Jurnal Teknik Informatika (JUTIF) Vol. 6, No. 4, August 2025, Page. 1881-1896
P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4878

Comparative Study of BiLSTM and GRU for Sentiment Analysis on Indonesian E-Commerce Product Reviews Using Deep Sequential Modeling

Khairunnisa Nasution[1], Khairun Saddami*[2], Roslidar Roslidar[3], Akhyar Akhyar[4], Fathurrahman Fathurrahman[5], Niza Aulia[6]

[1]Master Program of Electrical Engineering, Universitas Syiah Kuala, Indonesia
[2,3,6]Department of Electrical and Computer Engineering, Universitas Syiah Kuala, Indonesia
[4]Department Electrical, Electronic and Systems Engineering, Universiti Kebangsaan Malaysia, Malaysia
[5]Electrical Engineering Department, King Fahd University of Petroleum and Minerals, Saudi Arabia

Email: [2]khairun.saddami@usk.ac.id

Received : Jun 13, 2025; Revised : Jun 27, 2025; Accepted : Aug 16, 2025; Published : Aug 18, 2025

Abstract

Sentiment analysis plays a crucial role in understanding customer perspectives, especially within Indonesian e-commerce platforms. Despite the success of deep learning in high-resource languages, its application to Indonesian sentiment data remains underexplored. Previous studies using models like BERT-CNN or fine-tuned IndoBERT achieved modest results, highlighting the need for more effective architectures for Indonesian language. This study aims to investigate the effectiveness of Bidirectional Long Short-Term Memory (BiLSTM) and Gated Recurrent Unit (GRU) models in classifying buyers' sentiment from Indonesian product reviews on the PREDECT-ID dataset comprising 5,400 annotated product reviews. Standard NLP preprocessing techniques—including text normalization, tokenization, stopword removal, and stemming—were applied. Both models were trained using Adam and Stochastic Gradient Descent (SGD) optimizers, and their performance was evaluated using accuracy, precision, recall, and F1-score metrics. The GRU model trained with SGD achieved the highest performance, with an accuracy of 94.07%, precision of 93.84%, recall of 94.53%, and F1-score of 94.18%. Notably, the BiLSTM model combined with SGD resulted in competitive results, achieving 93.61% accuracy and 93.84% F1-score. The results confirm that GRU with SGD optimizer, are highly effective for sentiment classification in Indonesian language datasets. By leveraging deep sequential modeling for a low-resource language, this study contributes to the advancement of scalable sentiment analysis systems in underrepresented linguistic domains. The results contribute to the advancement of NLP systems for Indonesian by providing a benchmark for the future development of sentiment analysis tools in low-resource languages.

Keywords : *BiLSTM, Deep sequential representation, GRU, Indonesian product review, Sentiment Analysis.*

1. INTRODUCTION

Product reviews, both in traditional and electronic commerce, offer valuable indicators of customer satisfaction and expectations, helping prospective buyers make informed purchase decisions [1]. The perceived usefulness of a review is influenced by factors such as its length, the reviewer's experience, and the type of product evaluated. Additionally, reviews are critical for assessing reviewer credibility and identifying misleading or fraudulent content, which is essential for maintaining trust in online transactions [2]. With the rapid growth of e-commerce in Indonesia, analyzing customer reviews is increasingly important. Sentiment analysis, a branch of NLP, helps businesses extract insights to gauge satisfaction, spot product issues, and refine marketing strategies [3].

Deep learning is widely used for sentiment classification, especially via Recurrent Neural Networks (RNN). Long Short-Term Memory (LSTM), an RNN variant, effectively captures long-term dependencies in sentiment analysis tasks [4]. Bidirectional LSTM (BiLSTM) enhances this by processing data in both directions for better context understanding [5]. Alternatively, the Gated Recurrent Unit (GRU) offers a simpler, efficient architecture that addresses the vanishing gradient issue while remaining computationally lightweight and effective for sequential data tasks [6].

Previous studies have shown that deep learning models like BERT-CNN achieve accuracy levels between 70–79% [7], while fine-tuned IndoBERT reaches an F1-Score of 71% [8]. Meanwhile, GRU-based approaches show promising results, with accuracy rates ranging from 65% [9] to 96% [10], and BiLSTM has achieved accuracy of 90.5% on the IMDb dataset [11]. While transformer-based models such as IndoBERT and BERT-CNN have demonstrated strong performance in various sentiment analysis tasks, they often require substantial computational resources and large-scale training data, which may not be optimal for moderately sized or domain-specific datasets. In contrast, BiLSTM and GRU architectures offer a more lightweight yet effective alternative for capturing sequential dependencies [12], [13], [14]. Anam demonstrated that a hybrid GRU–BiLSTM model with SMOTE significantly improves emotion detection accuracy on social media text [12], while Rahman showed the efficacy of transformer–RNN hybrids—specifically RoBERTa–BiLSTM—in enhancing sentiment analysis performance across multiple English datasets [13]. These results making them particularly suitable for Indonesian sentiment analysis using datasets like PREDECT-ID. The relevance of PREDECT-ID has been highlighted in recent studies focusing on contextual sentiment understanding in e-commerce settings, demonstrating its utility for benchmarking.

Despite growing interest in Indonesian-language sentiment analysis, prior research remains limited in its comparative evaluation of deep sequential architectures under consistent experimental settings. This study addresses this gap by investigating the effectiveness of BiLSTM and GRU models in classifying binary sentiments from Indonesian product reviews. This research not only aims to improve classification accuracy but also contributes to advancing NLP methodologies for low-resource languages such as Indonesian, providing a foundation for future scalable applications.

2. METHOD

To investigate the performance of Biderictional Long Short-Term Memory (BiLSTM) and Gate Recurrent Unit (GRU) architectures in Indonesian-language sentiment classification, this study employed a structured experimental pipeline consists of data preparation, preprocessing, model construction, and evaluation. The PREDECT-ID dataset, comprising thousands of annotated product reviews, served as the primary data source [15]. Standard Natural Language Processing (NLP) techniques were applied to prepare the textual data, followed by training and testing of the deep learning models using two different optimization strategies. This section outlines the dataset characteristics, preprocessing steps, model configurations, training procedures, and evaluation metrics used in this study.

The PREDECT-ID dataset includes five emotion categories; Love, Happiness, Anger, Fear, and Sadness. For the purpose of simplicity, these emotions were recategorized into two sentiment classes: Anger, Fear, and Sadness are grouped under negative sentiment, while Love and Happiness are grouped under positive sentiment. This dataset comprises a total of 2,579 of positive and 2,821 samples of negative sentiment, respectively. The data were split into 70% for training, 10% for validation, and 20% for testing. Of the 2,579 positive sentiment samples, 2,047 were allocated for training and validation, while 532 were used for testing. Similarly, out of the 2,821 negative sentiment samples, 2,273 were assigned to training and validation, and 548 were reserved for testing.

