Fine-Tuning Hybrid Deep Learning for Sentiment Analysis of Indonesian Product Reviews

Arwin Halim[1*], Roni Yunis[2], and Erlina Halim[3]
[1,3]Department of Informatics Engineering, Faculty of Informatics, Universitas Mikroskil Sumatera Utara, Indonesia 20212
[2]Department of Information Systems, Faculty of Informatics, Universitas Mikroskil Sumatera Utara, Indonesia 20212
Email: [1]arwin@mikroskil.ac.id, [2]roni@mikroskil.ac.id, [3]erlina.halim@mikroskil.ac.id

***Abstract*—The research aims to build a hybrid deep learning model for sentiment analysis of Indonesian e-commerce product reviews, which represent the expressed opinions of customers. A major challenge in the domain is the presence of non-standard language and highly imbalanced sentiment classes, which hinder accurate classification. Most existing Indonesian sentiment analysis studies rely on relatively small and balanced datasets and primarily use attention mechanisms, an ensemble model, as well as a sequential fusion method. In the research, a large-scale dataset of Indonesian product reviews is collected from the largest e-commerce site in the country. The dataset consists of review text and corresponding product ratings. After preprocessing, semantic features are extracted using a pre-trained Indonesia Bidirectional Encoder Representations from Transformers (IndoBERT) model. The features are then fed into a hybrid model combining Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) layers through parallel feature-level fusion. Model hyperparameters are optimized using the Tree-Structured Parzen Estimator (TPE), while data imbalance is addressed through resampling methods. Regularization strategies are also applied to mitigate overfitting, and the model is evaluated using stratified k-fold cross-validation. The model hyperparameters are validated using a learning curve, showing a stable and consistent curve following the trend. The results show that the hybrid CNN-LSTM model, combined with Support Vector Machine Synthetic Minority Oversampling Technique (SVMSMOTE), achieves superior performance in distinguishing positive and negative reviews. This outcome reaches Receiver Operating Characteristic - Area Under the Curve (ROC AUC) score of 92.48%, outperforming baseline and conventional machine learning models. These results also show good generalization ability, characterized by consistent values with a very low standard deviation of 0.0009 for each fold.**

***Index Terms*—Convolutional Neural Network (CNN), Hybrid Deep Learning, Long Short-Term Memory (LSTM), Resampling Method, Tree-structure Parzen Estimator (TPE)**

I. Introduction

ONLINE shopping is a major consumer activity alongside rapid advances in information technology. A typical example is e-commerce platforms, which offer a wide range of products at competitive prices, enabling customers to easily compare options and read reviews from other users on the desired product before making a purchase. These reviews provide valuable feedback and subjective opinions about the purchased product [1], and are reported to influence 73%–87% of purchase decisions among potential buyers [2]. Consequently, customer-generated reviews have driven the development of automated sentiment analysis methods [3]. In real-world applications, review datasets are often highly imbalanced. Popular or high-quality products typically receive a large volume of positive reviews, while low-quality goods generate fewer, predominantly negative reviews. This imbalance creates a significant challenge for sentiment analysis systems. Moreover, sentiment analysis should be capable of processing large-scale, unstructured text that contains diverse and non-standard grammar, increasing the complexity of the task [4].

A variety of studies have been conducted on sentiment analysis in the Indonesian language across many different application domains. These analyses use several methods, including lexicon-based strategies, machine learning, and deep learning [5]. In the e-commerce domain, sentiment analysis studies are dominated by machine learning (48.1%), deep learning (44.4%), and hybrid (7.4%) methods with commonly used data sources such as Amazon and Twitter [1]. However, studies specifically focusing on

Received: June 20, 2025; received in revised form: Oct. 24, 2025; accepted: Oct. 24, 2025; available online: April 02, 2026.
*Corresponding Author

the analysis in the Indonesian language remain relatively limited [5]. For example, a previous study has evaluated consumer opinions on the Shopee e-commerce platform using a lexicon-based method [6]. It applies the SentiStrength algorithm to determine sentiment polarity and measure the intensity of expressed opinions. Another study has proposed a sentiment analysis model based on Naïve Bayes combined with Particle Swarm Optimization for tweet classification in the Indonesian language [7]. Following the reviews, a deep learning method is reported to achieve superior performance in sentiment classification and rating prediction tasks using a large Indonesian product review dataset [8]. The effectiveness of a hybrid method combining IndoBERT and LSTM for sentiment analysis in Indonesian-language skincare reviews is also shown in previous research [9]. Similarly, good accuracy with a hybrid model incorporating a customized IndoBERT model with BiLSTM, BiGRU, and attention mechanism layers is achieved [10]. Another research has proposed an ensemble framework by combining bagging, boosting, and multinomial Naive Bayes for multi-level sentiment analysis [11]. Previous studies have also explored a sequential fusion method using deep learning models for sentiment analysis in both English and Indonesian texts [8–11].

Several studies have explored deep learning methods for sentiment analysis across different languages and dataset conditions. For example, Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) models are applied to Indonesian hotel reviews, achieving high performance on 2,500 class-balanced samples [12]. Then, a hybrid deep learning model is evaluated on four imbalanced English datasets and reported good accuracy [13]. Another research has combined RoBERTa with Bi-LSTM, improving sentiment analysis accuracy across various large English-language datasets with balanced classes [14]. Addressing the challenge of imbalanced classes in Indonesian texts, previous researchers have applied a hybrid model with an under-sampling method to improve performance on 14,975 reviews [15].

