International Journal of Innovative Computing, Information and Control
Volume **20**, Number **3**, June **2024**

ICIC International ©2024 ISSN 1349-4198
pp. **929–944**

MODELING EMOTIONS RECOGNITION ON INDONESIAN PRODUCT REVIEW BY COMBINING BERT, CNN, AND LSTM ARCHITECTURE

Andry Chowanda[1,*], Rhio Sutoyo[1], Said Achmad[1]
Esther Widhi Andangsari[2], Sani Muhamad Isa[3] and Tin-Kai Chen[4]

[1]Computer Science Department, School of Computer Science
[2]Psychology Department, Faculty of Humanities
[3]Computer Science Department, BINUS Graduate Program – Master of Computer Science
Bina Nusantara University
Jl. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia
{ rsutoyo; said.achmad; esther }@binus.edu; sani.m.isa@binus.ac.id
[*]Corresponding author: achowanda@binus.edu

[4]Department of Comic Art
Tainan University of Technology
No. 529, Zhongzheng Road, Yongkang District, Tainan City 710302, Taiwan
t40118@mail.tut.edu.tw

Received September 2023; revised January 2024

Abstract. *The product review dataset is fast-growing and exciting data to exploit. The increasing number of Internet users and customer shopping habits through online stores significantly impact product review data growth, especially for online stores in Indonesia, such as Tokopedia. PRDECT-ID is an Indonesian language product review dataset with emotional and sentiment labels. The existence of an emotional label in the dataset makes the use of product review data more meaningful. By implementing deep learning architectures, computers can learn and recognize contextual data stored in review sentences. This study aims to obtain an effective deep-learning architecture by combining BERT-Based, Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM) for recognizing emotions through the PRDECT-ID dataset. Three stages of model exploration, layer exploration, and hyperparameter tuning experiments were implemented to display in-depth tests of the deep learning architecture used and combinations of layers and parameters suitable to obtain high emotion recognition performance. In addition, several techniques for dealing with unbalanced data were also applied to this experiment. The main contribution of this research is to present in-depth experiments related to product review datasets and to provide a deep learning architecture along with a combination of layers and parameters that have the best performance in recognizing emotions in product review datasets. The experimental results in this study show that the BERT-based deep learning architecture combined with the CNN layer has the highest performance in recognizing emotions in the PRDECT-ID dataset. It achieves 69.26% accuracy and 65.49% F1-Score.*

Keywords: Emotions classification, Natural language processing, Deep learning, BERT, Product review

1. **Introduction.** Online product reviews are the most accessible word-of-mouth information in electronic commerce [1]. They are an evaluation medium from customers to sellers regarding their products and services [2]. Moreover, positive product reviews and social content engagement are essential factors influencing customers' buying decisions

DOI: 10.24507/ijicic.20.03.929

and maintaining brand loyalties [3]. With more than 278 million people, Indonesia is the third most populous nation in Asia [4]. Furthermore, 76.3% of the Indonesian population has Internet access (around 212 million people). The increased adoption of online shopping due to the coronavirus pandemic produces massive product review data for emotion mining tasks [5]. Tokopedia online marketplace is one of the most popular online shopping in Indonesia[1]. With more than 10 million monthly active users [6] and 12 million registered merchants [7], Tokopedia is one of the market leaders in Indonesia's online shopping. Tokopedia also provides a product review feature that lets consumers share their opinions and shopping experiences on their platforms.

The popularity of the online marketplace attracts researchers to utilize its data for computational linguistics research. For instance, the research from Stephenie et al. [8] builds a sentiment analysis model with three labels: positive, negative, and neutral. It utilizes TF-IDF and machine learning classifiers for building the model. They utilized their dataset by collecting product reviews from Tokopedia. Furthermore, the research from Hadju and Jayadi [9] follows a similar approach with [8] by using three labels as the sentiment analysis class. They collected their dataset from five e-commerce in Indonesia and utilized four different machine-learning classifiers to build their model. Another research by Saputri et al. [10] resulted in an emotion recognition model using data from X (formerly known as Twitter).

