International Journal of Machine Learning
P-ISSN: xxxx-xxxx
E-ISSN: xxxx-xxxx
Vol. 1, No. 1, June 2026, Page. 12-22
https://ijoml.com/
DOI: https://doi.org/10.52436/1.ijoml.1.1.id

Benchmarking Modern Optimizers for IndoBERT-Based Sentiment Analysis on Indonesian Gojek Reviews

Randi Rizal*[1], Hidayatulah Himawan[2]

[1]Department of Informatics, Faculty of Engineering, Siliwangi University, Indonesia
[2]Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Malaysia

Email: [1]randirizal@unsil.ac.id

Received: Jan 9, 2026; Revised: Feb 28, 2026; Accepted: Feb 28, 2026; Published: June 1, 2026

Abstract

User reviews on platforms like Gojek serve as critical data for business intelligence, necessitating robust automated sentiment analysis models. While IndoBERT is the standard architecture for Indonesian natural language processing, the comparative impact of emerging optimizers on its performance remains underexplored, as most existing studies default to AdamW without investigating modern alternatives. This research comprehensively benchmarks five optimizers—AdamW, Muon, AdaMuon, Lion, and Sophia—by fine-tuning IndoBERT on 29,851 Indonesian Gojek reviews to identify the most effective training strategy. The study evaluates classification metrics alongside computational efficiency indicators, including training duration and peak memory usage. Empirical results demonstrate that AdamW, AdaMuon, and Lion achieve statistically equivalent superior performance, attaining an average accuracy of 91.6% and an F1-macro of 91.5%. Conversely, Muon and Sophia exhibit slightly lower predictive capability with higher resource demands. Regarding computational cost, AdamW and Lion provide the optimal balance of rapid convergence and memory efficiency, whereas Sophia demands significantly higher VRAM and matrix-based optimizers like Muon extend training duration. These findings confirm that AdamW remains the most robust and efficient choice for analyzing informal Indonesian text, indicating that the complex update mechanisms of newer optimizers do not yield necessary marginal gains for this specific classification task.

Keywords: *AdamW, Computational Efficiency, IndoBERT, Lion Optimizer, Sentiment Analysis.*

1. INTRODUCTION

Ride-hailing and on-demand services such as Gojek have become vital elements within Indonesia's digital ecosystem, where business sustainability and competitiveness are highly dependent on the quality of user experience, which is dynamically reflected in digital reviews [1]. User reviews that are massively available on application distribution platforms such as the Google Play Store constitute a primary data source rich in information regarding satisfaction levels, specific complaints, and technical issues encountered by users [1], [2], [3]. Given the exponentially increasing volume of reviews, manual analysis approaches become inefficient and prone to subjectivity, rendering automated sentiment analysis based on text mining and Natural Language Processing (NLP) a crucial approach for extracting user aspirations and appreciations in a rapid, consistent, and measurable manner compared to conventional methods [3], [4], [5], [6].

In recent developments in natural language processing technology, Transformer-based model architectures have dominated various text classification tasks due to their superior ability to capture long-range dependencies and complex semantic contexts through self-attention mechanisms [7], [8], [9]. Although multilingual models are available, the use of monolingual pretrained models specifically tailored to the target language has been shown to provide more accurate linguistic representations [5],

[9]. IndoBERT, as a BERT-based model trained on a massive Indonesian-language corpus (Indo4B), offers linguistic representations that are substantially more suitable for handling the unique characteristics of the Indonesian language compared to multilingual models [9], [10], [11].

While model architecture plays a central role, the final performance of Transformer models on downstream tasks is also strongly influenced by the choice of optimization algorithm and hyperparameter configuration during the fine-tuning stage [4], [11], [12]. To date, AdamW, a variant of Adam with a decoupled weight decay mechanism, remains the most commonly used optimizer in training Transformer models, including the BERT family [7], [13], [14], [15], [16], [17]. The majority of sentiment analysis studies utilizing IndoBERT continue to rely exclusively on AdamW as the primary optimizer [4], [16], [18], [19], without conducting in-depth exploration of alternative optimization algorithms.

Along with advances in neural network optimization research, new optimizers such as Muon [20], AdaMuon [21], Lion [22], and Sophia [23] have emerged, offering promising computational efficiency and competitive performance. The Lion optimizer (EvoLved Sign Momentum), discovered through symbolic program search, provides higher memory efficiency by tracking only momentum and employing sign-based operations for uniform parameter updates [22], [24], [25]. Meanwhile, Sophia (Second-order Clipped Stochastic Optimization) is introduced as a lightweight second-order optimizer that utilizes diagonal Hessian estimates to achieve faster convergence and lower validation loss compared to AdamW [23]. More recently, Muon has emerged as an optimizer specifically designed for matrix parameters by applying orthogonal updates, which has empirically demonstrated higher data efficiency in large-scale training [20], [26], [27] Its subsequent development, AdaMuon, attempts to combine the geometric stability of Muon with coordinate-wise adaptivity to further enhance training efficiency [21].