Based on the discussion mentioned earlier, study gaps concerning the issue remain. In particluar, relatively few studies evaluate hybrid deep learning models for sentiment analysis of Indonesian-language product reviews using large-scale data and imbalanced datasets. The research aims to develop a sentiment analysis model for large-scale Indonesian-language data, with a focus on product reviews. The dataset is obtained from online reviews on Shopee, an e-commerce platform that has surfaced as the most visited online shopping site in Indonesia by 2024 [16]. The proposed hybrid model incorporates the IndoBERT pre-trained language model, CNN, stacked LSTM, and the resampling methods. This model uses parallel feature-level fusion, allowing more comprehensive incorporation of complementary information. In addition, the model is evaluated on a large-scale Indonesian e-commerce dataset that has been resampled through a combination of over-under sampling method. The sampling strategy proposed in previous research [17] that has effectively addressed the class imbalance problem. Improving data diversity can increase variability in the dataset and advance the generalizability of the model, allowing a more comprehensive evaluation of hybrid deep learning methods in sentiment analysis tasks. The hyperparameter tuning process and regularization method are also applied to ensure optimal model performance and prevent overfitting [18]. The research focuses on developing a model capable of accurately detecting both positive and negative sentiment classes. The main contribution of this research is to provide a large Indonesian language dataset for product reviews. Additionally, the research proposes a combination of reliable and consistent hybrid deep learning models in analyzing Indonesian language sentiment from product reviews with imbalanced classes.

II. Research Method

Figure 1 shows the proposed deep learning framework. It incorporates a pre-trained language model, a hybrid CNN-LSTM architecture, and a resampling method for sentiment analysis on Indonesian language reviews. The proposed framework starts with a data preprocessing phase, followed by the transformation of inputs into embedding vectors. These features are subsequently processed in parallel through a hybrid architecture consisting of CNN and LSTM layers. To enhance model generalizability, hyperparameter optimization and resampling techniques are integrated. Finally, a comprehensive evaluation is conducted to identify the most optimal hybrid configuration.

A. Dataset

The dataset consists of product reviews obtained from Shopee Indonesia. The review includes textual comments posted on the platform as well as numerical ratings given by buyers. Data collection is restricted to the most popular product categories on Shopee, namely fashion and beauty [19]. In each category, products containing customer comments section are identified. These sections include customer-related information, comment text, and corresponding rating. The data crawling process is conducted repeatedly across multiple products.

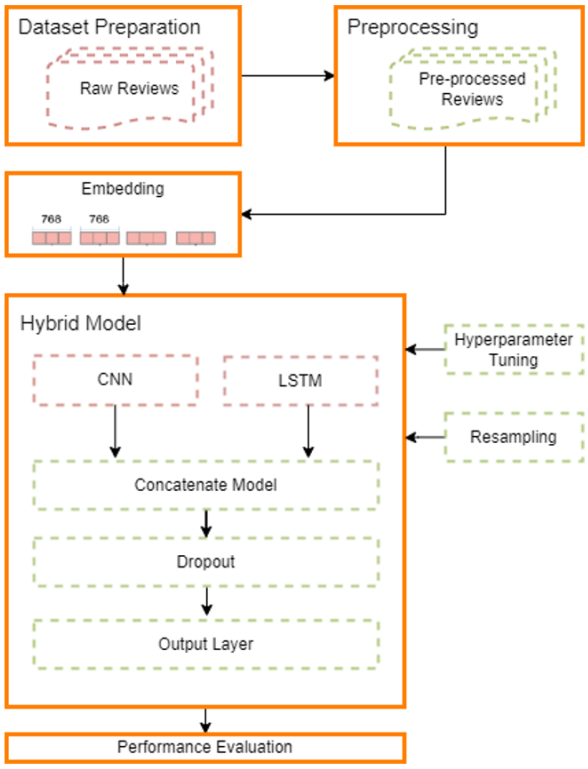

Fig. 1. Proposed hybrid deep learning model. Note: Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN).

B. Preprocessing

Preprocessing is a crucial step for text classification accuracy in the research [20, 21]. The process transforms the raw dataset into a structured format, which is suitable as input for the classification model. The preprocessing stages include case folding, tokenization, stopword removal, stemming, and duplicate [22]. During case folding, the comment text is converted to lowercase to ensure uniform representation. The next process is tokenization, which separates texts into individual words or tokens. Stopword removal eliminates meaningless words that do not contribute significant semantic meaning. Stemming reduces words by removing prefixes and suffixes to obtain root forms. The application of stemming has been commonly used in sentiment analysis studies, such as the evaluation of Twitter user sentiment in Lazada and Shopee services [23]. Other studies have included Indonesian public opinion on sentiment analysis [15], sentiment analysis of consumer reviews, film assessments, and social media with a combination of deep learning algorithms [13], and analysis of Virtual Reality (VR) social community in user evaluations using a combination of feature extraction [24]. In the research, the MPStemmer algorithm is used for stemming [25]. The final process includes duplicate removal, where comments containing identical words and ratings are deleted.

C. Embedding

Feature extraction is conducted at this stage from words to obtain vector values, which are often called word embedding. The analysis uses a derivative of the Bidirectional Encoder Representations from Transformers (BERT) method [26]. It is capable of capturing the semantic meaning of words, reducing dimensions, adding contextual information, and inspiring efficient learning by transferring linguistic knowledge through previously trained embeddings [27]. The word embedding process for Indonesian comments uses the BERT method specifically for citizens of the country, namely the IndoBERT method [28]. In addition, the output of IndoBERT is a vector of 768 data points for each comment processed [29].