Based on our findings, previous research [8, 9, 10] has not investigated emotion recognition models using Indonesian product review datasets. Two things need to be accomplished to create such emotion recognition models. First, build a dataset of Indonesian product reviews. We have done this step in previous research by building PRDECT-ID [11]. It comprises 5,400 carefully curated product reviews from Tokopedia annotated with Shaver's emotions model. Second, build an emotion recognition model using the PRDECT-ID dataset. This paper aims to answer this by proposing an effective architecture for emotion recognition tasks on the PRDECT-ID dataset by implementing various scenarios of data preprocessing, exploring BERT models, and performing hyperparameter tuning.

There are several state-of-the-art architectures to recognize emotions from text, such as using pre-trained models of transformer-based architecture. This research combines the best architectures to provide an effective architecture for emotion recognition tasks using Indonesian product reviews, i.e., PRDECT-ID dataset [11]. Currently, the PRDECT-ID dataset has an imbalanced dataset problem. Hence, this paper also presents several techniques to deal with the imbalance dataset problem. The results demonstrate that Combination 3 of Model A provides the best model among all the approaches (see Table 6). The best accuracy was achieved by the Hyper-parameter Tuning approach (69.26%); However, the best precision, recall, and F1-Score were achieved by the data augmentation approach (66.19%, 66.97%, and 66.13% for the precision, recall, and F1-Score, respectively). Although the weighted class approach did not significantly improve accuracy, it provided a balanced score for the model's precision, recall, and F1-Score.

The organization of this paper is structured as follows. The background and objectives are presented in the Introduction section. This paper summarizes previous studies on emotion recognition in product reviews in the Recent Work section. The Proposed Methods section explains the proposed research methods and experiment steps. The flow of this research is discussed in the Proposed Architectures section. Then, the Results and Discussion section presented the experimental results and our findings. Finally, the Conclusion and Future Work section provides a summary and ideas for future work.

[1]https://www.tokopedia.com/

2. **Recent Work.** This section discusses product reviews, datasets, and previous works on computational linguistic tasks. It also specifically finds related works of emotion recognition literature on product reviews. Table 1 lists the publicly available datasets from the previous works. Moreover, Table 2 presents the performance summary from the previous literature.

Table 1. Publicly available datasets from previous work

Name	Author	Year	Total label	Size	Type of data	Language
Indonesian Twitter Emotion Dataset [10]	Saputri et al.	2018	5	4,403	Tweet	Indonesian
EmotionLines Datasets (EmotionPush) [15]	Shmueli and Ku	2019	8	36,077	Dialogue	English
Tokopedia Product Online Reviews [8]	Stephenie et al.	2020	3	40,706	Product Review	Indonesian
Indonesian E-Commerce Product Reviews [9]	Hadju and Jayadi	2021	3	14,742	Product Review	Indonesian
PRDECT-ID [11]	Sutoyo et al.	2022	5	5,400	Product Review	Indonesian

Table 2. Performance summary from previous literature

Year	Corpus	Total label	Method	P	R	F1	Acc
2018	Indonesian Twitter Emotion Dataset [10]	5	Machine learning, features combination	0.700	0.680	0.680	N/A
2019	EmotionLines Datasets (EmotionPush) [15]	4	BERT, speaker, transfer, context	N/A	N/A	0.885	N/A
2020	Tokopedia Product Online Reviews [8]	3	Machine learning	N/A	N/A	N/A	0.971
2021	Indonesian E-Commerce Product Reviews [9]	3	Machine learning	N/A	N/A	N/A	0.959
2021	Emotions Recognition on Social Media Conversation [17]	4	Machine learning, deep learning	0.902	0.902	0.901	0.92

**P: Precision, R: Recall, F1: F1-Score, Acc: Accuracy*

2.1. **Product reviews.** In general, product review has two types of input: numerical (i.e., ratings of 1-5 stars) and textual (i.e., comments/posts in written text) [12]. Emotion mining from the textual content in product reviews can potentially capture customer satisfaction and emotions in greater detail. Because the textual reviews contain more comprehensive evaluative opinions than ratings of products [13]. It also often contains customers' emotions that show their feelings and perceptions toward certain situations [14].