Despite the promise of these modern optimizers, systematic comparative studies evaluating their performance on Indonesian-language review data with unique characteristics remain very limited. This gap in the literature raises a critical question regarding whether AdamW remains the optimal choice, or whether modern optimizers such as Muon, AdaMuon, Lion, and Sophia can deliver superior performance when handling Indonesian review data rich in slang, emojis, and noise. Therefore, this study aims to comprehensively compare the performance of AdamW, Muon, AdaMuon, Lion, and Sophia on the task of sentiment analysis of Gojek application reviews using the IndoBERT model. This research is expected to provide a structured benchmark on the effectiveness of modern optimizers within the IndoBERT architecture and to offer practical guidance for NLP practitioners in selecting the most efficient optimization strategies for sentiment analysis in the Indonesian language environment.

2. METHOD

The research methodology was designed to compare the performance and efficiency of five modern optimizers in fine-tuning IndoBERT for sentiment analysis of Gojek reviews, using a controlled workflow that includes data curation, text pre-processing, modeling, optimization, hyperparameter tuning, and evaluation of performance and statistical significance. The overall research workflow is summarized in Figure 1.

International Journal of Machine Learning
P-ISSN: xxxx-xxxx
E-ISSN: xxxx-xxxx
Vol. 1, No. 1, June 2026, Page. 12-22
https://ijoml.com/
DOI: https://doi.org/10.52436/1.ijoml.1.1.id

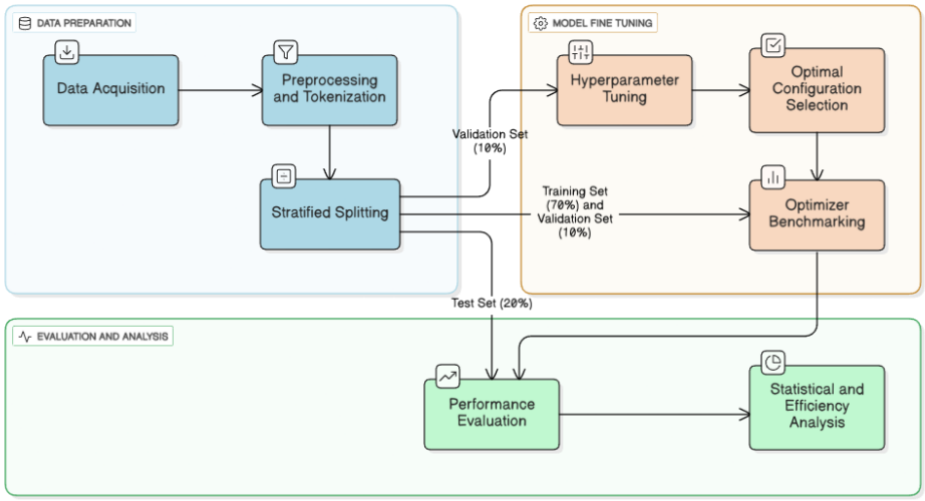

Figure 1. Research methodology workflow diagram

2.1. Dataset

The research dataset consists of Indonesian-language user reviews of the Gojek application from the Google Play Store, obtained via a public Kaggle dataset with an initial corpus of 387.645 reviews. All entries were cleaned by removing empty reviews and reviews containing ≤3 words, resulting in 29.851 reviews; subsequently, the text length of the preprocessed_content column was analyzed across the training, validation, and test sets, showing similar length distributions that are representative of the original corpus. Sentiment labels were automatically assigned based on ratings, with ratings 1–3 mapped to the negative class and 4–5 mapped to the positive class, yielding 15.903 negative reviews and 13.948 positive reviews. The final corpus was then stratified into 70% training data, 10% validation data, and 20% test data while preserving class proportions in each subset, following common practice in machine learning- and deep learning-based classification studies [28], [29], [30].

2.2. Pre-Processing

The text pre-processing stage is a foundation of sentiment analysis because it transforms raw reviews into clean, standardized representations that are ready to be processed by machine learning and deep learning models [2], [11], [18]. The process begins by removing noise such as HTML tags, URL links, user mentions (@username), hashtags, repeated punctuation, special characters, and numbers, so that the model focuses on relevant textual features [13], [16], [31], [32]. Emojis are not removed; instead, they are converted into text tokens (demojize) to preserve the affective information and sentiment nuances they convey [33], [34]. Representation consistency is strengthened through case folding to lowercase, which reduces feature redundancy due to capitalization variations [13], [16], [31]. Given the dominance of informal varieties in Indonesian-language reviews, normalization of non-standard words and abbreviations into formal forms is performed using an external slang dictionary to improve the model's semantic understanding of the text [19], [35].