D. Hybrid Deep Learning

CNN is a neural network that is often used for image processing. However, the CNN model has proven effective in the field of text classification [30]. The architecture of this model consists of three different layers, namely convolutional, pooling, and flattening. The pooling layer reduces computational complexity, with max pooling being the primary method. The flattened layer receives the output from the pooling layer and prepares it for input into the next layer of the model [19]. In the domain of text processing, a word is typically represented as a complete vector. The width of the convolutional kernel in the convolutional layer generally corresponds to the dimensionality of this word vector. For the input matrix $V = [v'_1, v'_2, \ldots, v'_i, \ldots, v'_n]$, the convolution operation is in Eq. (1) [31].

$$v_i^n = f(W \cdot V_{[i:i+k-1]} + b). \tag{1}$$

where $W \in R^{k*m}$ signifies a weight matrix, k and m represent the height and width of the convolution kernel, b shows the offsets, and f represents the activation function. The Rectified Linear Unit (ReLu) activation function is applied to all layers, except the final layer, which uses a SoftMax activation. After completing the convolution operation, the eigenvector matrix V' is represented in Eq. (2) [20].

$$v' = [v''_1, v''_2, \ldots, v''_i, \ldots, v''_{n-k+1}]. \tag{2}$$

LSTM networks are initially introduced by Hochreiter and Schmidhuber, providing a model capable of capturing long-term dependencies in sequential

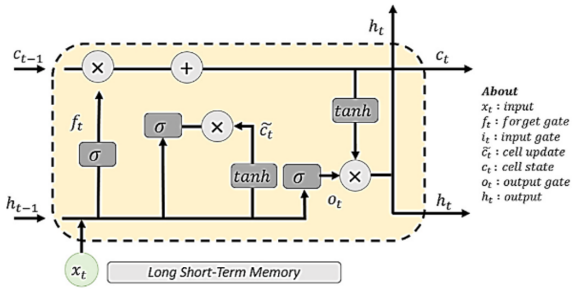

Fig. 2. Long Short-Term Memory (LSTM) unit [33].

data [32], as shown in Fig. 2. Different from conventional recurrent components, the LSTM unit retains the current memory $c_t \in R^n$ at time t. The input at time t is x_t, h_{t-1}, c_{t-1}, and the output is h_t, c_t, which can be updated by Eqs. (3)–(8).

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i), \quad (3)$$
$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f), \quad (4)$$
$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o), \quad (5)$$
$$g_t = \tanh(W_g x_t + U_g h_{t-1} + b_g), \quad (6)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot g_t, \quad (7)$$
$$h_t = o_t \odot \tanh(c_t). \quad (8)$$

where $\sigma(.)$ represents the logistic sigmoid function, and the operation $\odot$ signifies the element-wise vector product. At every time step t, there exists an input gate i_t, a forget gate f_t, an output gate o_t, a memory cell c_t and a hidden unit h_t. The h_0 and c_0 can be set to 0, and the parameters of the LSTM are W, U, and b. The research tests the use of stacked LSTM and several LSTM layers in sequence, where the output of one LSTM layer becomes the input for the next layer.

E. Hyperparameter Tuning and Resampling

Determining the best hyperparameters is needed to ensure a robust and generalizable deep learning model with the best performance. Many studies [34–36] have showed that the Tree-structured Parzen Estimator (TPE) algorithm, a Bayesian hyperparameter optimization method, is capable of improving the accuracy of machine learning and deep learning models. In the research, TPE is used to obtain the optimal hyperparameter values in the CNN and LSTM hybrid model.

The existence of imbalanced datasets profoundly influences the performance of learning models. This skewed distribution leads to biased model training, causing poor predictive performance. Addressing the challenges that arise from imbalanced datasets to improve both the reliability and the accuracy of the results obtained from learning models is essential [37]. Several resampling methods have been developed recently to handle the class imbalance. In the research, the resampling methods used are SMOTE, BorderlineSMOTE, Support Vector Machine Synthetic Minority Oversampling Technique (SVMSMOTE), Adaptive Synthetic Sampling (ADASYN), Synthetic Minority Oversampling Technique Ensemble Neural Networks (SMOTENN), and Synthetic Minority Over-sampling Technique and Tomek Links (SMOTETomek).

TABLE I
EVALUATION METRIC IN THE RESEARCH.

Evaluation	Formula
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
Specificity	$\frac{TN}{TN+FP}$
F1-Score	$2 \times \frac{Precision \times Recall}{Precision + Recall}$
ROC AUC Score	N/A

Note: Receiver Operating Characteristic - Area Under the Curve (ROC AUC), True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

F. Combination Layer

The combination layer (fully connected layer) combines the outputs generated from the CNN and LSTM levels to understand consumer reviews. During the process, CNN extracts local features from text and spatial patterns. Then, LSTM handles sequence as well as temporal dependencies.

G. Output Layer

The output layer is the classification stage using the sigmoid function to process the combined layers. The layer produces a probability distribution across all classes during the process. In this case, the sigmoid function is used for binary classification into positive or negative sentiment.

H. Evaluation

The performance of the learning model is assessed using evaluation metrics. A confusion matrix is used to quantify correct and incorrect predictions based on the actual values [38], including True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Additionally, the ROC value is calculated from the probability outputs to evaluate the overall performance of the model, with the corresponding formula shown in Table I.

TABLE II
THE ORIGINAL DATASET INFORMATION.

Rating	Reviews
1	14,287
2	10,165
3	31,999
4	47,111
5	325,438

TABLE III
THE ORIGINAL DATASET SENTIMENT AFTER CLEANSING PROCESS.