For instance, the sentence "*pengiriman cepat, packing rapi, kualitas mantap*" (fast delivery, good packing, good quality) expresses a happy emotion from customers toward the smooth delivery process and excellent product quality provided by the sellers. And the sentence "*Barang SAMPAH ini tidak seharusnya dikirimkan ke alamat saya*" (This

GARBAGE item should not be sent to my address) shows an anger emotion toward the quality of the product.

2.2. **Datasets.** Several datasets are available for emotion recognition tasks [10, 15]. The Indonesian Twitter Emotion Dataset [10] collects 4,403 Indonesian tweets and annotates them into five classes of emotions. Moreover, the EmotionX [15] consists of 36,077 dialogues from Facebook Messenger chats. Both datasets are built for general topics and are publicly available to download. In other words, they are not product review datasets.

Other researchers have also collected product review datasets for sentiment analysis tasks [8, 9]. The Tokopedia Product Online Reviews [8] labels their data into three classes: positive, negative, and neutral. The labeling process was performed using a systematic algorithm by calculating sentiment score using the Lexicon-Based. The Indonesian E-Commerce Product Reviews [9] collects three major product categories in five e-commerce in Indonesia. The product reviews were classified into positive, negative, and neutral. However, neither dataset is publicly available to download. The documentation of the experiment is also not available. In addition, they do not use the popular emotion model for data annotation, such as Shaver et al.'s [16].

Lastly, the PRDECT-ID [11] consists of 5,400 Indonesian product reviews from Tokopedia. It has 29 product categories and is annotated with Shaver et al.'s emotion models [16]. Based on the Indonesian emotion lexicon, emotions categories in Indonesia have five clusters of the basic level: *sedih* (sadness), *takut* (fear), *marah* (anger), *senang* (happiness), and *cinta* (love). Although the number of examples is relatively small compared with [8, 9], the PRDECT-ID dataset [11] was collected selectively to ensure the data quality for each data label. Each product review was annotated with a single emotion label based on Shaver et al.'s emotion model [16].

2.3. **Previous literature.** Existing literature reviews have discussed building sentiment analysis on Indonesian product reviews [8, 9]. Moreover, other existing literature also has discussed building emotion recognition models on natural language text [10, 15, 17]. However, we have not found publications discussing building emotion recognition models for Indonesian product reviews.

2.3.1. *Emotion recognition.* Saputri et al. proposed an annotated Indonesian Twitter dataset with Shaver et al.'s basic emotion [16] later popularized by Parrott as Parrott's basic emotion [18]. It achieved 91.7% for its Kappa score. They presented their findings from their feature engineering experiment to find the best feature. Based on the report, emotion word lists cannot recognize the emotion expressed in rich-textual data [10]. Moreover, they discovered that combining features, e.g., emotional word lists, word embedding, and Bag-of-Words features, achieves better performance. In their paper, Shmueli and Ku presented the result of the 7th International Workshop on Natural Language Processing for Social Media that builds emotion models using augmented EmotionLines datasets [15]. Eleven teams submitted their technical report. The top-scoring team, IDEA, achieved a micro-F1 score of 79.5%. They utilized Twitter data for pre-training and adding context. In addition, the BERT model with a weighted approach was utilized to handle the class imbalance issue [19]. Moreover, several works have been done in recognizing emotions from the text (e.g., social media or chat apps) in other languages (e.g., English); the research of Chowanda et al. [17] trained emotions recognition model using AffectiveTweets dataset and some conventional machine learning algorithms (e.g., Naïve Bayes, Generalized Linear Model, Fast-Large Margin, Decision Tree, Random Forest, and Support Vector Machine). The best performance of the model was a 92% of accuracy score, 90.2% of recall score, 90.2% of precision score, and 90.1% of F1-Score.

