Validation-Based Hyperparameter Optimization of a Hybrid IndoBERT-BiGRU Model for Indonesian Sentiment Analysis

Handoko[1)], Elly[2)], Rivaldi Lubis[3)]*

[1, 2] Department of Information System, Universitas Mikroskil
[3] Department of Information Technology, Universitas Mikroskil
* Corresponding author : rivaldi.lubis@mikroskil.ac.id

Abstract

Public sentiment across digital platforms spans multiple domains, each with different linguistic characteristics. This study investigated the hyperparameter configurations selected for a hybrid IndoBERT-BiGRU model across three Indonesian-language datasets and an aggregated dataset. Sentiment pseudo-labels were generated using the Indonesian Sentiment Lexicon (InSet), while Optuna with a tree-structured Parzen estimator (TPE) optimized BiGRU hidden size, number of layers, and dropout rate for the four training settings. Validation Macro F1 was used as the optimization objective. The Global and MyPertamina models selected 384 hidden units, two BiGRU layers, and a dropout rate of 0.2465; the Shin Tae-yong model selected 256 hidden units, one layer, and a dropout rate of 0.2301; and the One For All model selected 128 hidden units, one layer, and a dropout rate of 0.3543. Their validation Macro F1-scores were 81.25%, 82.16%, 61.40%, and 65.71%. Agreement between InSet-generated pseudo-labels and Human_Sentiment labels across 14,526 records, measured using Cohen's kappa, yielded a value of 0.2511. On the held-out test sets, evaluated against InSet-derived pseudo-labels, the models achieved Macro F1-scores of 49.19%, 46.86%, 22.71%, and 30.00% for Global, MyPertamina, Shin Tae-yong, and One For All, respectively. Per-class evaluation showed consistently low F1-scores for the Neutral class. These findings demonstrate differences in the validation-selected configurations but do not establish that domain-specific tuning is superior or that dataset size determines model complexity. The results should be interpreted in light of the limited agreement between InSet and Human_Sentiment labels and the very small One For All test set, which contains only eight records.

Keywords- BiGRU, Domain-aware Model, Hybrid Sentiment Analysis, Hyperparameter Optimization, Pseudo-labeling

1. INTRODUCTION

The volume of public opinion expressed in Indonesian-language text across digital platforms has grown rapidly, spanning heterogeneous domains such as public digital services, sports and social discourse, and cultural or media commentary [1], [2]. Similar patterns of domain-specific opinion growth have also been observed in local policy discourse, such as sentiment surrounding higher-education policy shared through Indonesian social media platforms [3]. Analyzing this opinion at scale to support monitoring of public services and policy evaluation has motivated a substantial body of research on automated sentiment classification [4], [5], including studies directed at emerging societal issues such as climate-related public discourse [6]. Deep learning approaches based on pretrained contextual language representations, in particular Bidirectional Encoder Representations from Transformers (BERT), have demonstrated strong performance in modeling context and semantic nuance in natural language text [7]. For Indonesian, IndoBERT provides a pretrained transformer specifically adapted to the linguistic characteristics of the language, including informal and non-standard expressions common in social media text [8], [9].

A recurring challenge in applying contextual embeddings to sentiment classification is that transformer representations alone do not always capture sequential and polarity-specific patterns efficiently, motivating hybrid architectures that combine transformer embeddings with recurrent layers such as the Bidirectional Gated Recurrent Unit (BiGRU) [10]. Hybrid approaches that further incorporate lexicon-based polarity information alongside deep contextual embeddings have also been shown to improve robustness on short or informal text where deep models alone may struggle to capture explicit sentiment cues [11], [12]. However, opinion data drawn from different domains differ considerably in vocabulary, sentence structure, topic focus, and dataset size, and several survey studies have noted that sentiment analysis performance is highly sensitive to such domain-specific characteristics [4], [5], [13].

DOI : https://doi.org/10.21063/jtif.2026.V14.2.139-148
Creative Commons Attribution-Share Alike 4.0 International License (CC BY-SA 4.0)

Previous studies on cross-domain sentiment classification have investigated how sentiment knowledge learned from one or more source domains can be transferred to a target domain. Fu and Liu proposed a contrastive transformer-based domain adaptation approach for multi-source cross-domain sentiment classification [14], while Zhu et al. developed a topic-driven adaptive network to capture domain-specific information for cross-domain sentiment classification [15]. These studies address cross-domain classification through explicit adaptation or transfer mechanisms. In contrast, the present study does not evaluate cross-domain transfer or generalization. Instead, it independently optimizes the same hybrid IndoBERT-BiGRU architecture for an aggregated dataset and three domain-specific datasets, then compares the resulting validation-optimal hyperparameter configurations.