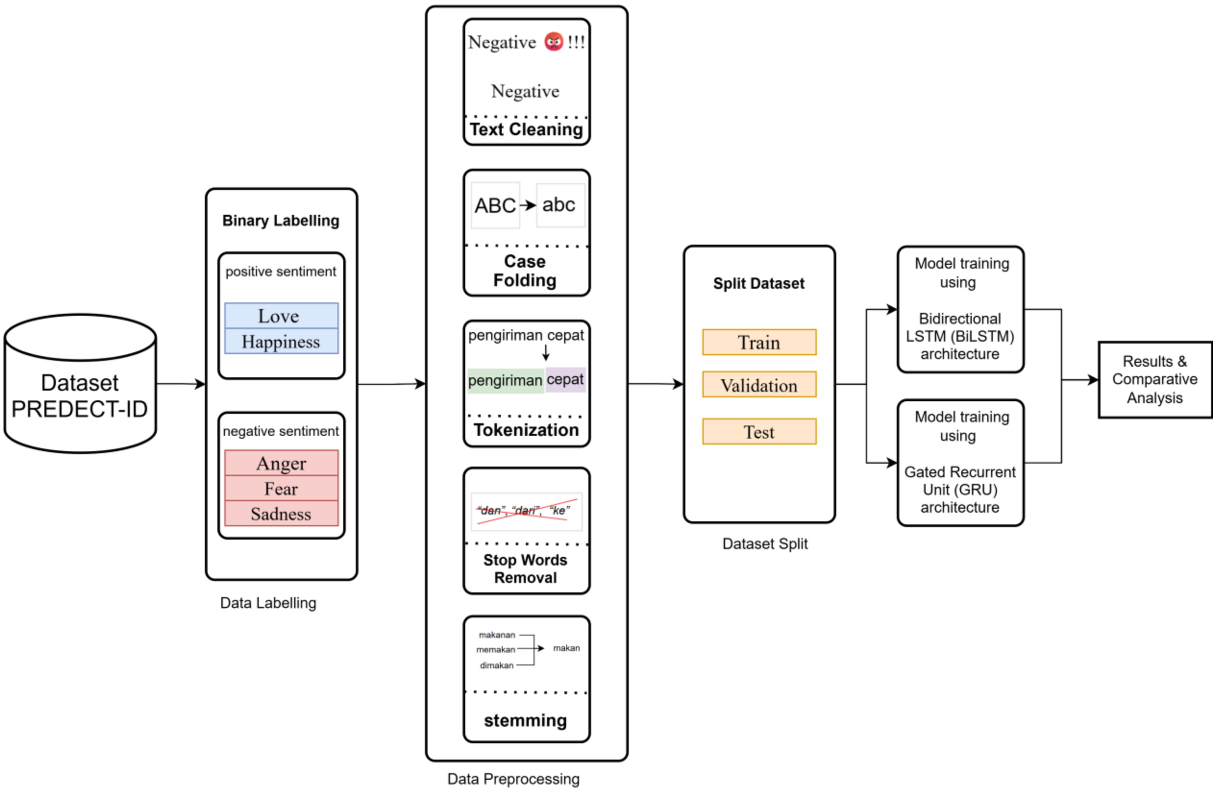


Figure 1. The research framework

Figure 1 showed the research methodology of the proposed system. The research workflow illustrated in the diagram shown a systematic approach to sentiment classification using the PREDICT-ID dataset and deep learning models.

2.1. Data Collection

The raw dataset consists of Indonesian-language product reviews collected from various e-commerce platforms. These reviews are annotated into two sentiment categories: positive (e.g., *love, happiness*) and negative (e.g., *anger, fear, sadness*), forming a binary classification task. The dataset used in this study is PREDECT-ID, comprising 5,400 labeled sentences [15].

2.2. Preprocessing and Data Splitting

The preprocessing stage is essential to prepare the textual data for model training. It begins with text cleaning to eliminate irrelevant symbols, punctuation, and noise from the raw reviews. Next, case folding is applied to standardize the text by converting all characters to lowercase. This is followed by tokenization, which breaks each sentence into individual words or tokens. To reduce non-informative content, stop word removal is performed, eliminating frequently occurring words that do not contribute significantly to sentiment (e.g., *yang*, *dan*). Finally, stemming is applied to reduce words to their root forms, ensuring consistency and reducing vocabulary size. These preprocessing steps collectively help normalize the data, improve learning efficiency, and enhance the model's ability to capture sentiment-relevant patterns. The preprocessed data are divided into training, validation, and testing sets to ensure robust model development and fair performance evaluation.

2.3. Deep Sequential Representation

Deep sequential representations have been proven effective for modeling sequential data. In this study, we employ two models based on deep sequential representation, BiLSTM and GRU [14]. The

BiLSTM model incorporates a trainable embedding layer that transforms each input token into a dense vector of dimension 128. This layer transforms each word in the input sequence into a dense, continuous-valued vector representation of length 128. The embedding layer learns semantic relationships between words during training. Following the embedding layer, a Bidirectional LSTM layer with 64 hidden units in each direction. This layer processes the input sequence in both forward and backward directions, enabling the network to capture contextual information from past and future tokens simultaneously. The sequence output from the BiLSTM is passed to a GlobalMaxPooling1D layer, which reduces the temporal dimension across time steps for each feature map, thereby retaining the most salient features while reducing model complexity. The pooled feature vector is passed through a dense layer with 64 units and ReLU activation, enabling non-linear transformations that enhance the model's capacity to learn discriminative features. To mitigate overfitting, a Dropout layer with a rate of 0.5 is applied. Finally, a dense output layer with a single unit and sigmoid activation is employed to generate a probability score in the range [0,1], indicating the likelihood that the input text expresses a positive sentiment. This configuration is appropriate for binary classification tasks. Figure 2 showed the architecture of the BiLSTM and GRU.