2.3. Model and Tokenization

The IndoBERT model variant indobenchmark/indobert-base-p1 is used within a transfer learning framework as the primary backbone by employing the BertForSequenceClassification architecture, where a linear layer is placed on top of the [CLS] token representation to produce probabilities for two sentiment classes [12]. The selection of IndoBERT is based on empirical evidence indicating superior performance on Indonesian NLP tasks because it is trained on the large Indo4B corpus, which spans formal to colloquial language varieties [10]. The maximum sequence length is determined through an

analysis of the token distribution at the 99th percentile to maintain computational efficiency while preserving information coverage [14]. Tokenization is performed using the IndoBERT tokenizer to convert text into input IDs and attention masks, with dynamic padding and truncation mechanisms to ensure consistent tensor dimensions across batches during training on Transformer architectures [16], [19].

2.4. Optimization Configuration and Training Strategy

This study compares five modern optimizers—AdamW as the baseline and Lion, Sophia, Muon, and AdaMuon—which represent different optimization paradigms in fine-tuning Transformer models. For Muon and AdaMuon, the training strategy adopts a hybrid approach: Newton–Schulz orthogonalization updates are applied to two-dimensional parameters (weight matrix), whereas embeddings, biases, and LayerNorm parameters are optimized using AdamW to maintain training stability for non-matrix parameters [20], [21], [27]. The integration of decoupled weight decay is applied consistently to AdamW, Muon, and AdaMuon due to its role in maintaining parameter RMS scale and reducing overfitting in large-scale models [27]. The effectiveness of large batch sizes is maximized through gradient accumulation to improve data efficiency without burdening GPU memory [36]. The entire fine-tuning process is conducted with mixed precision bfloat16 to accelerate computation and reduce memory usage [20], [36].

Hyperparameter tuning is conducted via grid search on a 10% subset of the training data, with model re-initialization in each trial to maintain evaluation consistency. Learning rate ranges are calibrated according to the characteristics of each optimizer: AdamW uses 2e-5–5e-5 following standard fine-tuning practice [7], [8], [14], [16], Muon and AdaMuon use 1e-3–5e-3 to accommodate the dynamics of orthogonal matrix updates [21], [27], [37], Lion uses 3e-6–1e-5 due to the magnitude of the update norm induced by sign operations [22], [25], and Sophia uses 1e-5–5e-5 consistent with its diagonal Hessian estimation design [23]. Each configuration is evaluated through short training for 20 steps to prevent overfitting on the proxy subset, and the best model is selected based on minimizing validation loss. In the final training stage, each model–optimizer combination is trained on the full training data for 3 epochs and repeated three times with different random seeds to reduce the influence of training stochasticity and yield more stable performance estimates.

2.5. Evaluation and Analysis Protocol

Model performance evaluation includes accuracy, precision, recall, as well as macro and per-class F1-scores to assess the model's ability to distinguish sentiment [6], [34]. The Area Under the Curve - Receiver Operating Characteristic (AUC-ROC) is computed as a threshold-independent indicator of discriminative capability [38]. Computational efficiency is measured through total training time, inference throughput (samples per second), and maximum GPU memory consumption (peak VRAM) during training as a basis for comparing memory footprints across modern optimizers such as Lion and AdamW. To ensure that performance differences are not merely random fluctuations, McNemar's significance test with continuity correction is applied to compare the baseline AdamW predictions with those of each other optimizer on the run with the best validation performance for each optimizer.

3. RESULT

This section presents the results of a comparative experiment involving five optimizers—AdamW, Muon, AdaMuon, Lion, and Sophia—on the task of sentiment classification of Gojek application reviews using IndoBERT. All models were evaluated on 5.971 test samples with three independent training runs, and the reported metrics are expressed as mean ± standard deviation to represent performance stability across runs.

Vol. 1, No. 1, June 2026, Page. 12-22
https://ijoml.com/
DOI: https://doi.org/10.52436/1.ijoml.1.1.id

Overall, AdamW, AdaMuon, and Lion exhibit nearly identical performance on aggregate metrics. AdamW achieves an accuracy of 0,916 ± 0,004 with an F1-macro of 0,915 ± 0,004 and an AUC-ROC of 0,963 ± 0,001. AdaMuon attains an accuracy of 0,916 ± 0,002, an F1-macro of 0,916 ± 0,002, and an AUC-ROC of 0,962 ± 0,001, while Lion yields an accuracy of 0,916 ± 0,000, an F1-macro of 0,915 ± 0,000, and an AUC-ROC of 0,963 ± 0,001. In contrast, Muon and Sophia perform slightly below the top trio: Muon achieves an accuracy of 0,903 ± 0,004 and an F1-macro of 0,902 ± 0,004 with an AUC-ROC of 0,955 ± 0,003, whereas Sophia produces an accuracy of 0,904 ± 0,014, an F1-macro of 0,902 ± 0,015, and an AUC-ROC of 0,957 ± 0,005. Despite these differences in aggregate performance, all configurations maintain AUC-ROC values above 0,95, indicating strong discriminative capability in distinguishing positive and negative sentiment. A summary of global performance alongside computational cost is presented in Table 1.

