on Text Classification: From Traditional to Deep Learning. In *ACM Transactions on Intelligent Systems and Technology* (Vol. 13, Number 2). Association for Computing Machinery. https://doi.org/10.1145/3495162

Ma, P., Jiang, B., Lu, Z., Li, N., & Jiang, Z. (2021). Cybersecurity Named Entity Recognition Using Bidirectional Long Short-Term Memory with Conditional Random Fields. *Tsinghua Science and Technology*, *26*(3), 259–265. https://doi.org/10.26599/TST.2019.9010033

Marreddy, M., Oota, S. R., Vakada, L. S., Chinni, V. C., & Mamidi, R. (2022). Am I a Resource-Poor Language? Data Sets, Embeddings, Models and Analysis for four different NLP Tasks in Telugu Language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *22*(1), 1–35. https://doi.org/10.1145/3531535

Menghani, G. (2023). Efficient Deep Learning: A Survey on Making Deep Learning Models Smaller, Faster, and Better. In *ACM Computing Surveys* (Vol. 55, Number 12). Association for Computing Machinery. https://doi.org/10.1145/3578938

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2022). Deep Learning-Based Text Classification. *ACM Computing Surveys*, *54*(3), 1–40. https://doi.org/10.1145/3439726

Mohamed, A., Najafabadi, M. K., Wah, Y. B., Zaman, E. A. K., & Maskat, R. (2020). The state of the art and taxonomy of big data analytics: view from new big data framework. *Artificial Intelligence Review*, *53*(2), 989–1037. https://doi.org/10.1007/s10462-019-09685-9

Murakami, E., Shionoya, T., Komenoi, S., Suzuki, Y., & Sakane, F. (2025). Cloning and characterization of novel testis-Specific diacylglycerol kinase η splice variants 3 and 4. *PLoS ONE*, *11*(9), 1–14. https://doi.org/10.1371/journal.pone

Nabiilah, G. Z., Alam, I. N., Purwanto, E. S., & Hidayat, M. F. (2024). Indonesian multilabel classification using IndoBERT embedding and MBERT classification. *International Journal of Electrical and Computer Engineering*, *14*(1), 1071–1078. https://doi.org/10.11591/ijece.v14i1.pp1071-1078

Olthof, A. W., van Ooijen, P. M. A., & Cornelissen, L. J. (2021). Deep Learning-Based Natural Language Processing in Radiology: The Impact of Report Complexity, Disease Prevalence, Dataset Size, and Algorithm Type on Model Performance. *Journal of Medical Systems*, *45*(10), 1–16. https://doi.org/10.1007/s10916-021-01761-4

O'Shaughnessy, D. (2026). An Overview of Recent Advances in Natural Language Processing for Information Systems. *Applied Sciences*, *16*(2), 1122. https://doi.org/10.3390/app16021122

Patel, P. (2025). Infrastructure Economics of Sparse Mixture-of-Experts in Cloud-Native NLP: Benchmarking Cost, Accuracy, and Performance. *Global Business & Economics Journal*, *27*(1), 1–18. https://doi.org/10.70924/f83n6wqz/vboj68ls

Pogiatzis, A., & Samakovitis, G. (2020). Using bilstm networks for context-aware deep sensitivity labelling on conversational data. *Applied Sciences (Switzerland)*, *10*(24), 1–17. https://doi.org/10.3390/app10248924

Qiu, Y., Dong, L., Zhang, W., Xing, H., & Huang, J. (2025a). A diffusion enhanced CRF and BiLSTM framework for accurate entity recognition. *Scientific Reports*, *15*(1), 1–25. https://doi.org/10.1038/s41598-025-04036-x

Qiu, Y., Dong, L., Zhang, W., Xing, H., & Huang, J. (2025b). A diffusion enhanced CRF and BiLSTM framework for accurate entity recognition. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-025-04036-x

Rahim, A., Zhong, Y., Ahmad, T., Ahmad, S., Pławiak, P., & Hammad, M. (2023). Enhancing Smart Home Security: Anomaly Detection and Face Recognition in Smart Home IoT Devices Using Logit-Boosted CNN Models. *Sensors*, *23*(15), 1–42. https://doi.org/10.3390/s23156979

Salmani, M., Ghafouri, S., Sanaee, A., Razavi, K., Mühlhäuser, M., Doyle, J., Jamshidi, P., & Sharifi, M. (2023). Reconciling High Accuracy, Cost-Efficiency, and Low Latency of Inference Serving Systems. *EuroMLSys '23: Proceedings of the 3rd Workshop on Machine Learning and Systems*, 78–86. https://doi.org/10.1145/3578356.3592578

Shidik, G. F., Saputra, F. O., Saraswati, G. W., Winarsih, N. A. S., Rohman, M. S., Pramunendar, R. A., Kusuma, E. J., Ratmana, D. O., Venus, V., Andono, P. N., & Hasibuan, Z. A. (2024). Indonesian disaster named entity recognition from multi source information using bidirectional LSTM (BiLSTM). *Journal of Open Innovation: Technology, Market, and Complexity*, *10*(3), 1–12. https://doi.org/10.1016/j.joitmc.2024.100358

Singgalen, Y. A. (2025). Performance Analysis of IndoBERT for Sentiment Classification in Indonesian Hotel Review Data. *Journal of Information System Research (JOSH)*, *6*(2), 976–986. https://doi.org/10.47065/josh.v6i2.6505

Surya Suwardi Ansyah, A., Oranova Siahaan, D., Rizqi Paradisiaca Darnoto, B., Nopember, S., Studi Informatika, P., Ilmu Komputer, F., Jember, U., Krajan Timur, J., Sumbersari, K., Jember, K., & Timur, J. (2025). Integrated Named Entity Recognition and Identical-Entity Detection for Extracting Unique Information Sources in News Articles. *Jurnal Teknologi Informasi Dan Komunikasi*, *16*(2), 72–83. https://doi.org/10.31849/digitalzone.v16i2

Wang, J., Yue, K., & Duan, L. (2023). Models and Techniques for Domain Relation Extraction: A Survey. In *Journal of Data Science and Intelligent Systems* (Vol. 1, Number 2, pp. 65–82). Bon View Publishing Pte Ltd. https://doi.org/10.47852/bonviewJDSIS3202973

Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J., & Wang, G. (2025). *Findings of the Association for Computational GPT-NER: Named Entity Recognition via Large Language Models*. https://github.com/

Zhang, Y., & Xiao, G. (2024). Named Entity Recognition Datasets: A Classification Framework. In *International Journal of Computational Intelligence Systems* (Vol. 17, Number 1). Springer Science and Business Media B.V. https://doi.org/10.1007/s44196-024-00456-1

Zhu, Y. (2024). A knowledge graph and BiLSTM-CRF-enabled intelligent adaptive learning model and its potential application. *Alexandria Engineering Journal*, *91*(1), 305–320. https://doi.org/10.1016/j.aej.2024.02.011