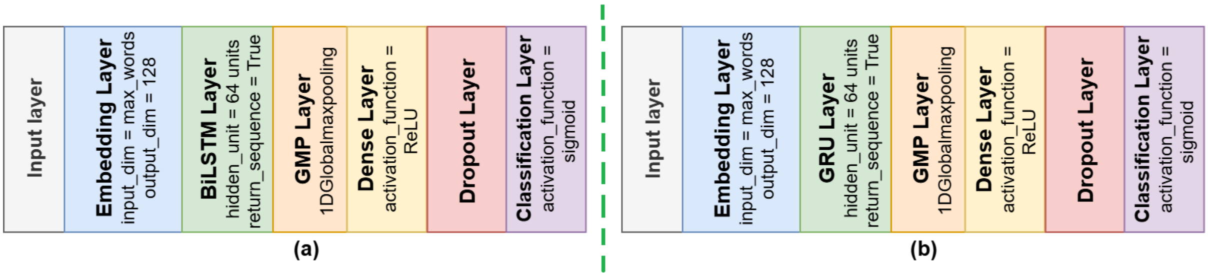


Figure 2. Deep sequencial representation, where: (a). BiLSTM architecture, (b). GRU architecture

The proposed GRU as well as BiLSTM using embedding layers with a vector dimension of 128. Then, a GRU layer with 64 hidden units follows the embedding layer. This recurrent layer captures temporal dependencies within the sequence, effectively learning contextual patterns in the text. To reduce the dimensionality and retain the most salient features across time steps, a GlobalMaxPooling1D layer is applied. This layer extracts the maximum activation from each feature map across the sequence length, effectively summarizing the entire input into a fixed-size representation. The output from the pooling layer is fed into a Dense layer with 64 units and ReLU activation, introducing non-linearity and enabling the model to learn complex decision boundaries. ReLU activation is formulated as Eq.1 [20]:

$$f(x) = \max\,(x, 0) \qquad (1)$$

where x is the input to the layer. To prevent overfitting, a Dropout layer with a rate of 0.5 is employed. This randomly deactivates half of the neurons during training, improving generalization. The final layer is a Dense layer with 1 output unit and a sigmoid activation function, which maps the output to a probability value between 0 and 1. This is appropriate for binary sentiment classification, where the output represents the probability of the input text being positive. Sigmoid activation function is calculated using Eq.2 [19]:

$$\sigma(x) = \frac{1}{1+e^{-x}} \qquad (2)$$

In this research, the binary cross-entropy loss function is used due to its simplicity and its widespread use in binary classification tasks. The binary cross-entropy loss function is formulated as Eq.3 [18]:

$$H_p\ (q) = -\frac{1}{N}\sum_{i=1}^{N} y_i.\log\big(p(y_i)\big) + (1 - y_i).\ \log(1 - p(y_i)) \qquad (3)$$

where $H_p\ (q)$ is the loss value, N is the number of dataset or batch, y_i is the true label of dataset, $p(y_i)$ is the probability of the prediction results.

2.4. Experiment Setup

In this section, we describe the experimental setup used to evaluate the performance of BiLSTM and GRU models on the PREDECT-ID dataset. All experiments were conducted using Python 3.8 on a machine equipped with an NVIDIA RTX 3060 GPU. The dataset was split into training, validation and testing sets in an 80:10:10 ratio, and standard preprocessing steps were applied to normalize the Indonesian-language text data. The BiLSTM and GRU models were trained using the hyperparameter settings presented in Table 1. These hyperparameters are the best fine-tuned during training process.

Table 1. Hyperparameters setting

Hyperparameters	Value
Batch Size	32
Embedding Dimension	128
Dropout Rate	0.2, 0.5, 0.75
Optimizer	SGD, Adam
Learning Rate	0.001
Epoch	100
Loss Function	binary_crossentropy
Activation Function	Sigmoid

To highlight the performance advantages of BiLSTM and GRU in sentiment analysis, this study compares the effectiveness of these models against several models, including LSTM [21, 22], [23], Support Vector Machine (SVM) [24], K-Nearest Neighbors (KNN) [25], Naïve Bayes [26], and Random Forest [27]. These models were selected to represent both deep and classical approaches in order to provide a comprehensive evaluation across different algorithmic paradigms. To ensure consistency in feature representation, all models except LSTM used the Term Frequency–Inverse Document Frequency (TF-IDF) technique for feature extraction [28]. TF-IDF is a widely adopted method that emphasizes terms that are informative while reducing the weight of frequently occurring, sentiment-neutral words. This representation allows machine learning models such as SVM and Random Forest to focus on discriminative features that are most relevant to sentiment classification. Through this comparative framework, we aim to underscore the empirical advantages of deep learning models in handling linguistic nuances that are often challenging for traditional classifiers relying on sparse and shallow feature representations.

2.5. Evaluation Metrics

We evaluated the performance of BiLSTM and GRU using accuracy, recall, precision and f1-score. These metrics are computed based on the confusion matrix, which records the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). By using these metrics, we aim to obtain a robust and interpretable evaluation of each model's effectiveness in real-world sentiment analysis scenarios.

Accuracy evaluates the overall correctness of the model's predictions by calculating the ratio of correctly predicted instances to the total number of samples. Accuracy is formulated as Eq.4 [21, 28]:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (4)$$

Recall quantifies the model's ability to correctly identify all relevant positive cases. Recall is calculated using Eq.5:

$$recall = \frac{TP}{TP+FN} \tag{5}$$

Precision measures the proportion of positive predictions that are truly positive, offering insight into the model’s reliability when predicting a specific class. Precision can be computed by Eq.6:

$$precision = \frac{TP}{TP+FP} \tag{6}$$

F1-score is the harmonic mean of precision and recall, providing a balanced metric especially useful in cases of class imbalance. F1-scores formulated as Eq.7:

$$F1 - score = \frac{2\times Recall \times Precision}{Recall+Precision} \tag{7}$$

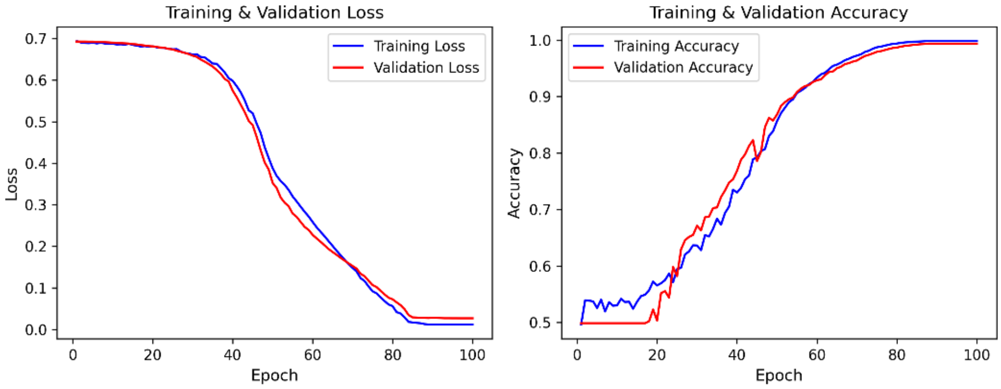
3. RESULT AND DISCUSSION

This section presents and discusses the empirical findings obtained from the training and evaluation of the proposed deep learning models for sentiment classification of Indonesian-language e-commerce product reviews. The analysis is structured into three main parts: training performance, testing results, and comparative evaluation with previous related studies.