Table 1. Mean ± standard deviation of overall performance and computational cost for each optimizer

Optimizer	Accuracy (mean ± std)	F1-Macro (mean ± std)	AUC-ROC (mean ± std)	Train Time (s)	Inference Speed (samp/s)	Peak VRAM (MB)
AdamW	0,916 ± 0,004	0,915 ± 0,004	0,963 ± 0,001	4.426,9	157,8	3.877,9
Muon	0,903 ± 0,004	0,902 ± 0,004	0,955 ± 0,003	6.075,2	161,0	3.885,5
AdaMuon	0,916 ± 0,002	0,916 ± 0,002	0,962 ± 0,001	6.151,6	159,7	4.847,2
Lion	0,916 ± 0,000	0,915 ± 0,000	0,963 ± 0,001	4.417,2	159,9	4.858,5
Sophia	0,904 ± 0,014	0,902 ± 0,015	0,957 ± 0,005	4.486,9	160,3	5.824,4

Table 1 shows that inference throughput across optimizers is practically equivalent (approximately 158–161 samples per second), indicating that optimizer choice in this setting does not meaningfully affect inference speed. More relevant variations emerge in GPU memory consumption and training time. AdamW and Muon are the most VRAM-efficient configurations (≈3,88 GB), whereas AdaMuon and Lion require higher VRAM (≈4,85 GB). Sophia is the most memory-intensive (≈5,82 GB), approximately 50% higher than AdamW. In terms of training time, AdamW and Lion are the fastest (≈4,4×10³ seconds), while Muon and AdaMuon require substantially longer durations (above 6×10³ seconds), consistent with the higher complexity of their parameter update mechanisms. Thus, the primary trade-off observed is not between accuracy and inference speed, but between marginal accuracy gains and computational cost in terms of VRAM usage and training duration.

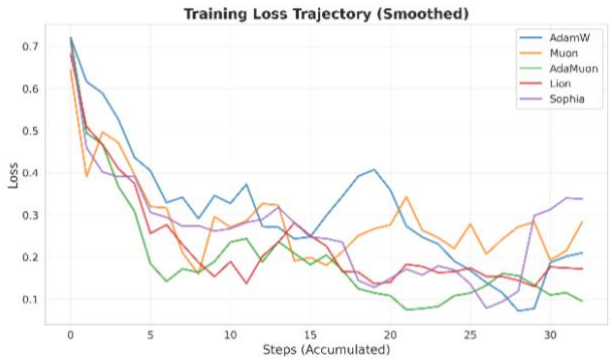

Figure 2. Training Loss Convergence Curves Across Optimizers

The training loss curves in Figure 2 show that all optimizers exhibit stable convergence without indications of numerical divergence during fine-tuning. This pattern is consistent with the final evaluation results: AdamW, AdaMuon, and Lion maintain lower final loss values, whereas Muon and

Vol. 1, No. 1, June 2026, Page. 12-22
https://ijoml.com/
DOI: https://doi.org/10.52436/1.ijoml.1.1.id

Sophia tend to converge at slightly higher loss levels, aligning with the observed gap of approximately 1–1,5 percentage points in accuracy and F1-macro relative to the top trio. From a practical perspective, these findings indicate that the observed performance differences are more closely related to the quality of the convergence point rather than issues of optimization stability.

Beyond aggregate metrics, per-class performance indicates that the models with the best overall performance also do not sacrifice performance on either class. Table 2 summarizes the mean F1-scores for the negative and positive classes. AdamW, AdaMuon, and Lion deliver balanced F1-scores across both classes (≈0,91–0,92), while Muon and Sophia achieve slightly lower scores for both classes. Inter-run variability is most pronounced for Sophia, particularly on the negative class (standard deviation 0,030), indicating higher sensitivity to training conditions.

Table 2. Mean ± standard deviation of per-class F1-scores for each optimizer

Optimizer	F1-Neg (mean ± std)	F1-Pos (mean ± std)
AdamW	0,917 ± 0,013	0,914 ± 0,009
Muon	0,905 ± 0,012	0,899 ± 0,005
AdaMuon	0,918 ± 0,008	0,913 ± 0,009
Lion	0,918 ± 0,010	0,912 ± 0,010
Sophia	0,902 ± 0,030	0,902 ± 0,004

From the perspective of probabilistic quality, the ROC and Precision–Recall curves in Figure 3 show ROC curves that lie well above the random baseline for all optimizers, consistent with the AUC-ROC values reported in Table 2. The Precision–Recall curves indicate that AdamW, AdaMuon, and Lion maintain a strong precision–recall balance across a wide range of thresholds, whereas Muon and Sophia experience a more rapid decline in precision as the threshold is lowered. The practical implication is that models based on the top trio tend to be more robust to changes in decision thresholds when deployed in operational scenarios that require adjustment of the false positive–false negative trade-off.