Studies on Indonesian-language sentiment analysis have investigated hybrid models and recurrent architectures in task-specific settings [12], [16]. Meanwhile, Optuna provides a systematic framework for hyperparameter optimization using methods such as the tree-structured Parzen estimator (TPE) [17], and previous research has examined hyperparameter optimization when fine-tuning pretrained language models [18]. Hyperparameter transfer learning has also investigated whether information obtained from tuning one dataset can guide optimization on related datasets [19]. These studies provide relevant methodological foundations for examining hyperparameter selection across different training settings.

Despite previous studies on cross-domain sentiment classification and hyperparameter optimization, the empirical question of how validation-selected hyperparameter configurations differ when the same hybrid IndoBERT-BiGRU architecture is independently optimized across an aggregated dataset and multiple Indonesian domain-specific datasets remains insufficiently examined. The present study addresses this question by applying the same predefined search space and validation Macro F1 objective to each of the four training settings. Its focus is on comparing the selected configurations and their validation performance, rather than evaluating cross-domain transfer or generalization.

The present study represents an initial stage of a broader research program toward developing a domain-aware hybrid sentiment analysis system. In a subsequent stage, neural topic modeling techniques such as BERTopic [20] may be investigated to identify finer-grained topical structures within Indonesian-language public opinion data, building on approaches that have been applied to extract topics from unstructured text corpora [21], [22]. However, BERTopic is not used to identify or assign domains in the experiments reported in this study. The current study uses an aggregated dataset and three predefined domain-specific datasets and focuses on independently optimizing and comparing their validation-selected hyperparameter configurations. The implementation, optimization, and evaluation of BERTopic are left for future research.

Accordingly, this study addresses two research questions. RQ1: Which hyperparameter configurations, in terms of BiGRU hidden size, number of recurrent layers, and dropout rate, are selected when the same hybrid IndoBERT-BiGRU architecture is optimized independently for an aggregated dataset and three domain-specific datasets? RQ2: How do the selected configurations and their corresponding validation Macro F1-scores differ across these four training settings under the same predefined search space? These questions focus on configuration selection and validation performance and do not establish whether a particular configuration generalizes better across domains.

The contributions of this study are twofold. Methodologically, the study applies an independent Optuna-based hyperparameter optimization procedure using the same predefined search space and validation Macro F1 objective to an aggregated dataset and three domain-specific datasets of a hybrid IndoBERT-BiGRU model. Sentiment training targets are generated using the InSet lexicon-based pseudo-labeling procedure. Experimentally, the study compares the validation-selected hyperparameter configurations and their corresponding validation performance across the four training settings. The findings provide a descriptive account of configuration differences under the specified experimental protocol. They do not establish that domain-specific tuning is necessary, that dataset size determines model complexity, or that any selected configuration generalizes better across domains. Since the sentiment labels were generated automatically using InSet, the reported validation results also do not independently establish agreement with human sentiment judgments.

DOI : https://doi.org/10.21063/jtif.2026.V14.2.139-148
Creative Commons Attribution-Share Alike 4.0 International License (CC BY-SA 4.0)

2. METHODS AND DATA

This study formed part of a broader effort toward a sentiment analysis approach capable of accommodating multiple topical domains within Indonesian-language public opinion text. As a preparatory stage toward that aim, the present study concerns only whether the hyperparameters of the underlying hybrid classification model differ across domains, prior to any consideration of how such differences might later be accommodated architecturally.

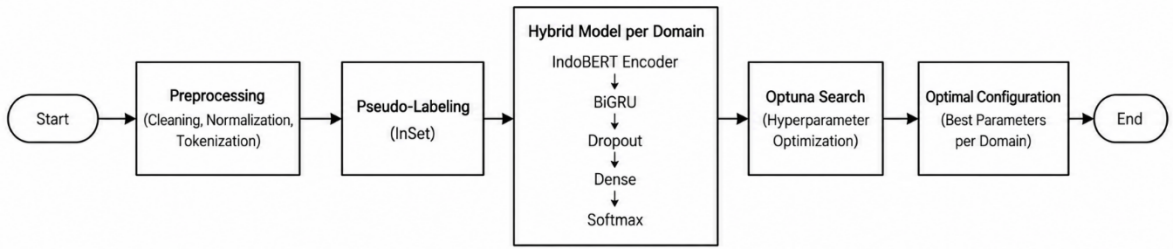

Figure 1. Research Procedure for Domain-specific Hyperparameter Optimization

2.1. Data and Preprocessing

Three Indonesian sentiment datasets were obtained from Kaggle: MyPertamina application reviews [23], public opinions concerning Shin Tae-yong [24], and opinions about *Merah Putih: One for All* [25]. After cleaning, the datasets were combined into the Global dataset.