2.3.2. *Sentiment analysis.* Stephenie et al. built a sentiment analysis model with the Tokopedia product online reviews dataset [8]. They proposed and compared several configurations of random forest classifiers. The best performance is achieved by configuring randomly sampled variables at each split to 73 and the number of trees to 50. The evaluation utilized 10-fold cross-validation with an average score of 97.05%. Hadju and Jayadi built their sentiment analysis model for Indonesian product reviews [9]. They proposed and compared four machine learning classifiers: decision trees, random forests, gradient boosts, and support vector machines. The support vector machine method achieves the best performance with 95.87% accuracy.

3. **Proposed Methods.** This research will employ deep learning techniques to perform emotion recognition tasks. Sentences taken from product review texts will be analyzed using deep learning techniques to identify emotional content. Deep learning will be employed on PRDECT-ID datasets and extract features relevant to the five recognized emotion labels, which allow for an automated product review classification process on the new data. The deep learning model named BERT is proposed as a method for emotion recognition in Indonesian product review datasets. The authors conducted trials on several BERT models and their combination of fine-tuning to get the best model for recognizing emotions. This experiment has three stages: pre-processing, model experiment, and evaluation. Figure 1 shows the methodology conducted in this research. The pre-processing stage aims to prepare the product review dataset before being processed by the model. In the preprocessing step, the abbreviation and slang dictionary are applied to replacing the words in the review sentence, and then data augmentation and a weighted approach are used to handle imbalanced data. The model experiment stage aims to prepare several BERT models and their fine-tuning combinations. Fine-tuning is employed by combining BERT and an additional layer of CNN or LSTM and fine-tuning its hyperparameters. Finally, the model evaluation stage is carried out to measure the model's performance that was set up in the previous stage. Several measurement metrics, such as accuracy and F1-Score macro, measure model performance.

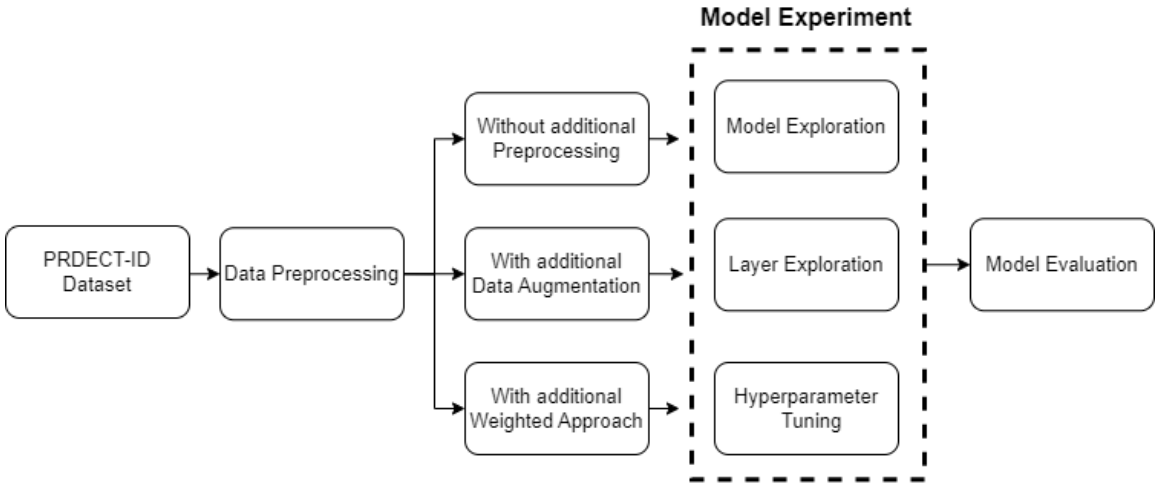


Figure 1. Methodology

3.1. **Preprocessing.** The product review data used is the PRDECT-ID dataset. The PRDECT-ID dataset contains 5,400 Indonesian product review data that have emotion labels. The emotion labels used in this dataset consist of 5 labels according to Shaver et al.'s emotional model, namely happy, love, anger, fear, and sadness. Figure 2 exhibits