Sentiment	Reviews
Positive	299,733
Negative	50,244

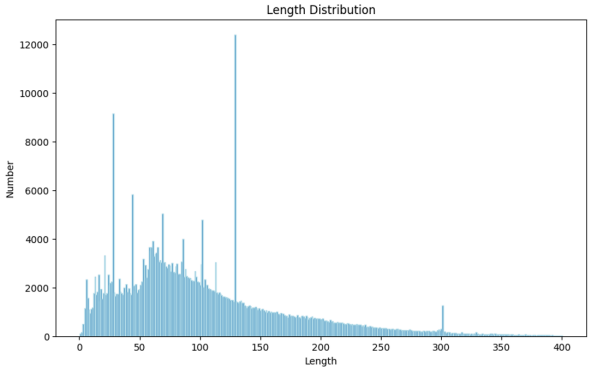

Fig. 3. Text length distribution in the dataset.

III. RESULTS AND DISCUSSION

A. Dataset Insight

Dataset is collected from the Shopee Indonesia platform, consisting of product reviews and ratings. Product reviews focus on the Indonesian language, and the total collected was 429,000 data points. The data are shown in Table II.

After the cleansing process, the original dataset comprises 349,977, which are grouped into two label categories. The sentiment labels of the dataset follow the predetermined criteria of the analysis. Reviews with a rating higher than three are categorized as positive sentiment. Those with a rating lower than three are categorized as negative sentiment. The details of the final dataset information are shown in Table III. This dataset has an imbalanced composition in the number of sentiment labels.

The characteristics of text reviews are shown in the form of a distribution of text lengths in Fig. 3. The collected dataset is processed in a pre-trained model before being used as input to the training model to convert review text into vectors. The pre-trained language model used is the BERT, specifically for Indonesian, namely indoBERT [17].

B. Experimental Results

The initial sentiment analysis model is formed by determining the appropriate hyperparameters for generalization. The dataset is split using a stratified method into training, validation, and testing with a ratio of 4:1:1. As a result, the training, validation, and testing sets have the same proportion for positive and negative class labels. TPE is applied to obtain the optimal hyperparameters with a total of 200 iterations, and the results are shown in Table IV.

The optimization process explores diverse search spaces for both CNN and LSTM layers to maximize model performance. For the CNN layer, the optimal parameter includes a kernel size of 3, tanh activation function, 32 filters, and a stride value of 8, which collectively enable effective feature extraction from the input sequences. The learning rate is optimized to 9.232776×10^{-5}, while dropout is set to 0.394 to prevent overfitting during training. However, the stacked LSTM model does not provide a significant influence in sentiment analysis on Indonesian text, as the hyperparameter tuning results show that the optimal performance is achieved with single layer containing 176 units, a recurrent dropout of 0.30119, and a batch size of 56. The model is trained for 10 epochs, which provide sufficient iterations for convergence without overfitting. These fine-tuned hyperparameters, selected from their respective search spaces through rigorous experimentation, demonstrate the importance of architecture-specific optimization in achieving superior performance for sentiment analysis tasks. The resulting configuration balances model complexity with generalization capability, ensuring robust performance on unseen data. Relating to this discussion, the hybrid deep learning model is obtained according to the optimal hyperparameter settings. The proposed model uses Adam optimizer, binary_crossentropy for loss calculation, and sigmoid in the classification layer.

Model performance is compared with baseline deep learning methods and other machine learning algorithms. Table V shows the comparison between the results of the proposed model and other models. The comparison is conducted by evaluating several machine learning methods, namely Decision Tree (DT), Gaussian Naive Bayes (GNB), Random Forest (RF), Stochastic Gradient Descent (SGD), and baseline deep learning methods namely CNN and LSTM. In general, the proposed model has slightly better F1-Sscore and Receiver Operating Characteristic - Area Under the Curve (ROC AUC) values compared to other learning

TABLE IV
THE HYPERPARAMETER TUNING RESULTS.

Layer	Hyper Parameter	Search Space	Optimal Value
-	Dropout	[0.2, 0.5]	0.39405608201176756
	Learning rate	[1e-5, 1e-3]	9.232776262466981e-05
CNN	Kernel size	[2, 5]	3
	Activation	['relu', 'tanh', 'swish', 'leakyrelu']	tanh
	Filters	[8, 128]	32
	Strides	[2, 32]	8
LSTM	Units	[32, 512]	176
	Layer number	[1, 3]	1
	Recurrent dropout	[0.2, 0.5]	0.3011980090275241
	Batch_size	[32, 128]	56
	Epochs	[5, 30]	10

TABLE V
THE LEARNING MODEL EVALUATION RESULTS.

Learning Model	Accuracy	Precision	Recall	Specificity	F1-Score	ROC AUC
Gaussian Naïve Bayes	77.470%	94.030%	78.690%	70.190%	85.680%	80.262%
Decision Tree	83.120%	90.660%	89.510%	44.980%	90.080%	67.245%
Random Forest	89.130%	90.110%	98.070%	35.760%	93.920%	89.093%
Convolutional Neural Network (CNN)	85.439%	86.122%	98.942%	4.896%	92.088%	75.772%
Long Short-Term Memory (LSTM)	90.888%	92.995%	96.640%	56.577%	94.782%	92.325%
Proposed Model (CNN-LSTM)	91.068%	93.272%	96.534%	58.468%	94.875%	92.429%

Note: Receiver Operating Characteristic - Area Under the Curve (ROC AUC).