Table 1. Summary of Domain Datasets

Domain Dataset	Data Volume	Domain Characteristic
MyPertamina application reviews	20,000 raw / 13,266 final	Public digital service
Shin Tae-yong public opinion	1,434 raw / 1,181 final	Sports and social discourse
Merah Putih: One for All film opinion	109 raw / 79 final	Culture and media

Data cleaning was performed in two stages. First, manual inspection was conducted to remove non-informative text, symbol-only content, and corrupted or unusable entries. Second, programmatic checks were applied to identify duplicate and empty-text records. After these procedures, the combined datasets were reduced from 21,543 to 14,526 records. Two text representations were subsequently prepared for different purposes. A conservatively cleaned representation was used as input to IndoBERT to preserve contextual and lexical information, whereas a normalized representation was prepared for InSet matching using stemming and stopword removal [26], [27]. The two representations were kept separate to prevent the normalization required for lexicon matching from unnecessarily altering the contextual input to IndoBERT.

2.2. Sentiment Pseudo-Labeling Using InSet

Because no manually annotated sentiment labels were available for model training, sentiment pseudo-labels were generated automatically using the Indonesian Sentiment Lexicon (InSet), a lexicon developed specifically for Indonesian-language sentiment analysis on microblog text [28]. Each document in the lexicon-oriented text representation was scored by matching tokens against the InSet dictionary and accumulating positive and negative polarity weights. The final score was calculated as the accumulated positive weight minus the accumulated negative weight. Documents with scores greater than zero were assigned positive labels, those with scores less than zero were assigned negative labels, and those with scores equal to zero were assigned neutral labels. Similar automated lexicon curation approaches for Indonesian sentiment analysis have been reported as practical alternatives to manual annotation at scale [29]. These pseudo-labels served as the training targets for the hybrid classification model described in the following subsection. The resulting pseudo-label distributions across the three source-defined datasets are presented in Table 2.

DOI : https://doi.org/10.21063/jtif.2026.V14.2.139-148
Creative Commons Attribution-Share Alike 4.0 International License (CC BY-SA 4.0)

Table 2. Distribution of InSet-Based Pseudo-Labels Across Datasets

Dataset	Negative (Train:Validation:Test)	Neutral (Train:Validation:Test)	Positive (Train:Validation:Test)	Total
MyPertamina	6,088 (4,870:609:609)	1,564 (1,251:156:157)	5,614 (4,491:562:561)	13,266
Shin Tae-yong	374 (299:37:38)	121 (97:12:12)	686 (549:69:68)	1,181
One For All	28 (22:3:3)	20 (16:2:2)	31 (25:3:3)	79
Global	6,490 (5,191:649:650)	1,705 (1,364:170:171)	6,331 (5,065:634:632)	14,526

The data were divided into training, validation, and test subsets using shuffled stratified sampling based on domain and pseudo-label category, with a random seed of 42 and an approximate 80:10:10 ratio. Table 2 summarizes the total pseudo-label counts and their distribution across the three subsets for each domain and the aggregated Global dataset.

Although lexicon-based pseudo-labeling provides a practical means of generating sentiment labels at scale, the resulting labels may contain noise that can affect model training and evaluation [30]. In this study, InSet-derived pseudo-labels were used as the sentiment targets for model training and as the reference labels for validation and held-out test evaluation. Validation Macro F1 was used as the optimization objective for selecting hyperparameter configurations. To assess agreement between automatically generated InSet pseudo-labels and Human_Sentiment labels, Cohen's kappa was calculated across all 14,526 records. Human_Sentiment was used exclusively for this agreement analysis and was not used as a training target, validation reference, hyperparameter-selection criterion, or test reference label. The agreement result is reported in Section 3.3.

2.3. Hybrid IndoBERT-BiGRU Architecture

The classification model used the pretrained IndoBERT checkpoint indobenchmark/indobert-base-p1 [8]. The corresponding tokenizer was loaded from the same checkpoint, with a vocabulary of 30,521 tokens and a maximum sequence length of 128 tokens. Input sequences were truncated when they exceeded the maximum length and padded to a fixed length. Attention masks were provided to distinguish valid tokens from padding tokens. The contextual token representations produced by IndoBERT were passed to the BiGRU layers to capture sequential dependencies in both forward and backward directions, followed by a dropout layer and a fully connected classification layer producing probabilities for three sentiment classes: positive, negative, and neutral. This general hybrid design follows an architectural pattern that has been reported to perform competitively on Indonesian-language sequence classification tasks [10], [16], [30]. An IndoBERT-BiGRU combination has also been applied to Indonesian-language entity extraction, although that task differs from sentiment classification [31].