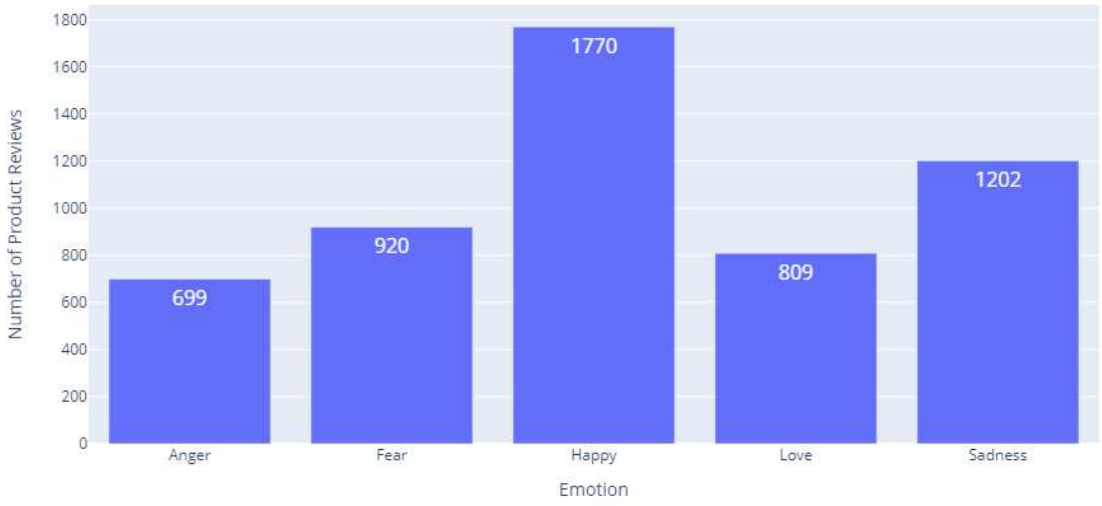


FIGURE 2. Distribution of emotion labels

the data distribution on emotion labels. Based on the proportion of the dataset, it can be seen that most of the review data have happy and sadness labels. The PRDECT-ID dataset will be divided into 80% training and 20% testing data. Based on the existing training data, the data train will be divided into 20% for evaluation data.

Pre-processing on the PRDECT-ID dataset is carried out to maximize the model's feature extraction. The PRDECT-ID dataset contains product reviews in the Indonesian language that contain many abbreviations and slang. Therefore, a word change process is needed to replace these words with more common words. The initial pre-processing stage is to remove spaces at the beginning and end of the sentence, and then change all existing letters to lowercase. The next stage is that each word in the review sentence is replaced with a more common word according to the abbreviation and slang dictionary. After the review sentence contained a better word. The next pre-processing is carried out by performing stopword removal and stemming. Figure 3 shows the word cloud of the dataset after pre-processing.