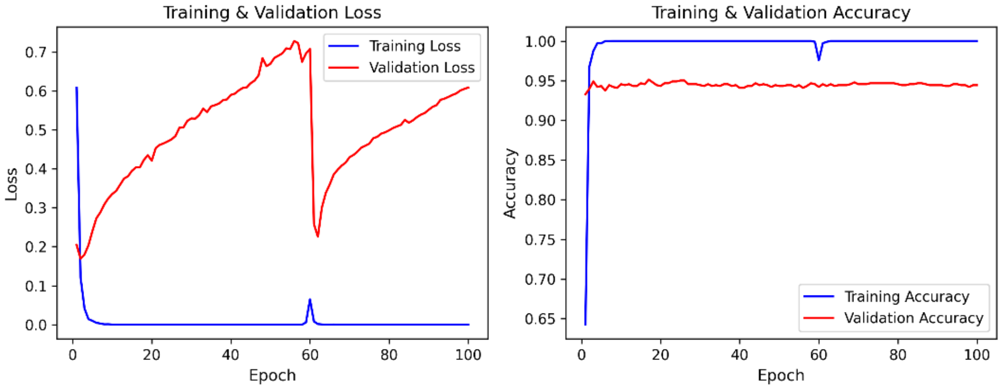
3.1. Training Results

The training result is represented by learning curve behaviour across epochs. Learning curve that represented training accuracy and loss are visualized and discussed to analyze the convergence trends and model capacity.

Figure 3 show training result on BiLSTM using SGD dan Adam optimizers. The learning curves indicate that SGD outperforms Adam in terms of stability, convergence, and generalization. SGD demonstrates a smooth and consistent reduction in loss during training, with training and validation losses closely following each other, which indicates effective generalization without overfitting. Validation accuracy also closely follows training accuracy, further emphasizing its robustness. In contrast, Adam exhibits rapid initial learning but struggles with overfitting. The rapid rise in training accuracy, followed by stagnation or decline in validation accuracy, points to Adam's inability to generalize well to unseen data. The fluctuations in the validation loss curve and the widening gap between training and validation accuracy highlight the instability in Adam's learning process.