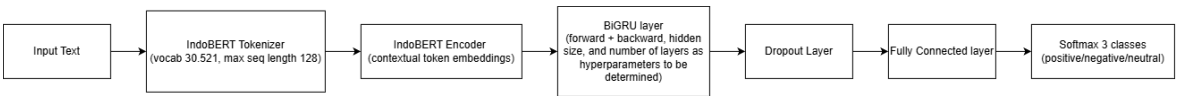

Figure 2. Hybrid IndoBERT-BiGRU Architecture

Four instances of this architecture were trained independently on the pseudo-labeled data: one aggregated (global) model trained on the combined dataset across all three domains, and three domain-specific models, each trained only on the data belonging to its respective domain. This design allowed the hyperparameter search described in the following subsection to be conducted separately for each domain, so that differences in optimal configuration could be observed directly.

2.4. Hyperparameter Optimization Using Optuna

Hyperparameter optimization was performed independently for the global model and for each of the three domain-specific models using Optuna, a hyperparameter optimization framework that employs

DOI : https://doi.org/10.21063/jtif.2026.V14.2.139-148
Creative Commons Attribution-Share Alike 4.0 International License (CC BY-SA 4.0)

sequential model-based search [17]. The default sampler used by Optuna, the tree-structured Parzen estimator (TPE), was retained without modification. Each model underwent a maximum of 20 optimization trials, with each trial trained for a maximum of 10 epochs. An early stopping mechanism with a patience of three epochs was applied, halting training for a given trial if the validation Macro F1-score did not improve for three consecutive epochs. Across the four models, the overall optimization process therefore comprised a maximum of 80 trials.

Three hyperparameters were treated as open search parameters: BiGRU hidden size, with candidate values of 128, 256, and 384; the number of BiGRU layers, either one or two; and the dropout rate, searched within the range of 0.20 to 0.50. Learning rate 0.00002, weight decay (0.01), and batch size (32) were held fixed across all trials to isolate the effect of the architecture-related hyperparameters. Model parameters were optimized using AdamW with cross-entropy loss. A linear learning-rate scheduler with warmup and gradient clipping were also used during training. The Macro F1-score computed on the validation subset of each respective training setting was used as the objective function to be maximized, since this metric assigns equal weight to each sentiment class regardless of class frequency, an important property given the class imbalance observed within several of the datasets.

To support reproducibility, a fixed random seed of 42 was used for Python, NumPy, PyTorch, dataset splitting, and the Optuna TPE sampler. Dataset splitting employed shuffled stratified sampling based on the combination of domain and pseudo-label category. The test subsets were not used during hyperparameter optimization or model selection and were retained for independent evaluation.

2.5. Implementation Environment

The experiments were implemented in Python using the PyTorch deep learning framework and the Hugging Face Transformers library. The computational environment recorded in the research notebook used Python 3.12.13, PyTorch 2.11.0+cu128, Transformers 5.13.1, Optuna 4.9.0, NumPy 2.0.2, Pandas 2.2.2, and scikit-learn 1.6.1. The recorded execution environment used an NVIDIA Tesla T4 GPU. These software and hardware details are reported to facilitate reproducibility of the experimental procedure.

3. RESULTS AND DISCUSSION

3.1. Best Validation-Selected Hyperparameter Configuration

Table 3 summarizes the validation-selected hyperparameter configurations obtained through independent Optuna optimization for the Global model and the three domain-specific models.

Table 3. Best Validation-Selected Hyperparameter Configurations and Validation Macro F1

Model	Hidden Size	BiGRU Layer	Dropout	Val. Loss	Accuracy	Precision	Recall	Macro F1
Global	384	2	0.2465	0.4108	86.51%	85.60%	78.85%	81.25%
MyPertamina	384	2	0.2465	0.4352	86.81%	84.05%	80.78%	82.16%
Shin Tae-yong	256	1	0.2301	1.4226	65.25%	69.44%	58.70%	61.40%
One For All	128	1	0.3543	1.0492	62.50%	66.67%	66.67%	65.71%

The results in Table 3 show that the validation-selected configuration was not uniform across the four training settings. The Global and MyPertamina models selected an identical configuration consisting of a hidden size of 384, two BiGRU layers, and a dropout rate of 0.2465, with validation Macro F1-scores of 81.25% and 82.16%, respectively. In contrast, the Shin Tae-yong model selected a smaller configuration with a hidden size of 256, one BiGRU layers, and a dropout rate of 0.2301, achieving a validation Macro F1-score of 61.40%. The One For All model selected the smallest hidden size of 128 with one BiGRU layers and the highest dropout rate of 0.3543, achieving a validation Macro F1-score

DOI : https://doi.org/10.21063/jtif.2026.V14.2.139-148
Creative Commons Attribution-Share Alike 4.0 International License (CC BY-SA 4.0)

of 65.71%.Although the selected configurations show different hidden sizes across training settings, these differences should not be interpreted as evidence that dataset size determines model complexity. The differences in selected configurations provide a descriptive answer to RQ1: under the same predefined search space and validation objective, independent hyperparameter optimization produced different validation-selected configurations across the four training settings. The differences were particularly evident in BiGRU hidden size and the number of recurrent layers. However, the results do not establish that dataset size alone caused these differences. The datasets also differ in domain characteristics, class distributions, and potential pseudo-label noise, which cannot be separated from dataset size within the present experimental design.