FIGURE 3. Word cloud PRDECT-ID

Based on the distribution of emotional labels, it can be seen that the dataset has an imbalanced label distribution. On the other hand, the majority of the data has a happy label. This study applies augmentation and weighting data to data train as a treatment for unbalanced datasets. First, train data that has been allocated is duplicated so that there are two train data, namely data train without augmentation and data train with

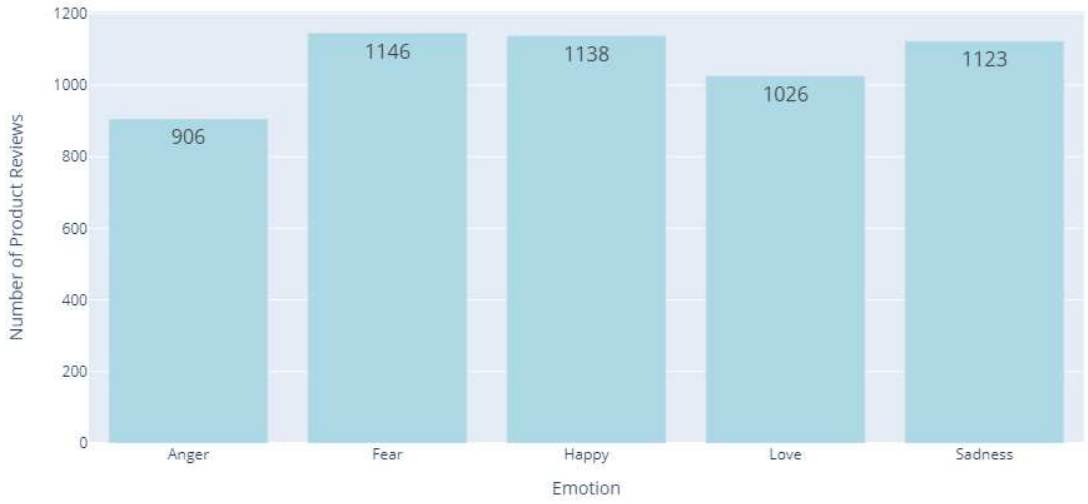


FIGURE 4. Labels distribution after augmentation

augmentation. Then, contextual augmentation is applied to multiplying the data in the minority class so that the amount of data is close to the data in the majority class. In this research, the minority class that was augmented was data with the labels of sadness, fear, love, and anger. Figure 4 shows the distribution of emotion labels after data augmentation.

Weighting is employed to generate class weights based on calculating the distribution of labels in the data. Equation (1) shows the formula for calculating the weighted class used in this research. The $Weight_{\mathrm{A}}$ represents the weight calculation for class A within the dataset. N_{A} denotes the count of data belonging to class A. $N_{\text{all sample}}$ is the overall dataset size and K is the total number of classes present in the dataset. The results of the calculations in the form of class weights will be used as additional parameters in the model that are trained using data without augmentation.

$$Weight_{\mathrm{A}} = \frac{1}{N_{\mathrm{A}}} \times \frac{N_{\text{all sample}}}{K} \tag{1}$$

3.2. **Model experiment.** The model experiment stage aims to prepare several BERT models and their fine-tuning combinations. Several Indobert models were used as part of the experimental model to determine the best-performing Indobert model for recognizing emotions in the product review dataset. This study processes the train data using several model setups, namely the indobert model without architectural modifications and the indobert model with architectural modifications, with each model setup consisting of several indobert models, namely indobert large phase 2 from [20], indobert base from [21] and indobert tweet from [22]. This study proposes BERT by adding a CNN layer and a BiLSTM layer to perform emotion recognition in the PRDECT-ID dataset.

Combining BERT with CNN and BiLSTM produces several model architectures: A and B. Model A uses the BERT architecture with the CNN layer. Model B uses the BERT architecture added with the BiLSTM layer. This study conducted three experiments to get the best model: model exploration, layer exploration, and hyperparameter tuning. First, the exploration model is an experiment with several BERT models employed without any modification to measure the base BERT performance on recognizing emotions from a product review. Then in the layer exploration stage, the three BERT models are implemented in Model A and Model B so that each model has variations of several BERT models. Next, layer exploration is a model experiment where tests on the number of layers in CNN and the number of units in the BiLSTM layer are applied to the proposed model.