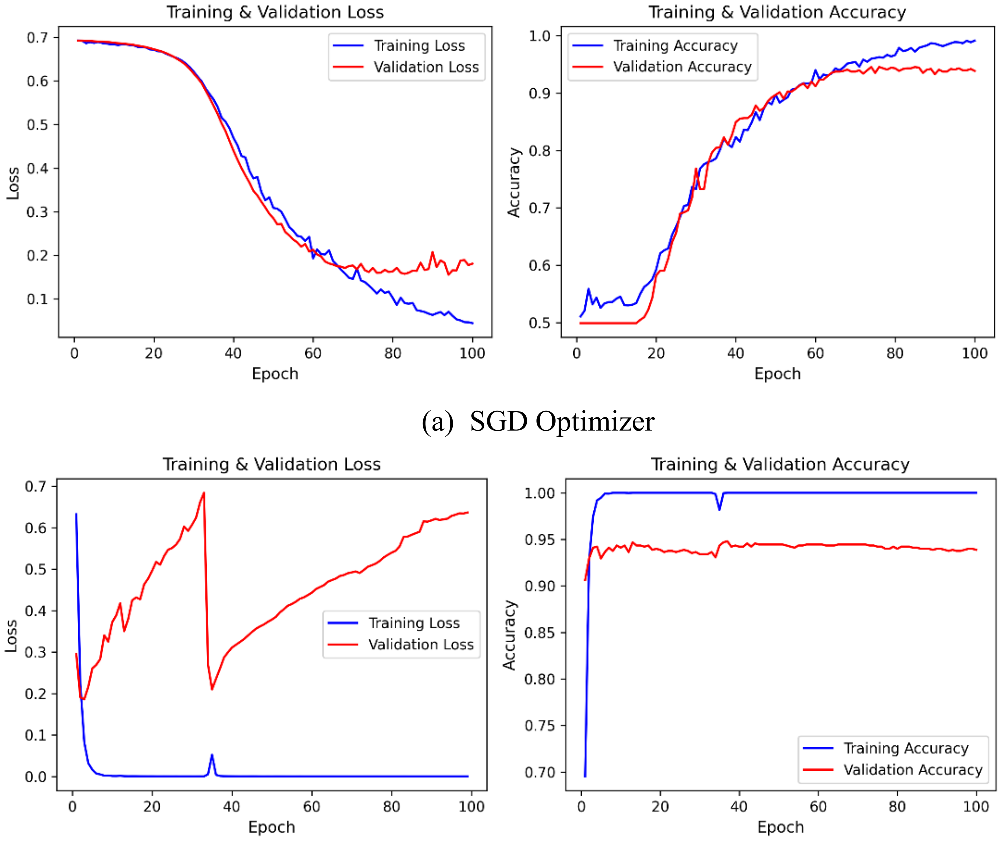


(a) SGD optimizer



(b) Adam Optimizer

Figure 3 Training Results using BiLSTM architecture



(a) SGD Optimizer

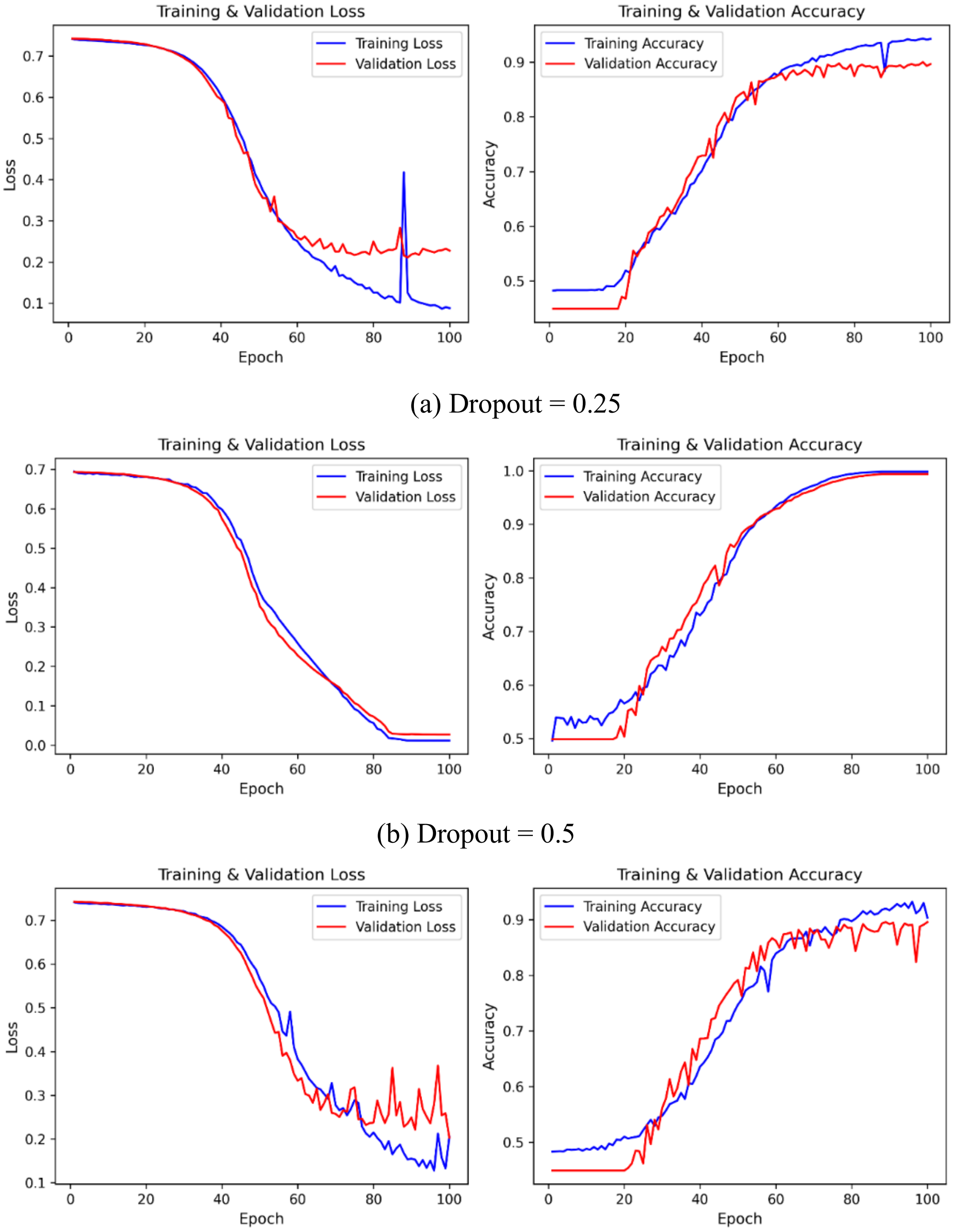
(b) Adam Optimizer

Figure 4 Training Results using BiLSTM architecture

Figure 4 shown the comparison between the SGD and Adam optimizers used in build GRU models that highlights significant differences in the training dynamics of the GRU model. When using the SGD optimizer, both training and validation loss resulted in a gradual and synchronized decline. The validation loss remains relatively stable toward the end of training, indicating that the model generalizes well to unseen data. Similarly, training and validation accuracy increase in parallel, reaching values above 0.95. This demonstrates that the model trained with SGD achieved a stable and balanced learning process, avoiding both overfitting and underfitting.

In contrast, the Adam optimizer shows unstable learning behavior. Training loss drops rapidly to near zero within a few epochs, while validation loss increases and fluctuates significantly, especially around epoch 40. Training accuracy reaches almost 100% early, but validation accuracy remains stagnant and considerably lower. This reflects severe overfitting, where the model memorizes the training data but fails to learn generalized representations.

We varied the dropout rate to identify the most effective model for learning patterns from the dataset. Figures 5 and 6 illustrate the training performance of models under different dropout configurations. The objective was to evaluate how regularization through dropout impacts model generalization. All models were trained using the SGD optimizer, as preliminary experiments revealed that the Adam optimizer failed to adequately learn the patterns in the dataset. Therefore, SGD was chosen for its more stable convergence behavior in this particular task and dataset context.



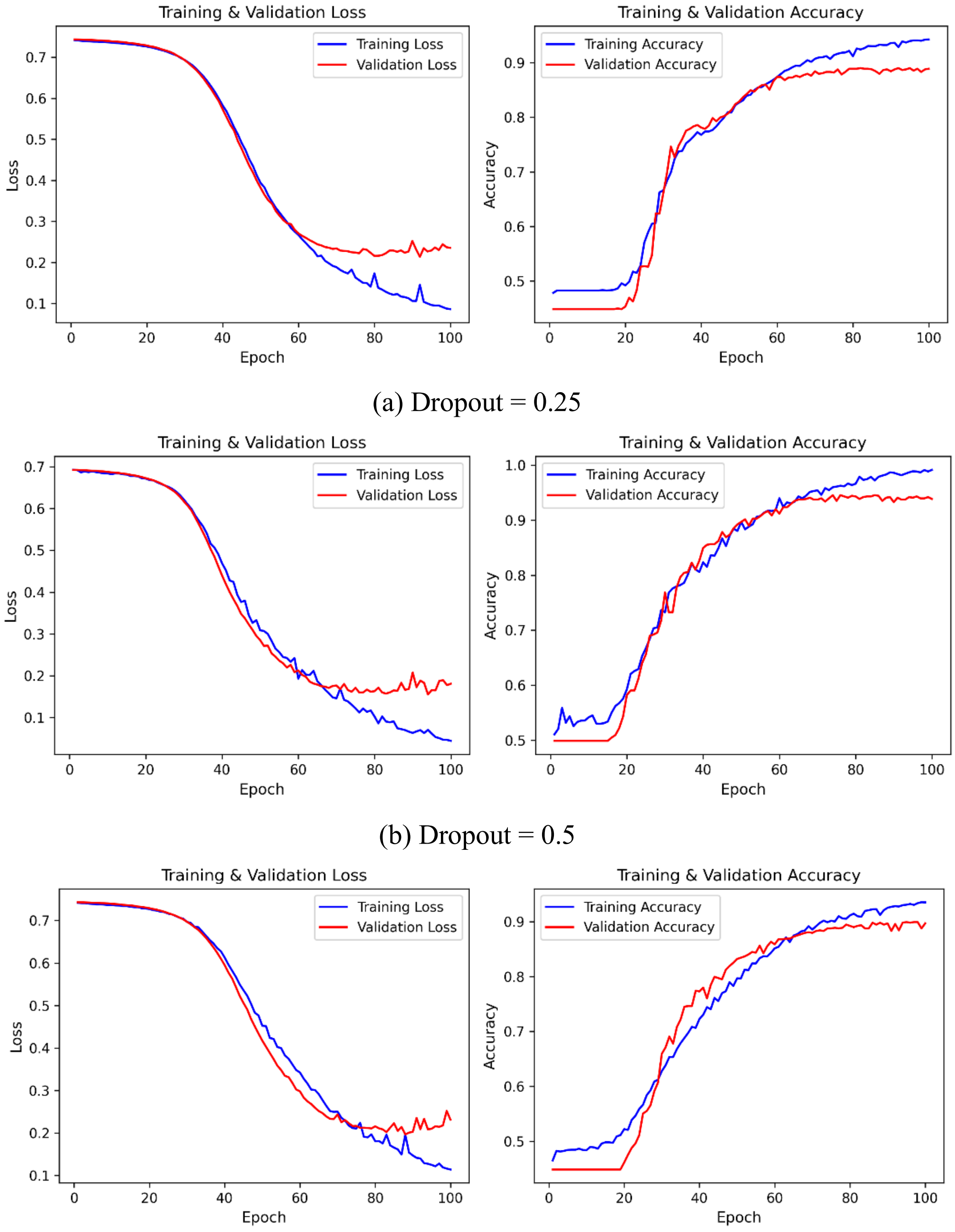
(a) Dropout = 0.25

(b) Dropout = 0.5

(c) Dropout = 0.75

Figure 5. Training results with various dropout on BiLSTM model

Based on Figure 5, the best model performance was achieved when the dataset was trained using a dropout rate of 0.5. In contrast, models trained with dropout values of 0.25 and 0.75 failed to converge within 100 epochs. Both models exhibit difficulties in learning from data, although for different reasons. In contrast, the model trained with a dropout rate of 0.5 exhibited the most balanced and stable learning curve. Training and validation loss decreased in sync, and both accuracies increased steadily to nearly 100%. This indicates strong generalization and effective learning. Therefore, a dropout rate of 0.5 can be considered optimal for BiLSTM architectures to prevent both underfitting and overfitting.