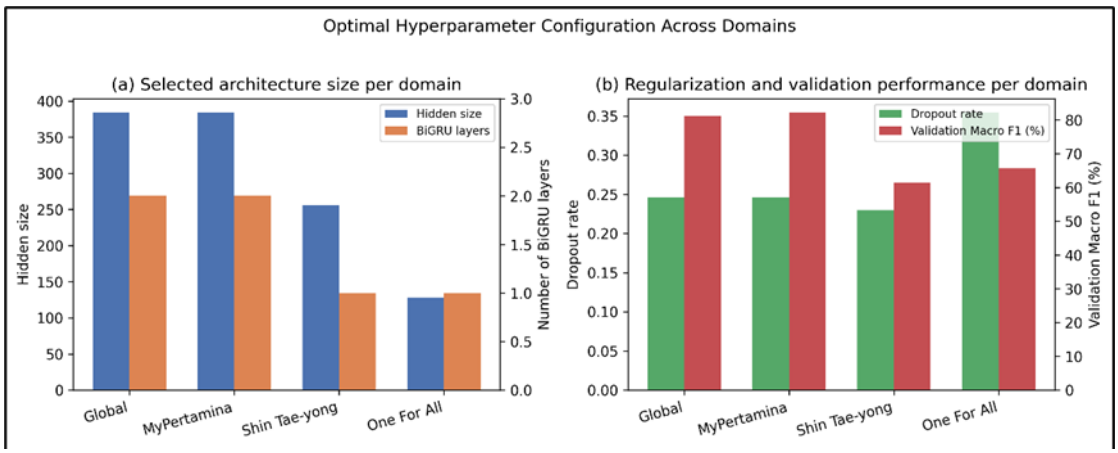

Figure 3. Domain Hyperparameter Comparison

The largest source dataset, MyPertamina, retained 13,266 records and selected the largest hidden size and two-layer BiGRU configuration available in the search space. The Shin Tae-yong and One For All datasets retained 1,181 and 79 records, respectively, and selected smaller hidden sizes with one BiGRU layers. Dropout, however, did not increase uniformly as dataset size decreased: Shin Tae-yong selected 0.2301, below the 0.2465 selected by Global and MyPertamina, while One For All selected 0.3543. Thus, the results support only a descriptive observation that the selected configurations differed; they do not establish that smaller datasets necessarily require stronger regularization. Broader work has examined relationships between training-set size and deep-learning model scale [32], but that literature does not establish the cause of the pattern observed here.

3.2. Independent Test Evaluation Against InSet-Derived Pseudo-Labels

To complement the validation-based optimization results, each selected model was evaluated on its corresponding held-out test set using InSet-derived pseudo-labels as reference labels. Therefore, the reported test metrics measure the models' agreement with the InSet-based labeling procedure rather than performance against independently human-annotated sentiment labels. The Global model was evaluated on the Global test set, while each domain-specific model was evaluated on the test set of its respective domain. The test subsets were not used during hyperparameter optimization or model selection and were retained for independent evaluation.

Table 4. Independent Test Performance of the Selected Models Against InSet-Derived Pseudo-Labels

Model	Test Samples	Accuracy	Macro Precision	Macro Recall	Macro F1
Global	1,453	64.42%	48.92%	54.47%	49.19%
MyPertamina	1,327	62.92%	47.79%	55.40%	46.86%
Shin Tae-yong	118	27.12%	61.22%	40.05%	22.71%

DOI : https://doi.org/10.21063/jtif.2026.V14.2.139-148
Creative Commons Attribution-Share Alike 4.0 International License (CC BY-SA 4.0)

Model	Test Samples	Accuracy	Macro Precision	Macro Recall	Macro F1
One For All	8	37.50%	24.44%	38.89%	30.00%

Table 4 shows the performance of the validation-selected models on their corresponding held-out test sets. Because both validation and test evaluation use InSet-derived pseudo-labels as reference labels, the resulting metrics provide a consistent basis for examining model performance across the two data subsets. However, the test results should be interpreted as measures of agreement with the InSet-based labeling procedure rather than as direct measures of agreement with human sentiment judgments. The Global model obtained the highest independent-test Macro F1-score at 49.19%, followed by MyPertamina at 46.86%, One For All at 30.00%, and Shin Tae-yong at 22.71%. The decrease between validation and independent-test performance indicates that validation performance should not be interpreted as a direct estimate of performance on unseen data.

The results should also be interpreted in light of the differences in dataset size, domain characteristics, and class distributions across the four training settings. In particular, the One For All test set contains only eight records, making its evaluation metrics highly sensitive to individual predictions. Consequently, its test results should be treated as descriptive rather than as a stable estimate of generalization performance.