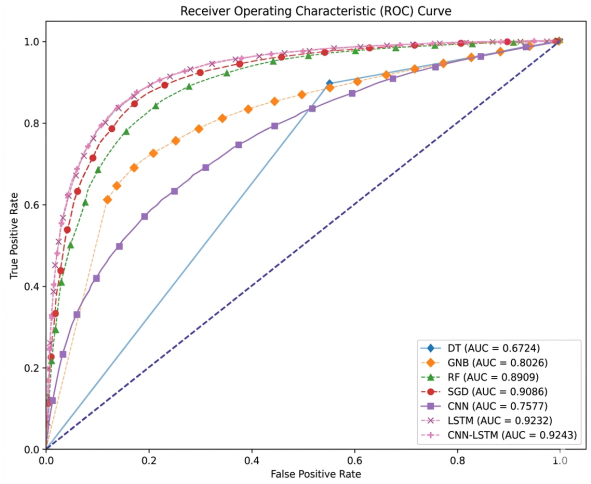

Fig. 4. Comparison of the ROC AUC score in the learning model.

models. F1-Score is the harmonic mean value of precision and recall, which shows the balance between positive and negative class predictions in this case. The proposed model also distinguishes positive and negative classes better, as shown by the higher ROC AUC value. Through the specificity value, the proposed model focuses extensively on predicting the majority class. It is reviewed by a TN value that is lower than the Naive Bayes model. It is capable of predicting the negative class with the best specificity value of 70.19%.

Figure 4 shows a comparison of the ROC AUC values in the learning model. The best ROC AUC evaluation results for machine learning models are obtained by the SGD model with 0.9086. Then, the deep learning model is obtained by the LSTM model with 0.9232. The lower results of the CNN model compared to many machine learning models indicate that it is not suitable if applied directly to natural language processing cases. The proposed model still has a slight advantage compared to other models, with 0.9243. The proposed model distinguishes positive and negative classes better, as shown by the higher ROC AUC value. These results indicate the superiority of the hybrid model compared to the standalone model.

Resampling methods are applied to the proposed model to generate new data for training the model. Table VI shows the results of the transformation of existing data and increasing the size of the training set. Resampling techniques, such as SMOTE, BorderlineSMOTE, SVMSMOTE, ADASYN, and SMOTETomek, successfully balance the negative and positive sentiment classes. While the negative class is augmented by approximately 180,000 synthetic samples to match the positive classes, SMOTENN yields different results. Due to outlier detection, SMOTENN reduces the overall data count by eliminating positive samples located near the negative decision boundary. Improving the diversity of the data contributes to a greater variability, augmenting the generalizability of the model. The resampling methods used consist of oversampling on the minority class and a combination of under- and

TABLE VI
THE DETAILS OF THE DATASET AFTER RESAMPLING.

Resampling Method	Original Training Set		After Resampling	
	Negative	Positive	Negative	Positive
Synthetic Minority Oversampling Technique (SMOTE)	40,195	239,786	239,786	239,786
BorderlineSMOTE	40,195	239,786	239,786	239,786
Support Vector Machine Synthetic Minority Oversampling Technique (SVMSMOTE)	40,195	239,786	239,786	239,786
Adaptive Synthetic Sampling (ADASYN)	40,195	239,786	240,879	239,786
Synthetic Minority Oversampling Technique Ensemble Neural Networks (SMOTENN)	40,195	239,786	238,693	122,893
Synthetic Minority Over-sampling Technique and Tomek Links (SMOTETomek)	40,195	239,786	239,760	239,760

TABLE VII
THE RESAMPLING EVALUATION RESULTS ON THE PROPOSED MODEL.

Learning Model	Accuracy	Precision	Recall	Specificity	F1-Score	ROC AUC
Proposed Model (CNN-LSTM)	91.068%	93.272%	96.534%	58.468%	94.875%	92.429%
Proposed Model (CNN-LSTM) + SMOTE	86.416%	96.569%	87.239%	81.512%	91.667%	91.904%
Proposed Model (CNN-LSTM) + BorderlineSMOTE	85.671%	96.780%	86.134%	82.905%	91.147%	91.851%
Proposed Model (CNN-LSTM) + SVMSMOTE	88.208%	96.286%	89.691%	79.363%	92.871%	92.480%
Proposed Model (CNN-LSTM) + ADASYN	86.476%	96.517%	87.362%	81.194%	91.711%	91.829%
Proposed Model (CNN-LSTM) + SMOTENN	73.553%	98.235%	70.383%	92.458%	82.009%	91.251%
Proposed Model (CNN-LSTM) + SMOTETomek	86.431%	96.635%	87.192%	81.891%	91.671%	92.024%

Note: Receiver Operating Characteristic - Area Under the Curve (ROC AUC), Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), Synthetic Minority Oversampling Technique (SMOTE), Support Vector Machine Synthetic Minority Oversampling Technique (SVMSMOTE), Adaptive Synthetic Sampling (ADASYN), Synthetic Minority Oversampling Technique Ensemble Neural Networks (SMOTENN), and Synthetic Minority Over-sampling Technique and Tomek Links (SMOTETomek).

oversampling. The methods only affect the amount of data trained on the model, without affecting those being tested. This condition is to ensure no bias when testing the model during the process.