(a) Dropout = 0.25

(b) Dropout = 0.5

(c) Dropout = 0.75

Figure 6. Training results with various dropout on GRU model

According to Figure 6, the best model was represented by model trained using droput value of 0.5. Models trained with dropout values of 0.25 and 0.75 failed to converge effectively within 100 epochs, indicating that both configurations struggled to learn from the data. A small dropout value (0.25) causes insufficient regularization, leading to overfitting. This is reflected by the steep decline in training loss while the validation loss increases near the end of training. Although training accuracy reaches high values, validation accuracy stagnates and becomes unstable, showing poor generalization. Table 2 show comparison of dropout comparison used in training the model.

Table 2 shows the effect of dropout variation on the performance of BiLSTM and GRU models. In BiLSTM, dropout 0.5 produces the best accuracy and loss, indicating a balance between generalization and overfitting. Meanwhile, GRU achieves the highest training accuracy at dropout 0.5, but the best validation occurs at dropout 0.25, despite the high training loss. Dropout 0.75 in both models decreases accuracy, indicating too many neurons are deactivated. Thus, dropout 0.5 in BiLSTM and GRU provides the most stable and accurate performance overall.

Table 2. Comparison of dropout value for training model

Dropout value	Training accuracy	Validation accuracy	Training loss	Validation loss
		BiLSTM		
0.25	0.9344	0.8826	0.1031	0.2658
0.5	0.9831	0.9792	0.0557	0.0611
0.75	0.8981	0.8283	0.2011	0.0231
		GRU		
0.25	0.9447	0.8832	0.2415	0.0732
0.5	0.9923	0.9324	0.0262	0.2014
0.75	0.8944	0.9201	0.1041	0.2421

3.2. Testing Results

The testing result section focuses on the evaluation of model performance on previously unseen data. Evaluation metrics including accuracy, precision, recall, and F1-score are presented to quantify the effectiveness of the models in binary sentiment classification. Additionally, confusion matrices are included to provide detailed insight into the distribution of correct and incorrect predictions. The findings are interpreted with respect to the models' generalization ability and robustness in handling real-world product review data in the Indonesian language.

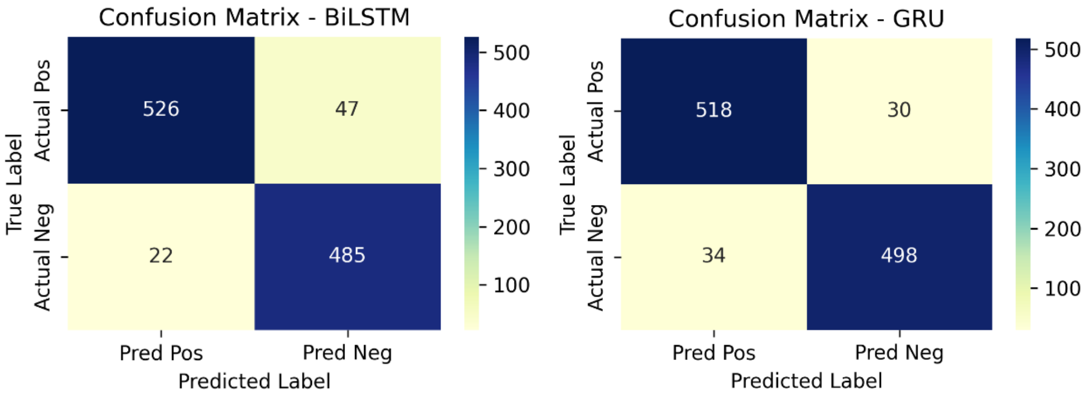


Figure 7. Confusion Matrix, where: (a). BiLSTM trained model, (b). GRU trained model

Based on Figure 7, it can be seen that 47 reviews that categorized as positive review is considered as negative. We analyzed and investigate that, Based on the confusion matrix results, it can be seen that

GRU has a better ability in detecting positive sentiment, where GRU only incorrectly predicted 30 positive sentiments while BiLSTM incorrectly detected up to 47 reviews. However, BiLSTM is very good at determining or predicting negative sentiment, where BiLSTM only incorrectly predicted 22 negative sentiments while GRU reached 34. The classification results reveal distinct performance tendencies between GRU and BiLSTM architectures in sentiment analysis tasks. Table 3 shown the example of GRU and BiLSTM correct and incorrect classify the reviews.

Table 3. Examples of misprediction by BiLSTM dan GRU

Reviews	Label	Predicted	
		BiLSTM	GRU
mantaapppppppppplpll gannnnnnnnnn	positive	positive	negative
mantaapppppppppplpll gannnnnnnnnn	positive	positive	negative
Awalnya saya kira bagus, tapi ternyata biasa saja.	negative	positive	negative
Barangnya oke, tapi tidak seperti yang saya harapkan.	negative	positive	negative

According to Table 3, the BiLSTM and GRU models reveal important architectural implications in sentiment analysis tasks, particularly in processing informal and contrastive language structures. In two instances, the review "mantaapppppppppplpll gannnnnnnnnn" was accurately classified as positive by BiLSTM but incorrectly by GRU.

3.3. Comparison with others

This subsection compares the performance of the proposed models BiLSTM and GRU with results from previous works or baseline methods reported in the literature. The comparison highlights both quantitative improvements and architectural distinctions. Benchmark metrics from models such as LSTM, Naïve Bayes, and SVM are discussed to contextualize the contributions of this research. This comparison validates the effectiveness of the chosen architectures and demonstrates their applicability for sentiment analysis in low-resource language scenarios.

Table 4. Comparison with others

No	Methods	Accuracy	Recall	Precision	F1-score
1	GRU	**94.07%**	**94.53%**	93.84%	**94.18%**
2	BiLSTM	93.61%	91.80%	**95.99%**	93.84%
3	LSTM [29]	92.59%	90.23%	94.49%	92.31%
4	Naive Bayes [30]	88.89%	84.59%	92.21%	88.24
5	KNN [25]	87.69%	84.40%	89.98%	87.10%
6	SVM [24]	90.46%	88.35%	91.98%	90.12%
7	Random Forest [31]	89.54%	87.22%	91.16%	89.14%

According to Table 4, the GRU model achieved the highest accuracy (94.07%) and F1-score (94.18%), with strong recall (94.53%) and balanced precision (93.84%), making it effective for general sentiment classification tasks. BiLSTM, on the other hand, attained the highest precision (95.99%) due to its bidirectional structure, which enhances its ability to capture contextual sentiment shifts. However, its lower recall (91.80%) suggests a tendency to over-predict positives, making GRU more suitable for balanced performance across varied scenarios.