The independent test evaluation complements the validation-based analysis by examining the selected configurations on held-out data that were not used for optimization or model selection. Nevertheless, these results do not establish that domain-specific hyperparameter optimization is superior to optimization on the aggregated dataset, because the four models were trained and evaluated under different dataset conditions rather than through a controlled comparison using identical test data.

3.3. Per-Class Performance on the Independent Test Sets

To examine class-specific behavior that may be obscured by Macro F1, Precision, Recall, and F1 were calculated separately for the Negative, Neutral, and Positive classes.

Table 5. Per-Class F1 Scores on the Independent Test Sets

Model	Negative F1	Neutral F1	Positive F1
Global	77.59%	20.53%	49.47%
MyPertamina	74.54%	17.39%	48.66%
Shin Tae-yong	28.07%	3.85%	36.22%
One For All	40.00%	0.00%	50.00%

The per-class results show that the Negative class achieved the highest F1-score for both the Global and MyPertamina models, at 77.59% and 74.54%, respectively. The Positive class achieved F1-scores of 49.47% for Global, 48.66% for MyPertamina, 36.22% for Shin Tae-yong, and 50.00% for One For All. In contrast, the Neutral class consistently exhibited substantially lower F1-scores. The Neutral F1-score was 20.53% for Global, 17.39% for MyPertamina, 3.85% for Shin Tae-yong, and 0.00% for One For All. These results indicate that the models had greater difficulty classifying Neutral instances than the other sentiment classes. The low Neutral-class F1-scores contributed to the relatively low Macro F1-scores observed on the independent test sets.

The class-level results should nevertheless be interpreted in relation to the available support in the corresponding test sets. Based on the InSet-derived pseudo-labels, the Global test set contained 171 Neutral samples, MyPertamina contained 157, Shin Tae-yong contained 12, and One For All contained only two. In particular, the One For All class-level metrics are highly sensitive to individual predictions because its entire test set contains only eight observations. Thus, the per-class results provide diagnostic

DOI : https://doi.org/10.21063/jtif.2026.V14.2.139-148
Creative Commons Attribution-Share Alike 4.0 International License (CC BY-SA 4.0)

information about model behavior rather than statistically stable estimates of class-specific generalization.

In addition to the InSet-based pseudo-labels, the dataset contained a separate Human_Sentiment label representing manually assigned sentiment categories. These labels were kept separate from the InSet labeling procedure and were not used as training targets, validation references, hyperparameter-selection criteria, or test reference labels. Human_Sentiment was used exclusively to assess the agreement between the InSet-generated pseudo-labels and manually assigned sentiment labels using Cohen's kappa across all 14,526 records.

The agreement between InSet-generated pseudo-labels and manually assigned Human_Sentiment labels was evaluated across 14,526 records, yielding a Cohen's kappa of 0.2511, which indicates a relatively low level of agreement between the two labeling methods. This assessment was performed independently of model training, hyperparameter tuning, and test-set evaluation. The findings suggest notable discrepancies between the InSet-based pseudo-labels and the manually assigned sentiment labels. Nevertheless, the kappa value alone is insufficient to determine whether the Human_Sentiment labels are entirely accurate or free from labeling errors.

4. CONCLUSION

This study investigated hyperparameter optimization of a hybrid IndoBERT-BiGRU model across four training settings. For RQ1, independent Optuna optimization selected different configurations: Global and MyPertamina used 384 hidden units, two BiGRU layers, and a dropout rate of 0.2465; Shin Tae-yong used 256 hidden units, one layer, and a dropout rate of 0.2301; and One For All used 128 hidden units, one layer, and a dropout rate of 0.3543. For RQ2, the main differences were observed in hidden size and the number of BiGRU layers, with validation Macro F1-scores of 81.25%, 82.16%, 61.40%, and 65.71%, respectively. The held-out test Macro F1-scores against the InSet-derived pseudo-labels were 49.19% for Global, 46.86% for MyPertamina, 22.71% for Shin Tae-yong, and 30.00% for One For All . Per-class evaluation showed that the Neutral class remained difficult to classify. The separate agreement analysis between InSet and Human_Sentiment across 14,526 records yielded a Cohen's kappa of 0.2511, indicating limited agreement between the two labeling approaches and highlighting limitations in the pseudo-labeling procedure. Because both validation and independent testing used InSet-derived pseudo-labels as reference labels, the observed difference between validation and test Macro F1-scores should be interpreted with caution, as these metrics reflect agreement with the same pseudo-labeling procedure rather than independently annotated human sentiment labels. Overall, the findings demonstrate that independent hyperparameter optimization can produce different validation-selected configurations across training settings. However, they do not establish that domain-specific tuning is superior or necessary, nor that dataset size determines model complexity. A controlled comparison of Global and domain-specific configurations on the same domain-specific test sets was not performed and remains future work. Further research should also examine the robustness of the results through repeated experiments and uncertainty analysis, while documenting and strengthening the human-annotation procedure. The One For All results should be interpreted cautiously because its independent test set contains only eight records.