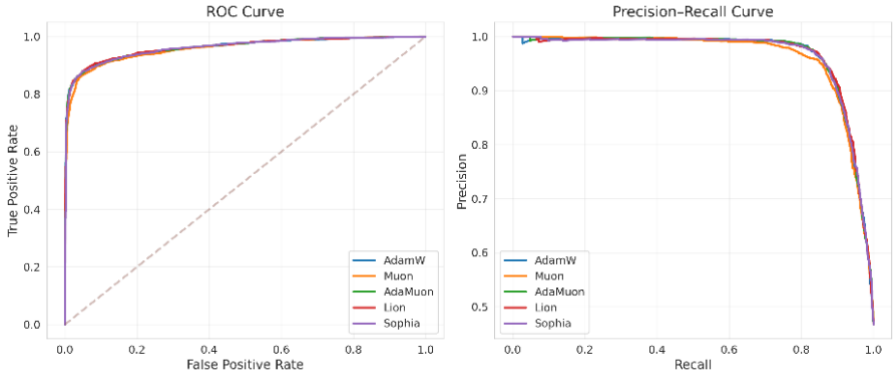

Figure 3. ROC and Precision–Recall Curves Across Optimizers

Differences in accuracy relative to the AdamW baseline were further analyzed using the McNemar test on paired predictions from the test set, as shown in Table 3. In one representative run, Muon yielded n_{01}=154 (only Muon correct) and n_{10}=212 (only AdamW correct), resulting in χ^2=8,877 with p=0,003, indicating that the accuracy decrease of Muon relative to AdamW in that run is significant at the 5% level. In contrast, AdaMuon, Lion, and Sophia show p-values far above 0,05, indicating that their accuracy differences relative to AdamW are not significant in that run. It should be noted that the test results in Table 3 are based on a single run, and thus their interpretation is limited to indicating

Vol. 1, No. 1, June 2026, Page. 12-22
https://ijoml.com/
DOI: https://doi.org/10.52436/1.ijoml.1.1.id

significance under specific training conditions rather than providing fully generalizable inferential conclusions across all inter-run variability.

Table 3. McNemar Test Results Comparing Optimizers Against AdamW on the Test Set

Optimizer	n_{01} (only opt correct)	n_{10} (only AdamW correct)	χ^2 McNemar	p-value
Muon	154	212	8,877	0,003
AdaMuon	123	121	0,004	0,949
Lion	126	112	0,710	0,399
Sophia	127	124	0,016	0,900

Computational efficiency aspects are visualized in Figure 4 through the relationship between F1-macro, training time, and peak VRAM. AdamW and Lion occupy regions characterized by high F1-macro and relatively short training durations, whereas Muon and AdaMuon shift toward longer training times. Sophia appears in a less favorable configuration, combining lower F1-macro with the highest VRAM consumption, such that no clear compensation is observed in terms of the trade-off between accuracy and computational cost under the experimental configuration.

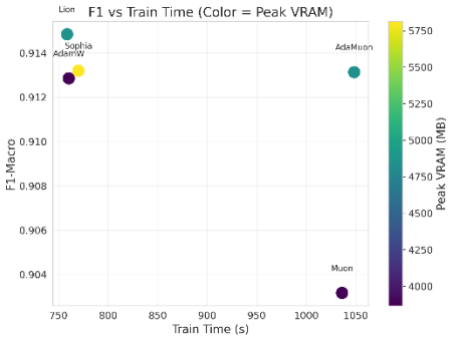

Figure 4. Performance–Computational Cost Trade-off Across Optimizers

Overall, for the IndoBERT-based sentiment analysis task on Gojek reviews, AdamW, AdaMuon, and Lion achieve nearly identical performance (accuracy/F1-macro ≈0,915–0,916 and AUC-ROC ≈0,962–0,963) with balanced per-class results. Muon and Sophia lag slightly behind, although they still demonstrate strong discriminative capability (AUC-ROC >0,95). Given that inference throughput is nearly identical across all optimizers, the choice of optimizer in this setting is primarily determined by the trade-off between marginal accuracy, VRAM requirements, and training time, with AdamW and Lion standing out as configurations that provide a combination of high performance and relatively efficient training cost in this experiment.