Layer exploration aims to improve BERT performance by adding additional layers and trying several combinations of the number of layers and units in the proposed architecture. Then, hyperparameter tuning is a model experiment where a combination of parameters is applied to each proposed model to find the combination of parameters that produces the best performance.

Each experiment, exploration model, exploration layer, and hyperparameter tuning were run to get the best-performing model and provide an in-depth exploration of models, layers, and hyperparameters. After the model setup is done, each model will process training data without augmentation, training data with augmentation, and training data with class weighting.

3.3. **Model evaluation.** Each model will perform emotion recognition on the test data, and then the accuracy, precision, recall, and F1-Score macro values will be calculated. Since the dataset has an imbalanced proportion of class labels, as in Figure 2, where the happy and sadness emotion labels are more dominant than other labels, the best model consideration will be seen based on the F1-Score value. F1-Score combines precision and recall in its calculation. Thus, this metric considers the accuracy of predicting positives (precision) and finding all positive instances (recall). In imbalanced datasets, the balance between precision and recall is essential. The model with the highest F1-Score macro value will be said to be the best model in recognizing emotions from product review data. The model with the best performance will be analyzed further by displaying the confusion matrix.

4. **Proposed Architectures.** This study proposes models A and B to recognize emotions in an Indonesian product review. Model A consists of a combination of BERT with CNN layers. The BERT variation used in this model is BERT with Indonesian language pre-train, from [20], [21], and [22]. The number of layers added consists of 2-3 CNN layers, with the number of hidden units ranging from 32 to 128. Each CNN layer is added by one max-pooling layer and one drop-out layer. Figure 5 is a detailed description of the existing architecture in model A. Model B consists of BERT with the same Indonesian pre-train model as model A. The difference in Model B is that the added layer is a BiLSTM layer. The number of layers added consists of 1-2 BiLSTM layers, with the number of LSTM units varying from 128 to 256. Each addition of one BiLSTM layer is followed by one layer drop out. Figure 6 is a detailed description of the existing architecture in model B.

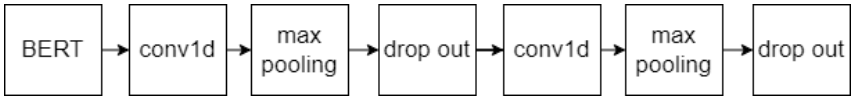


Figure 5. BERT CNN architecture

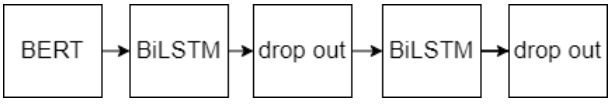


Figure 6. BERT BiLSTM architecture

5. **Results and Discussion.** This study employs three experiments to get the best model: model exploration, layer exploration, and hyperparameter tuning. The model exploration was run to measure BERT performance without any modifications in recognizing emotions from product reviews. The data train without augmentation is used in

Table 3. Model exploration result

Base model	Accuracy	Precision	Recall	F1-Score
indobenchmark/indobert-large-p2	**65.93%**	**62.24%**	**61.48%**	**65.49%**
indolem/indobert-base-uncased	62.96%	60.49%	55.57%	56.06%
indolem/indobertweet-base-uncased	64.63%	62.62%	57.63%	58.47%

the model exploration stage. This stage uses the same parameter for all BERT models: learning rate 5E-05, batch size 16, maximum length 128, and Adam optimizer. Table 3 shows the result of model exploration. From the result of model exploration, it can be seen that the indobert from indobenchmark has the highest F1-Score and accuracy. This is due to the model having a more significant number of parameters trained compared to the other models.