ACKNOWLEDGEMENTS

The authors would like to thank Kemdiktisaintek for the financial support provided for this research through the Penelitian Dosen Pemula (PDP) research grant scheme.

REFERENCES

[1] M. He, C. Ma, and R. Wang, "A Data-Driven Approach for University Public Opinion Analysis and Its Applications," *Applied Sciences*, vol. 12, no. 18, 2022, doi: https://doi.org/10.3390/app12189136.

Creative Commons Attribution-Share Alike 4.0 International License (CC BY-SA 4.0)

[2] T. Anderson, S. Sarkar, and R. Kelley, "Analyzing Public Sentiment on Sustainability: A Comprehensive Review and Application of Sentiment Analysis Techniques," *Natural Language Processing Journal*, vol. 8, p. 100097, 2024, doi: https://doi.org/10.1016/j.nlp.2024.100097.

[3] S. J. Pipin and H. Kurniawan, "Analisis Sentimen Kebijakan MBKM Berdasarkan Opini Masyarakat di Twitter Menggunakan LSTM," *Jurnal SIFO Mikroskil*, vol. 23, no. 2, pp. 197–208, Oct. 2022, doi: https://doi.org/10.55601/jsm.v23i2.900.

[4] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A Survey on Sentiment Analysis Methods, Applications, and Challenges," *Artif. Intell. Rev.*, vol. 55, no. 7, pp. 5731–5780, 2022, doi: https://doi.org/10.1007/s10462-022-10144-1.

[5] J. Cui, Z. Wang, S.-B. Ho, and E. Cambria, "Survey on Sentiment Analysis: Evolution of Research Methods and Topics," *Artif. Intell. Rev.*, vol. 56, no. 8, pp. 8469–8510, 2023, doi: https://doi.org/10.1007/s10462-022-10386-z.

[6] M. El Barachi, M. AlKhatib, S. Mathew, and F. Oroumchian, "A Novel Sentiment Analysis Framework for Monitoring the Evolving Public Opinion in Real-Time: Case Study on Climate Change," *J. Clean. Prod.*, vol. 312, p. 127820, 2021, doi: https://doi.org/10.1016/j.jclepro.2021.127820.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *North American Chapter of the Association for Computational Linguistics*, vol. 1, pp. 4171–4186, 2019, doi: 10.18653/v1/N19-1423.

[8] 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*, D. Scott, N. Bel, and C. Zong, Eds., Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 757–770. doi: https://doi.org/10.18653/v1/2020.coling-main.66.

[9] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, A. Zaenen and A. den Bosch, Eds., Prague, Czech Republic: Association for Computational Linguistics, Jun. 2007, pp. 440–447. [Online]. Available: https://aclanthology.org/P07-1056/

[10] Q. Zhu, X. Jiang, and R. Ye, "Sentiment Analysis of Review Text Based on BiGRU-Attention and Hybrid CNN," *IEEE Access*, vol. 9, pp. 149077–149088, 2021, doi: https://doi.org/10.1109/ACCESS.2021.3118537.

[11] G. Kaur and A. Sharma, "HAS: Hybrid Analysis of Sentiments for the Perspective of Customer Review Summarization," *J. Ambient Intell. Humaniz. Comput.*, vol. 14, no. 9, pp. 11971–11984, 2023, doi: https://doi.org/10.1007/s12652-022-03748-6.

[12] H. Ahmadian, T. F. Abidin, H. Riza, and K. Muchtar, "Hybrid Models for Emotion Classification and Sentiment Analysis in Indonesian Language," *Applied Computational Intelligence and Soft Computing*, vol. 2024, pp. 1–17, 2024, doi: https://doi.org/10.1155/2024/2826773.

[13] W. Zhang, X. Li, Y. Deng, L. Bing, and W. Lam, "A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges," *IEEE Transactions on Knowledge & Data Engineering* , vol. 35, no. 11, pp. 11019–11038, Nov. 2023, doi: https://doi.org/10.1109/TKDE.2022.3230975.

[14] Y. Fu and Y. Liu, "Contrastive transformer based domain adaptation for multi-source cross-domain sentiment classification," *Knowledge-Based Systems*, vol. 245, Art. no. 108649, 2022, doi: 10.1016/j.knosys.2022.108649.

[15] Y. Zhu, Y. Qiu, Q. Wu, F. L. Wang, and Y. Rao, "Topic Driven Adaptive Network for cross-domain sentiment classification," *Information Processing & Management*, vol. 60, no. 2, Art. no. 103230, 2023, doi: 10.1016/j.ipm.2022.103230.