4. DISCUSSIONS

This comparative analysis confirms that AdamW, AdaMuon, and Lion exhibit statistically and practically equivalent performance with minimal metric differences, thereby addressing the research question by showing that modern optimizers do not consistently outperform AdamW in fine-tuning IndoBERT for sentiment analysis of Gojek reviews. This finding reinforces the body of literature that positions AdamW as a very strong baseline for BERT-based models [7], [13], [14], [15], [16], [17], while at the same time contrasting claims of superiority for Lion [22], Sophia [23], Muon [20], and AdaMuon [21] reported in vision domains or large-scale pretraining, which are not replicated in this short-horizon binary fine-tuning task.

P-ISSN: xxxx-xxxx https://ijoml.com/
E-ISSN: xxxx-xxxx DOI: https://doi.org/10.52436/1.ijoml.1.1.id

From a technical perspective, the observed performance parity among AdamW, AdaMuon, and Lion, supported by non-significant McNemar test results, suggests that the loss landscape of this classification task is relatively smooth, such that variations in momentum update schemes or orthogonalization provide only marginal benefits over standard adaptive mechanisms. In contrast, the slightly inferior performance of Muon and Sophia implies that their diagonal Hessian-based updates or pure orthogonalized momentum updates are less well aligned with rapid adaptation on review datasets characterized by a high degree of linguistic informality without extensive hyperparameter tuning.

Given the identical inference throughput across models, the practical implication is that optimizer selection is primarily determined by the trade-off between accuracy, stability, training time, and VRAM consumption; AdamW and Lion emerge as the most rational primary choices, while AdaMuon offers an alternative when cross-run robustness is prioritized, whereas Muon and Sophia are not competitive in terms of cost-to-F1 ratio under this configuration. These conclusions apply specifically to the IndoBERT-base architecture in the Gojek application review domain, and further validation on larger language models, low-resource scenarios, continual learning settings, and integration with modern regularization techniques is therefore required to test the limits of the claim that the dominance of AdamW remains difficult to displace.

5. CONCLUSION

This comparative study empirically confirms that the dominance of AdamW in the fine-tuning of IndoBERT for sentiment analysis of Indonesian-language reviews remains unmatched by modern optimizers such as Lion, AdaMuon, Muon, and Sophia. Although AdamW, Lion, and AdaMuon demonstrate statistically equivalent classification performance with identical F1-macro values of 0,915, AdamW proves to be superior in terms of computational efficiency by maintaining an optimal balance between shorter training duration and lower GPU memory consumption. In contrast, the complexity of orthogonal update mechanisms in Muon and the diagonal Hessian estimation in Sophia fails to yield substantial adaptive benefits for the noisy characteristics of review data, instead resulting in marginal accuracy degradation and a disproportionate increase in resource overhead.

These findings indicate that the optimization landscape for short-horizon text classification tasks with IndoBERT is relatively stable and does not require complex higher-order algorithmic interventions, while also recommending that NLP practitioners continue to prioritize AdamW as an efficient and robust primary standard, or consider Lion as a competitive alternative when memory constraints become a secondary priority. Further validation on larger-scale language model architectures or in continual learning scenarios is required to assess the generalization limits of this baseline performance robustness in future work.

CONFLICT OF INTEREST

The authors declares that there is no conflict of interest between the authors or with research object in this paper.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the support and assistance of all parties involved in the conduct of this research and the writing of this article, particularly colleagues and reviewers, whose insights, comments, and suggestions contributed significantly to the refinement and completion of this work.

REFERENCES

[1] S. P. Yuliani, A. A. P. Muharani, R. Q. Fatmawati, and F. Fahmi, "Sentiment Analysis in User Reviews of Gojek Application using Natural Language Processing," *Journal of System and Computer Engineering (JSCE)*, vol. 6, no. 4, pp. 296–305, Oct. 2025, doi: 10.61628/jsce.v6i4.2062.

[2] S. Sanjaya, R. G. Guntara, and S. S. Maesaroh, "Sentiment Analysis of LinkAja Digital Wallet Application Reviews on Google Play Store using Transfer Learning IndoBERT," *INOVTEK Polbeng - Seri Informatika*, vol. 10, no. 3, pp. 1730–1740, Nov. 2025, doi: 10.35314/afjx7b16.

[3] A. R. Prananda and I. Thalib, "Sentiment Analysis for Customer Review: Case Study of GO-JEK Expansion," *Journal of Information Systems Engineering and Business Intelligence*, vol. 6, no. 1, p. 1, Apr. 2020, doi: 10.20473/jisebi.6.1.1-8.

[4] K. S. Nugroho, A. Y. Sukmadewa, H. Wuswilahaken DW, F. A. Bachtiar, and N. Yudistira, "BERT Fine-Tuning for Sentiment Analysis on Indonesian Mobile Apps Reviews," in *6th International Conference on Sustainable Information Engineering and Technology 2021*, New York, NY, USA: ACM, Sep. 2021, pp. 258–264. doi: 10.1145/3479645.3479679.