3.4. Discussion

The results showed that optimizers and dropout configurations affect the performance of deep learning architectures in sentiment analysis. SGD consistently yielded more stable convergence than Adam in both BiLSTM and GRU. Its use of averaged gradients ensures controlled updates, preventing drastic shifts in parameter space, and facilitating smoother generalization. Conversely, Adam's adaptive learning and momentum caused rapid convergence but often led to overfitting, particularly visible in GRU, where training accuracy soared while validation accuracy stagnated. These observations align with previous findings highlighting Adam’s sensitivity to small datasets and gradient noise in sequence models.

Dropout regularization also played a critical role in model generalization. A dropout rate of 0.5 provided the most balanced performance in both architectures. Lower dropout (0.25) resulted in inadequate regularization, causing the model to overfit, while higher dropout (0.75) impaired learning due to excessive neuron deactivation—leading to underfitting. The optimal performance at 0.5 reflects its ability to retain enough model capacity while reducing reliance on specific neurons, enhancing generalization.

Based on model comparison, GRU performed better in identifying direct, explicit positive sentiment, particularly in syntactically simple reviews. However, its performance declined with informal, elongated expressions (e.g., “mantapppp”) or contrastive structures. BiLSTM, with its bidirectional design, effectively captured long-range dependencies and reversed sentiment patterns, making it more robust in identifying nuanced or implicit negative sentiment. This supports the conclusion that GRU favors recency and lexical clarity, while BiLSTM excels in processing structurally complex or noisy data.

Compared with baseline methods, both GRU and BiLSTM demonstrated superior performance. GRU achieved the best overall balance across metrics, while BiLSTM achieved the highest precision. Traditional machine learning models like SVM and Naïve Bayes had lower performance, constrained by their reliance on manual feature extraction and limited adaptability-further validating the effectiveness of deep learning approaches for sentiment analysis in low-resource, real-world Indonesian datasets.

4. CONCLUSION

This study demonstrates the effectiveness of deep sequential models-particularly BiLSTM and GRU-for sentiment analysis in Indonesian-language e-commerce reviews. This research highlights the importance of context-aware architectures in capturing linguistic nuances within user-generated content. The experimental results reveal that GRU achieves the highest performance on most evaluation measures confirming its strength in processing bidirectional contextual information. Meanwhile, the BiLSTM model showing superior sensitivity in identifying positive sentiment. BiLSTM demonstrates comparative strength in recognizing explicit positive expressions, whereas GRU excels in interpreting complex or contrastive reviews-particularly those involving implicit or negative sentiments. Additionally, dropout regularization was found to significantly influence model performance, with 0.5 being the most optimal for both architectures. When compared to classical machine learning models such as Naïve Bayes, SVM, KNN, and Random Forest, deep learning approaches demonstrated clear superiority, particularly in handling informal language and semantic shifts. The results confirm that deep sequential representations not only enhance classification performance in low-resource languages like

Indonesian but also serve as a foundational step toward building scalable, domain-specific NLP applications. In future work, we will explore cross-domain sentiment transfer techniques could further enhance model robustness when applied to varying product categories or contextual domains in Indonesian e-commerce reviews.

CONFLICT OF INTEREST

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

ACKNOWLEDGEMENT

This research is funded by Directorate of Research and Community Service, Directorate General of Research and Development, Ministry of Higher Education, Science, and Technology, the Republic of Indonesia and the Institute for Research and Community Service of Universitas Syiah Kuala (LPPM USK) under Postgraduate scheme 2025 with contract number: 113/C3/DT.05.00/PL/2025 and 188/UN11.L1/PG.01.03/DPPM/2025.

REFERENCES

[1] P. Campos, E. Pinto, and A. Torres, "Rating and perceived helpfulness in a bipartite network of online product reviews," *Electron. Commer. Res.*, vol. 25, no. 3, pp. 1607–1639, 2023, doi: 10.1007/s10660-023-09725-1.

[2] H. Md Altab, M. Yinping, H. Md Sajjad, A. Nkrumah Kofi Frimpong, M. Frempomaa Frempong, and S. Sarfo Adu-Yeboah, "Understanding Online Consumer Textual Reviews and Rating: Review Length With Moderated Multiple Regression Analysis Approach," *SAGE Open*, vol. 12, no. 2, 2022, doi: 10.1177/21582440221104806.

[3] U. Singh, A. Saraswat, H. K. Azad, K. Abhishek, and S. Shitharth, "Towards improving e-commerce customer review analysis for sentiment detection," *Sci. Rep.*, vol. 12, no. 1, pp. 1–15, 2022, doi: 10.1038/s41598-022-26432-3.

[4] K. Poonam and T. Ramakrishnudu, "Dual Bi-LSTM-GRU based stance detection in tweets ordered classes," *Neural Comput. Appl.*, vol. 37, no. 17, pp. 10439–10463, 2025, doi: 10.1007/s00521-024-10549-9.

[5] J. Sangeetha and U. Kumaran, "A hybrid optimization algorithm using BiLSTM structure for sentiment analysis," *Meas. Sensors*, vol. 25, p. 100619, 2023, doi: 10.1016/j.measen.2022.100619.

[6] Z. Yu, Y. Sun, J. Zhang, Y. Zhang, and Z. Liu, "Gated recurrent unit neural network (GRU) based on quantile regression (QR) predicts reservoir parameters through well logging data," *Front. Earth Sci.*, vol. 11, pp. 1–8, 2023, doi: 10.3389/feart.2023.1087385.

[7] A. R. Abas, I. Elhenawy, M. Zidan, and M. Othman, "Bert-cnn: A deep learning model for detecting emotions from text," *Comput. Mater. Contin.*, vol. 71, no. 2, pp. 2943–2961, 2022, doi: 10.32604/cmc.2022.021671.

[8] R. Pramana, M. Jonathan, H. S. Yani, and R. Sutoyo, "A Comparison of BiLSTM, BERT, and Ensemble Method for Emotion Recognition on Indonesian Product Reviews," *Procedia Comput. Sci.*, vol. 245, no. C, pp. 399–408, 2024, doi: 10.1016/j.procs.2024.10.266.

[9] Y. Aliyu, A. Sarlan, K. Usman Danyaro, A. S. B. A. Rahman, and M. Abdullahi, "Sentiment Analysis in Low-Resource Settings: A Comprehensive Review of Approaches, Languages, and Data Sources," *IEEE Access*, vol. 12, pp. 66883–66909, 2024, doi: 10.1109/ACCESS.2024.3398635.