Table VII shows the proposed model combined with resampling methods on the train data for sentiment analysis of Indonesian language product reviews. The resampling results present an increase in the ability of the proposed model to detect minority classes. In addition, the model accuracy decreases slightly, showing that the models have a fairer assessment between the majority and minority classes. The oversampling method significantly increases specificity by approximately 20% compared to the model without resampling. However, combining the under-over sampling method in SMOTENN and SMOTETomek does not produce any performance improvement. The Ensemble Neural Networks (ENN) method in SMOTENN appears overly aggressive in altering the data distribution by removing instances that do not associate with the majority. Moreover, SMOTETomek achieves relatively good results, outperforming several other oversampling methods. The Tomek links in SMOTETomek helps to clarify feature boundaries between sentiment classes. In general, the proposed model combined with the SVMSMOTE method provides the best performance in separating positive and negative classes, with an ROC AUC score of 92.48%. The combination of the proposed model and SVMSMOTE succeeds in increasing the specificity value by around 9% from the Naive Bayes model. The best model for detecting negative classes is the proposed model combined with SMOTENN. The model detects 92.458% of negative classes and obtains the best precision value. Generally, the use of resampling methods successfully improves the ability of the model to detect negative classes.

C. Validation Model

The selection of optimal hyperparameter values plays an important role in the deep learning model. In the research, hyperparameters include control over the training process, model complexity, regularization, optimization process, and activation function. The training process control parameter consists of the learning rate, batch size, and epoch values. The model complexity control parameters comprises the number of layers, the number of units, kernel size, stride, and number of filters. Additionally, the activation function control parameters include the type of activation in the layer. The model also implements the Adam optimization function and regularization using Dropout.

The hyperparameters with the best classification effect of the model are studied using TPE. The data is divided into a training and validation set from the original dataset. The evaluation metric is the loss value to find the optimal hyperparameter. Figure 5 shows the learning curve of the proposed model using the optimal hyperparameters searched using the TPE method. The curve signifies that the loss values in training and validation tends to be stable and consistently follows

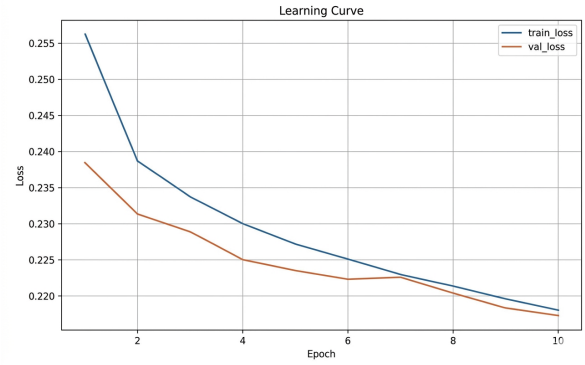

Fig. 5. Training set vs. validation set - Loss function curves for Tree-Structured Parzen Estimator (TPE) optimizations.

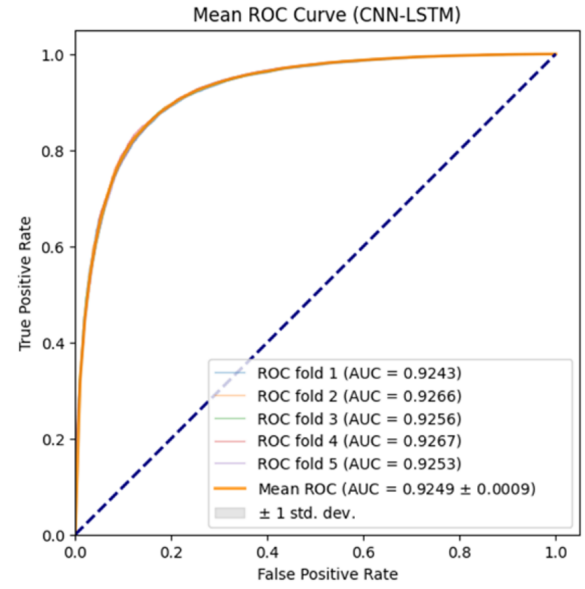

Fig. 6. Mean Receiver Operating Characteristic - Area Under the Curve (ROC AUC) curve for proposed model with 5-fold cross validation. Note: Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM).

the trend. It shows that the hyperparameters in the proposed model captures the underlying patterns and sentiment of the reviews, implying its reliability as well as predictive ability.

The stratified k-fold cross-validation method is used to validate the model performance, where the k value in the research is 5. Figure 6 shows the mean ROC AUC curve for the proposed model with 5-fold cross-validation. The ROC AUC results for each fold signify consistent values with a very low standard deviation of 0.0009. These results also show good generalization ability of the proposed model using optimal hyperparameters. The average value of the ROC AUC in the proposed model is 0.9249.

IV. Conclusion

In conclusion, the research examines the hybrid deep learning model for training Indonesian product reviews, optimizing the model through hyperparameter tuning and resampling methods. The main objective of the analysis is to develop a model capable of reliably predicting sentiment in the Indonesian language dataset. To support this, a large Indonesian sentiment dataset is constructed with multiple resampling variations, enabling further investigations into handling data imbalance in deep learning. The hybrid model, combining IndoBERT and a CNN-LSTM architecture with six resampling methods, is trained using parallel feature-level fusion and optimized with TPE tuning for sentiment analysis. Experimental results show that the hybrid model achieves an F1-Score of 94.875%, outperforming the baseline model and other machine learning models. On an imbalanced customer review dataset, combining the hybrid model with the SVMSMOTE resampling method produces promising results, achieving ROC AUC score of 92.48% in distinguishing positive and negative sentiment. During the process of this analysis, many customer reviews contain mixed Indonesian and English text.

Despite showing promising results, the research still has limitations. First, while the hybrid approach addresses class imbalance, the computational cost of the CNN-LSTM model is higher than traditional methods, posing challenges for real-time applications. Second, the model's performance on mixed-language texts, such as Indonesian and English, requires further optimization. Future work should focus on optimizing computational efficiency and expanding the training data to include diverse domains.