[5] D. G. Mandhasiya, H. Murfi, and A. Bustamam, "The hybrid of BERT and deep learning models for Indonesian sentiment analysis," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 33, no. 1, p. 591, Jan. 2024, doi: 10.11591/ijeecs.v33.i1.pp591-602.

[6] A. A. P. Simarmata and T. B. Sasongko, "Sentiment Analysis on BRImo Application Reviews Using IndoBERT," *Journal of Applied Informatics and Computing*, vol. 9, no. 3, pp. 851–862, Jun. 2025, doi: 10.30871/jaic.v9i3.8162.

[7] O. El Azzouzy, T. Chanyour, and S. J. Andaloussi, "Transformer-based models for sentiment analysis of YouTube video comments," *Sci Afr*, vol. 29, p. e02836, Sep. 2025, doi: 10.1016/j.sciaf.2025.e02836.

[8] Y. Li, "Building and Optimizing Deep Learning Models for Sentiment Analysis in English Text," *Journal of Cases on Information Technology*, vol. 27, no. 1, pp. 1–17, Jun. 2025, doi: 10.4018/JCIT.382380.

[9] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," in *Proceedings of the 28th International Conference on Computational Linguistics*, Stroudsburg, PA, USA: International Committee on Computational Linguistics, 2020, pp. 757–770. doi: 10.18653/v1/2020.coling-main.66.

[10] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 843–857. doi: 10.18653/v1/2020.aacl-main.85.

[11] G. Z. Nabiilah, I. N. Alam, E. S. Purwanto, and M. F. Hidayat, "Indonesian multilabel classification using IndoBERT embedding and MBERT classification," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 14, no. 1, p. 1071, Feb. 2024, doi: 10.11591/ijece.v14i1.pp1071-1078.

[12] E. Yulianti and N. K. Nissa, "ABSA of Indonesian customer reviews using IndoBERT: single-sentence and sentence-pair classification approaches," *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 5, pp. 3579–3589, Oct. 2024, doi: 10.11591/eei.v13i5.8032.

[13] A. Maretta and A. Meiriza, "Aspect-Based Sentiment Analysis of Hospital Service Reviews Using Fine-Tuned IndoBERT," *Journal of Applied Informatics and Computing*, vol. 9, no. 5, pp. 2541–2551, Oct. 2025, doi: 10.30871/jaic.v9i5.10765.

[14] A. Öcal, "BERT-Based Sentiment Analysis of Turkish e-Commerce Reviews: Star Ratings Versus Text," *Sakarya University Journal of Computer and Information Sciences*, vol. 8, no. 4, pp. 677–687, Oct. 2025, doi: 10.35377/saucis...1747068.

[15] A. S. Saud and A. Dhakal, "Optimizing BERT for Nepali Text Classification: The Role of Stemming and Gradient Descent Optimizers," *Aadim Journal of Multidisciplinary Research*, vol. 1, pp. 25–38, Jul. 2025, doi: 10.3126/ajmr.v1i1.82292.

[16] L. Afuan, N. Hidayat, H. Hamdani, H. Ismanto, B. C. Purnama, and D. I. Ramdhani, "Optimizing BERT Models with Fine-Tuning for Indonesian Twitter Sentiment Analysis," *J Wirel Mob Netw*

Ubiquitous Comput Dependable Appl, vol. 16, no. 2, pp. 248–267, Jun. 2025, doi: 10.58346/JOWUA.2025.I2.016.

[17] U. Yagci, E. Iscan, and A. Kolcak, "ReBERT at HSD-2Lang 2024: Fine-Tuning BERT with AdamW for Hate Speech Detection in Arabic and Turkish," in *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, A. Hürriyetoğlu, H. Tanev, S. Thapa, and G. Uludoğan, Eds., St. Julians, Malta: Association for Computational Linguistics, Mar. 2024, pp. 195–198. [Online]. Available: https://aclanthology.org/2024.case-1.27/

[18] V. D. Setiawan, D. U. Iswavigra, and E. Anggiratih, "Implementation of IndoBERT for Sentiment Analysis of the Constitutional Court's Decision Regarding the Minimum Age of Vice Presidential Candidates," *Scientific Journal of Informatics*, vol. 12, no. 3, pp. 397–406, Aug. 2025, doi: 10.15294/sji.v12i3.26320.

[19] M. N. Zaidan, Y. Sibaroni, and S. S. Prasetyowati, "Learning Rate And Epoch Optimization In The Fine-Tuning Process For Indobert's Performance On Sentiment Analysis Of Mytelkomsel App Reviews," *Jurnal Teknik Informatika (Jutif)*, vol. 5, no. 5, pp. 1443–1450, Oct. 2024, doi: 10.52436/1.jutif.2024.5.5.2396.