[10] R. Sudheesh *et al.*, "Bidirectional encoder representations from transformers and deep learning model for analyzing smartphone-related tweets," *PeerJ Comput. Sci.*, vol. 9, 2023, doi: 10.7717/peerj-cs.1432.

[11] Z. Hameed and B. Garcia-Zapirain, "Sentiment Classification Using a Single-Layered BiLSTM Model," *IEEE Access*, vol. 8, pp. 73992–74001, 2020, doi: 10.1109/ACCESS.2020.2988550.

[12] M. K. Anam et al., "Improved performance of hybrid GRU-BiLSTM for detection emotion on

Twitter dataset," *J. Appl. Data Sci.*, vol. 6, no. 1, pp. 354–365, 2025, https://doi.org/10.47738/jads.v6i1.459.

[13] P. K. Jain, V. Saravanan, and R. Pamula, "A hybrid CNN-LSTM: A deep learning approach for consumer sentiment analysis using qualitative user-generated contents," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 20, no. 5, pp. 1–15, 2021, https://doi.org/10.1145/3457206.

[14] I. D. Mienye, T. G. Swart, and G. Obaido, "Recurrent Neural Networks: A Comprehensive Review of Architectures, Variants, and Applications," *Information*, vol. 15, no. 9, p. 517, 2024, doi: 10.3390/info15090517.

[15] R. Sutoyo, S. Achmad, A. Chowanda, E. W. Andangsari, and S. M. Isa, "PRDECT-ID: Indonesian product reviews dataset for emotions classification tasks," *Data Br.*, vol. 44, p. 108554, 2022, doi: 10.1016/j.dib.2022.108554.

[16] K. K. Putri and E. B. Setiawan, "Depression Detection in Indonesian X Social Media Text using Convolutional Neural Networks and Long Short-Term Memory with TF- IDF and FastText Methods," vol. 6, no. 2, pp. 557–574, 2025, doi: 10.52436/1.jutif.2025.6.2.4206.

[17] I. P. A. Pratama, N. Wayan, and J. Kusama, "Comparative Analysis of Gradient-Based Optimizers in Feedforward Neural Networks for Titanic Survival Prediction," vol. 6, no. 1, pp. 90–102, 2025, doi: 10.56705/ijodas.v6i1.219.

[18] A. Mao, M. Mohri, and Y. Zhong, "Cross-Entropy Loss Functions: Theoretical Analysis and Applications," *Proc. Mach. Learn. Res.*, vol. 202, pp. 23803–23828, 2023, doi: 10.48550/arXiv.2304.07288.

[19] S. Kiliçarslan and M. Celik, "RSigELU: A nonlinear activation function for deep neural networks," *Expert Syst. Appl.*, vol. 174, p. 114805, 2021, doi: https://doi.org/10.1016/j.eswa.2021.114805.

[20] S. R. Dubey, S. K. Singh, and B. B. Chaudhuri, "Activation functions in deep learning: A comprehensive survey and benchmark," *Neurocomputing*, vol. 503, pp. 92–108, 2022, doi: 10.1016/j.neucom.2022.06.111.

[21] A. M. Iddrisu, S. Mensah, F. Boafo, G. R. Yeluripati, and P. Kudjo, "A sentiment analysis framework to classify instances of sarcastic sentiments within the aviation sector," *Int. J. Inf. Manag. Data Insights*, vol. 3, no. 2, p. 100180, 2023, doi: 10.1016/j.jjimei.2023.100180.

[22] N. Nurdin, K. Kluza, M. Fitria, K. Saddami, and R. S. Utami, "Analysis of Social Media Data Using Deep Learning and NLP Method for potential use as Natural Disaster Management in Indonesia," in *2023 2nd International Conference on Computer System, Information Technology, and Electrical Engineering (COSITE)*, 2023, pp. 143–148. doi: 10.1109/COSITE60233.2023.10249849.

[23] Richard, J. R. Andres, J. P. Soetandar, R. Sutoyo, and H. Riza, "Emotion Recognition Model using Product Review from Indonesia Marketplace," in *2023 2nd International Conference on Computer System, Information Technology, and Electrical Engineering (COSITE)*, 2023, pp. 67–71. doi: 10.1109/COSITE60233.2023.10249811.

[24] R. Obiedat *et al.*, "Sentiment Analysis of Customers' Reviews Using a Hybrid Evolutionary SVM-Based Approach in an Imbalanced Data Distribution," *IEEE Access*, vol. 10, pp. 22260–22273, 2022, doi: 10.1109/ACCESS.2022.3149482.

[25] E. O. Kiyak and B. Ghasemkhani, "High-Level K-Nearest Neighbors (HLKNN): A Supervised," *Electronics*, vol. 12, pp. 1–20, 2023, doi: 10.3390/electronics12183828.

[26] C. Dewi and R.-C. Chen, "Complement Naive Bayes Classifier for Sentiment Analysis of Internet Movie Database BT - Intelligent Information and Database Systems," in *Proc. Asian Conference on Intelligent Information and Database Systems*, pp. 81–93, 2022, doi: 10.1007/978-3-031-21743-2_7.

[27] B. Jlifi, C. Abidi, and C. Duvallet, "Beyond the use of a novel Ensemble based Random Forest-BERT Model (Ens-RF-BERT) for the Sentiment Analysis of the hashtag COVID19 tweets," *Soc. Netw. Anal. Min.*, vol. 14, no. 1, p. 88, 2024, doi: 10.1007/s13278-024-01240-x.

[28] H. Liu, X. Chen, and X. Liu, "A Study of the Application of Weight Distributing Method Combining Sentiment Dictionary and TF-IDF for Text Sentiment Analysis," *IEEE Access*, vol. 10, pp. 32280–32289, 2022, doi: 10.1109/ACCESS.2022.3160172.

[29] M. S. Hossen, A. H. Jony, T. Tabassum, M. T. Islam, M. M. Rahman, and T. Khatun, "Hotel review analysis for the prediction of business using deep learning approach," in *Proc. 2021 Int. Conf. Artif. Intell. Smart Syst. (ICAIS)*, 2021, pp. 1489–1494, https://doi.org/10.1109/ICAIS50930.2021.9395757.

[30] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim, "RoBERTa-LSTM: A hybrid model for sentiment analysis with transformer and recurrent neural network," *IEEE Access*, vol. 10, pp. 21517–21525, 2022, https://doi.org/10.1109/ACCESS.2022.3152828.

[31] S. S. M. M. Rahman, K. B. M. B. Biplob, M. H. Rahman, K. Sarker, and T. Islam, "An investigation and evaluation of N-Gram, TF-IDF and ensemble methods in sentiment classification," in *Cyber Security and Computer Science: Proc. 2nd EAI Int. Conf. ICONCS*, Dhaka, Bangladesh, 2020, pp. 391–402, 10.1007/978-3-030-52856-0_31.