Future studies also need to develop a multilingual sentiment analysis model that incorporates Large Language Models (LLMs) and a Generative Adversarial Networks (GANs) to improve information extraction and address class imbalance challenges. The ability of LLMs to capture semantic knowledge is demonstrated by their strong performance in natural language understanding tasks across multiple languages. Meanwhile, GANs can generate synthetic samples for minority classes in multilingual datasets. This integration offers a robust solution for sentiment analysis on imbalanced datasets, leveraging GANs for data augmentation and LLMs for maintaining semantic fidelity.

Acknowledgement

The authors acknowledge the financial support from the Ministry of Higher Education, Science, and Technology of Indonesia (103/E5/PG.02.00.PL/2024).

Author Contribution

Designed research experiments, A. H. and R. Y.; Crawled the data from website and application, A. H. and E. H.; Built program code (tools) to run experiments, A. H. and E. H.; Conducted preprocessing process and hybrid model experiments, A. H. and E. H.; Wrote the paper, A. H., R. Y., E. H.; Reviewed the paper, A. H., and R. Y.; and Resampled process analysis, R. Y.

Data Availability

The data that support the findings of the research are available from the corresponding author, Arwin Halim, upon reasonable request. The data are not publicly available due to institutional data sharing policies.

References

[1] H. Huang, A. A. Zavareh, and M. B. Mustafa, "Sentiment analysis in e-commerce platforms: A review of current techniques and future directions," *IEEE Access*, vol. 11, pp. 90 367–90 382, 2023.

[2] V. O. Tama, Y. Sibaroni, and Adiwijaya, "Labeling analysis in the classification of product review sentiments by using multinomial Naive Bayes algorithm," *Journal of Physics: Conference Series*, vol. 1192, pp. 1–11, 2019.

[3] R. Catelli, S. Pelosi, and M. Esposito, "Lexicon-based vs. BERT-based sentiment analysis: A comparative study in Italian," *Electronics*, vol. 11, no. 3, pp. 1–20, 2022.

[4] C. Fiarni, H. Maharani, and R. Pratama, "Sentiment analysis system for Indonesia online retail shop review using hierarchy Naive Bayes technique," in *2016 4th International Conference on Information and Communication Technology (ICoICT)*. Bandung, Indonesia: IEEE, May 25–27, 2016, pp. 1–6.

[5] A. Daza, N. D. G. Rueda, M. S. A. Sánchez, W. F. R. Espíritu, and M. E. C. Quiñones, "Sentiment analysis on e-commerce product reviews using machine learning and deep learning algorithms: A bibliometric analysis, systematic literature review, challenges and future works," *International Journal of Information Management Data Insights*, vol. 4, no. 2, pp. 1–20, 2024.

[6] E. Halim, R. Purba, and A. Andri, "Consumer opinion extraction using text mining for product recommendations on e-commerce," *Indonesian Journal of Artificial Intelligence and Data Mining*, vol. 4, no. 1, pp. 19–28, 2021.

[7] N. Hayatin, G. I. Marthasari, and L. Nuarini, "Optimization of sentiment analysis for indonesian presidential election using Naïve Bayes and Particle Swarm Optimization," *Jurnal Online Informatika*, vol. 5, no. 1, pp. 81–88, 2020.

[8] A. Romadhony, S. Al Faraby, R. Rismala, U. N. Wisesti, and A. Arifianto, "Sentiment analysis on a large indonesian product review dataset." *Journal of Information Systems Engineering & Business Intelligence*, vol. 10, no. 1, pp. 167–178, 2024.

[9] J. H. Computer, S. M. Honova, V. P. Computer, C. A. Setiawan, I. H. Parmonangan, and Diana, "Sentiment analysis of skincare product reviews in Indonesian language using IndoBERT and LSTM," in *2023 IEEE 9th Information Technology International Seminar (ITIS)*. Batu Malang, Indonesia: IEEE, Oct. 18–20, 2023, pp. 1–6.

[10] 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, no. 1, pp. 1–17, 2024.

[11] W. F. Satrya, R. Aprilliyani, and E. H. Yossy, "Sentiment analysis of Indonesian police chief using multi-level ensemble model," *Procedia Computer Science*, vol. 216, pp. 620–629, 2023.

[12] R. Kusumaningrum, I. Z. Nisa, R. Jayanto, R. P. Nawangsari, and A. Wibowo, "Deep learning-based application for multilevel sentiment analysis of Indonesian hotel reviews," *Heliyon*, vol. 9, no. 6, pp. 1–12, 2023.

[13] A. K. Gogineni, S. K. Sai Reddy, H. Kakarala, Y. C. Gavini, M. P. Venkat, K. Hajarathaiah, and M. K. Enduri, "A hybrid deep learning framework for efficient sentiment analysis," *International Journal of Advanced Computer Science & Applications*, vol. 14, no. 12, pp. 1032–1038, 2023.

[14] M. M. Rahman, A. I. Shiplu, Y. Watanobe, and M. A. Alam, "RoBERTa-BiLSTM: A context-aware hybrid model for sentiment analysis," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 9, no. 6, pp. 3788–3805, 2025.

[15] C. H. Lin and U. Nuha, "Sentiment analysis of indonesian datasets based on a hybrid deep-learning strategy," *Journal of Big Data*, vol. 10, pp. 1–19, 2023.