[16] S. Z. Pranida and A. Kurniawardhani, "Sentiment Analysis of Expedition Customer Satisfaction using BiGRU and BiLSTM," *Indonesian Journal of Artificial Intelligence and Data Mining*, vol. 5, no. 1, pp. 44–53, Jun. 2022, doi: https://doi.org/10.24014/ijaidm.v5i1.17361.

[17] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, in KDD '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 2623–2631. doi: https://doi.org/10.1145/3292500.3330701.

[18] X. Liu and C. Wang, "An Empirical Study on Hyperparameter Optimization for Fine-Tuning Pre-trained Language Models," in *Proc. 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 2286–2300, doi: 10.18653/v1/2021.acl-long.178.

[19] D. Salinas, H. Shen, and V. Perrone, "A Quantile-based Approach for Hyperparameter Transfer Learning," in *Proc. 37th International Conference on Machine Learning*, Proc. Mach. Learn. Res., vol. 119, 2020, pp. 8438–8448, doi: https://doi.org/10.48550/arXiv.1909.13595

[20] M. R. Grootendorst, "BERTopic: Neural Topic Modeling With a Class-Based Tf-Idf Procedure," *ArXiv*, vol. abs/2203.05794, 2022, doi: https://doi.org/10.48550/arXiv.2203.05794.

DOI : https://doi.org/10.21063/jtif.2026.V14.2.139-148
Creative Commons Attribution-Share Alike 4.0 International License (CC BY-SA 4.0)

[21] M. Ali *et al.*, "Investigating Mental Wellbeing Self-Care in Higher Education Using BERTopic Modeling," *Discover Mental Health*, vol. 5, no. 1, p. 204, Nov. 2025, doi: https://doi.org/10.1007/s44192-025-00323-1.
[22] A. Rejeb, K. Rejeb, S. Simske, and E. Süle, "Industry 5.0 Research: An Approach Using Co-word Analysis and BERTopic Modeling," *Discover Sustainability*, vol. 6, no. 1, p. 402, 2025, doi: https://doi.org/10.1007/s43621-025-01252-3.
[23] sagitasantia, "Sentiment Analysis MyPertamina Application," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/sagitasantia/sentiment-analisis-mypertamina-application. Accessed: Feb. 18, 2026.
[24] wildanabid19, "Analisis Sentiment STY Raw," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/wildanabid19/analisis-sentiment-sty-raw. Accessed: Feb. 18, 2026.
[25] ldausl, "Merah Putih: One for All Sentiment," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/ldausl/merah-putih-one-for-all-sentiment. Accessed: Feb. 18, 2026.
[26] N. Garg and K. Sharma, "Text Pre-processing of Multilingual for Sentiment Analysis Based on Social Network Data," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 1, pp. 776–784, 2022, doi: https://doi.org/10.11591/ijece.v12i1.pp776-784.
[27] L. Hickman, S. Thapa, L. Tay, M. Cao, and P. Srinivasan, "Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations," *Organ. Res. Methods*, vol. 25, no. 1, pp. 114–146, 2022, doi: https://doi.org/10.1177/1094428120971683.
[28] F. Koto and G. Y. Rahmaningtyas, "Inset Lexicon: Evaluation of a Word List for Indonesian Sentiment Analysis in Microblogs," in *2017 International Conference on Asian Language Processing (IALP)*, 2017, pp. 391–394. doi: https://doi.org/10.1109/IALP.2017.8300625.
[29] R. Wijayanti and A. Arisal, "Automatic Indonesian Sentiment Lexicon Curation with Sentiment Valence Tuning for Social Media Sentiment Analysis," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 20, no. 1, Mar. 2021, doi: https://doi.org/10.1145/3425632.
[30] Y. Zhang, M. Jiang, Y. Meng, Y. Zhang, and J. Han, "PIEClass: Weakly-*Supervised Text Classification with Prompting and Noise-Robust Iterative Ensemble Training," in Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP),* 2023, pp. 12655–12670, doi: 10.18653/v1/2023.emnlp-main.780.
[31] E. Dave and A. Chowanda, "IPerFEX-2023: Indonesian Personal Financial Entity Extraction Using indoBERT-BiGRU-Crf Model," *J. Big Data,* vol. 11, no. 1, p. 139, 2024, doi: https://doi.org/10.1186/s40537-024-00987-6.
[32] J. Hestness *et al.*, "Deep Learning Scaling is Predictable, Empirically," *ArXiv*, vol. abs/1712.00409, 2017, doi: https://doi.org/10.48550/arXiv.1712.00409.

DOI : https://doi.org/10.21063/jtif.2026.V14.2.139-148
Creative Commons Attribution-Share Alike 4.0 International License (CC BY-SA 4.0)