[20] K. Jordan *et al.*, "Muon: An optimizer for hidden layers in neural networks," 2024. [Online]. Available: https://kellerjordan.github.io/posts/muon/

[21] C. Si, D. Zhang, and W. Shen, "AdaMuon: Adaptive Muon Optimizer," 2025. [Online]. Available: https://arxiv.org/abs/2507.11005

[22] X. Chen *et al.*, "Symbolic discovery of optimization algorithms," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, in NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2023.

[23] H. Liu, Z. Li, D. Hall, P. Liang, and T. Ma, "Sophia: A Scalable Stochastic Second-order Optimizer for Language Model Pre-training," 2024. [Online]. Available: https://arxiv.org/abs/2305.14342

[24] T. Do, T. T. Doan, K. Le, T. Nguyen, D.-D. Le, and T. D. Ngo, "Key Information Extraction and Recognition from Rich Text Images," *Vietnam Journal of Computer Science*, vol. 11, no. 04, pp. 569–594, Nov. 2024, doi: 10.1142/S2196888824500131.

[25] S. Kumar, M. Pande, and A. Y. Damle, "Comparative Analysis of Lion and AdamW Optimizers for Cross-Encoder Reranking with MiniLM, GTE, and ModernBERT," 2025. [Online]. Available: https://arxiv.org/abs/2506.18297

[26] M. Crawshaw, C. Modi, M. Liu, and R. M. Gower, "An Exploration of Non-Euclidean Gradient Descent: Muon and its Many Variants," 2025. [Online]. Available: https://arxiv.org/abs/2510.09827

[27] J. Liu *et al.*, "Muon is Scalable for LLM Training," 2025. [Online]. Available: https://arxiv.org/abs/2502.16982

[28] S. Al-Dabet, B. Alomar, S. Turaev, and A. N. Belkacem, "Dual-Task Learning for AI-Generated Medical Text Detection and Named Entity Recognition," in *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, M. Abbas and A. A. Freihat, Eds., Trento: Association for Computational Linguistics, Oct. 2024, pp. 157–167. [Online]. Available: https://aclanthology.org/2024.icnlsp-1.18/

[29] A. Qasim, G. Mehak, N. Hussain, A. Gelbukh, and G. Sidorov, "Detection of Depression Severity in Social Media Text Using Transformer-Based Models," *Information*, vol. 16, no. 2, p. 114, Feb. 2025, doi: 10.3390/info16020114.

[30] M. D. S. Antariksa, A. Sugiharto, and B. Surarso, "BERT Model Fine-tuned for Scientific Document Classification and Recommendation," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 9, no. 4, pp. 754–764, Aug. 2025, doi: 10.29207/resti.v9i4.6789.

[31] G. Kaur, S. Haraldsson, and A. Bracciali, "Comparative analysis of transformer models for sentiment classification of UK CBDC discourse on X," *Discover Analytics*, vol. 3, no. 1, p. 7, Jun. 2025, doi: 10.1007/s44257-025-00035-4.

[32] L. P. Mudarakola, R. K. Gatla, A. S. N. Raju, A. Y. Jaffar, A. Alzahrani, and A. A. Dessalegn, "Multi stage sentiment analysis for product reviews on Twitter using optimized machine learning algorithm," *Sci Rep*, vol. 15, no. 1, p. 39777, Nov. 2025, doi: 10.1038/s41598-025-23451-8.

[33] A. Khan, D. Majumdar, and B. Mondal, "Sentiment analysis of emoji fused reviews using machine learning and Bert," *Sci Rep*, vol. 15, no. 1, p. 7538, Mar. 2025, doi: 10.1038/s41598-025-92286-0.

[34] M. Pota, M. Ventura, R. Catelli, and M. Esposito, "An Effective BERT-Based Pipeline for Twitter Sentiment Analysis: A Case Study in Italian," *Sensors*, vol. 21, no. 1, p. 133, Dec. 2020, doi: 10.3390/s21010133.

[35] J. F. Kusuma and A. Chowanda, "Indonesian Hate Speech Detection Using IndoBERTweet and BiLSTM on Twitter," *JOIV : International Journal on Informatics Visualization*, vol. 7, no. 3, pp. 773–780, Sep. 2023, doi: 10.30630/joiv.7.3.1035.

[36] J. Schlotthauer, C. Kroos, C. Hinze, V. Hangya, L. Hahn, and F. Küch, "Pre-Training LLMs on a budget: A comparison of three optimizers," 2025. [Online]. Available: https://arxiv.org/abs/2507.08472

[37] I. Shah *et al.*, "Practical Efficiency of Muon for Pretraining," 2025. [Online]. Available: https://arxiv.org/abs/2505.02222

[38] A. Rahman *et al.*, "Multilingual sentiment analysis in restaurant reviews using aspect focused learning," *Sci Rep*, vol. 15, no. 1, p. 28371, Aug. 2025, doi: 10.1038/s41598-025-12464-y.