[16] TMO Group, "SEA eCommerce: Sales data by country & industry (+ free reports)," 2024. [Online]. Available: https://www.tmogroup.asia/insights/southeast-asia-ecommerce-data-monthly-updates/

[17] A. Hussain, V. Dhanawat, A. Aslam, N. Iqbal, and S. Tripura, "Credit card fraud detection using machine learning techniques: Dealing with imbalanced data using over-sampling and under-sampling methods," in *2024 Beyond Technology Summit on Informatics International Conference (BTS-I2C)*. East Java, Indonesia: IEEE, Dec. 19, 2024, pp. 676–681.

[18] C. F. G. D. Santos and J. P. Papa, "Avoiding overfitting: A survey on regularization methods for convolutional neural networks," *ACM Computing Surveys (CSUR)*, vol. 54, no. 10s, pp. 1–25, 2022.

[19] TMO Group, "10 largest online marketplaces in Southeast Asia (2024)," 2024. [Online]. Available: https://www.tmogroup.asia/insights/must-know-southeast-asia-online-marketplaces/

[20] Z. Rahimi and M. M. Homayounpour, "The impact of preprocessing on word embedding quality: A comparative study," *Language Resources and Evaluation*, vol. 57, no. 1, pp. 257–291, 2023.

[21] Q. Lu, X. Sun, Y. Long, Z. Gao, J. Feng, and T. Sun, "Sentiment analysis: Comprehensive reviews, recent advances, and open challenges," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 11, pp. 15 092–15 112, 2023.

[22] K. Smelyakov, D. Karachevtsev, D. Kulemza, Y. Samoilenko, O. Patlan, and A. Chupryna, "Effectiveness of preprocessing algorithms for natural language processing applications," in *2020 IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S&T)*. Kharkiv, Ukraine: IEEE, Oct. 2–9, 2020, pp. 187–191.

[23] J. Y. B. Yin, N. H. M. Saad, and Z. Yaacob, "Exploring sentiment analysis on e-commerce business: Lazada and Shopee," *TEM Journal*, vol. 11, no. 4, pp. 1508–1519, 2022.

[24] A. Singh and J. O'Hagan, "Exploring topic modelling of user reviews as a monitoring mechanism for emergent issues within social VR communities," 2024. [Online]. Available: https://arxiv.org/abs/2406.03994

[25] A. G. Prabono, "Mpstemmer: A multi-phase stemmer for standard and nonstandard Indonesian words," 2020. [Online]. Available: https://github.com/ariaghora/mpstemmer

[26] H. Murfi, T. Gowandi, G. Ardaneswari, and S. Nurrohmah, "BERT-based combination of convolutional and recurrent neural network for Indonesian sentiment analysis," *Applied Soft Computing*, vol. 151, 2024.

[27] C. Zhou, Q. Li, C. Li, J. Yu, Y. Liu, G. Wang, K. Zhang, C. Ji, Q. Yan, L. He, H. Peng, J. Li, J. Wu, Z. Liu, P. Xie, C. Xiong, J. Pei, P. S. Yu, and L. Sun, "A comprehensive survey on pretrained foundation models: A history from BERT to ChatGPT," *International Journal of Machine Learning and Cybernetics*, vol. 16, no. 12, pp. 9851–9915, 2025.

[28] 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*. Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020, pp. 757–770.

[29] B. Wilie, K. Vincentio, G. I. Winata, S. Cahyawijaya, X. Li, Z. Y. Lim, S. Soleman, R. Mahendra, P. Fung, S. Bahar, and A. Purwarianti, "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*. Suzhou, China: Association for Computational Linguistics, 2020, pp. 843–857.

[30] A. U. Rehman, A. K. Malik, B. Raza, and W. Ali, "A hybrid CNN-LSTM model for improving accuracy of movie reviews sentiment analysis," *Multimedia Tools and Applications*, vol. 78, no. 18, pp. 26 597–26 613, 2019.

[31] L. Yang, Y. Li, J. Wang, and R. S. Sherratt, "Sentiment analysis for e-commerce product reviews in Chinese based on sentiment lexicon and deep learning," *IEEE Access*, vol. 8, pp. 23 522–23 530, 2020.

[32] X. Wang, W. Jiang, and Z. Luo, "Combination of convolutional and recurrent neural network for sentiment analysis of short texts," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan: The COLING 2016 Organizing Committee, 2016, pp. 2428–2437.

[33] J. Sangeetha and U. Kumaran, "A hybrid optimization algorithm using BiLSTM structure for sentiment analysis," *Measurement: Sensors*, vol. 25, pp. 1–7, 2023.

[34] S. Tao, P. Peng, Y. Li, H. Sun, Q. Li, and H. Wang, "Supervised contrastive representation learning with Tree-Structured Parzen Estimator Bayesian optimization for imbalanced tabular data," *Expert Systems with Applications*, vol. 237, 2024.

[35] G. Békési, L. Barancsuk, and B. Hartmann, "Deep neural network based distribution system state estimation using hyperparameter optimization," *Results in Engineering*, vol. 24, pp. 1–14, 2024.

[36] N. Zhou, B. Shang, M. Xu, L. Peng, and G. Feng, "Enhancing photovoltaic power prediction using a CNN-LSTM-attention hybrid model with Bayesian hyperparameter optimization," *Global Energy Interconnection*, vol. 7, no. 5, pp. 667–681, 2024.

[37] P. Mooijman, C. Catal, B. Tekinerdogan, A. Lommen, and M. Blokland, "The effects of data balancing approaches: A case study," *Applied Soft Computing*, vol. 132, pp. 1–32, 2023.

[38] U. B. Mahadevaswamy and P. Swathi, "Sentiment analysis using bidirectional LSTM network," *Procedia Computer Science*, vol. 218, pp. 45–56, 2023.