The next stage is layer exploration. This stage was run to improve BERT performance by adding an additional layer. The data train without augmentation is used in the layer exploration stage. Table 4 shows the combination scenario for layer exploration. Column **C** indicates the combination number, and column **M** illustrates the Model scenario. This stage uses the same parameter for all combinations: learning rate 5E-05, batch size 16,

Table 4. Layer exploration scenario

C	M	Variation	Base model	Architecture
1	A	BERT+CNN	indobenchmark/indobert-large-p2	2L CNN (128,64)
2	A	BERT+CNN	indobenchmark/indobert-large-p2	3L CNN (128,64,32)
3	A	BERT+CNN	indolem/indobert-base-uncased	2L CNN (128,64)
4	A	BERT+CNN	indolem/indobert-base-uncased	3L CNN (128,64,32)
5	A	BERT+CNN	indolem/indobertweet-base-uncased	2L CNN (128,64)
6	A	BERT+CNN	indolem/indobertweet-base-uncased	3L CNN (128,64,32)
7	B	BERT+BiLSTM	indobenchmark/indobert-large-p2	1L LSTM (128)
8	B	BERT+BiLSTM	indobenchmark/indobert-large-p2	2L LSTM (128)
9	B	BERT+BiLSTM	indolem/indobert-base-uncased	1L LSTM (128)
10	B	BERT+BiLSTM	indolem/indobert-base-uncased	2L LSTM (128)
11	B	BERT+BiLSTM	indolem/indobertweet-base-uncased	1L LSTM (128)
12	B	BERT+BiLSTM	indolem/indobertweet-base-uncased	2L LSTM (128)

Table 5. Layer exploration result

Combination	Accuracy	Precision	Recall	F1-Score
1	32.96%	86.59%	20.00%	9.91%
2	32.96%	66.59%	20.00%	9.92%
3	60.28%	59.98%	56.47%	54.83%
4	56.94%	52.83%	49.83%	47.23%
5	**68.89%**	65.72%	**65.43%**	**65.49%**
6	38.79%	80.99%	29.70%	22.68%
7	32.96%	**86.59%**	20.00%	9.92%
8	32.96%	86.59%	20.00%	9.92%
9	66.85%	62.75%	62.36%	62.39%
10	67.22%	63.17%	63.06%	62.97%
11	67.68%	64.36%	63.65%	63.93%
12	66.48%	62.72%	62.06%	62.28%

maximum length 128, and Adam optimizer. Table 5 shows the result of the layer exploration. From the result of layer exploration, it can be seen that indobert form indobenchmark got a performance drop. However, vice versa, the indobert tweet with the combination of CNN layers (Combination 5, Model A with indolem/indobertweet-base-uncased pre-trained model) has the highest F1-Score and accuracy (68.89% and 65.49% of accuracy and F1-Score respectively). The highest precision was achieved by combining BERT with LSTM layers (Combination 7, Model B, ndobenchmark/indobert-large-p2 with pre-trained model) with a precision score of 86.59%. However, this model has quite an imbalanced score of precision (86.59%) and recall (20.00%). Hence, the most stable model was achieved by the one trained with Combination 5 and Model A settings. The models trained with the indobenchmark/indobert-large-p2 pre-trained model provide a relatively high trade-off between precision and recall score.

The next stage is hyperparameter tuning. This stage was run to find the combination of parameters that produces the best performance. The data train without augmentation is used in the hyperparameter tuning stage. This stage uses the same parameter for all combinations: learning rate 5E-05 and Adam optimizer. This stage eliminates the combination of indobert from indobechmark since the model combination showed poor performance in the layer exploration stage. Table 6 shows the combination scenario for hyperparameter tuning. This stage uses the same parameter for all combinations: learning rate 5E-05 and Adam optimizer. Column **C** indicates the combination number, and column **M** illustrates the Model scenario. Moreover, Column **BS** demonstrates the Batch Size setting, and column ML indicates the Maximum Length of the input. The base models explored in this experiment are from indolem (e.g., indolem/indobert-base-uncased). Table 7 shows the result of the hyperparameter tuning. From the result of hyperparameter tuning, it can be seen that the indobert tweet with the combination of CNN layers

Table 6. Hyperparameter tuning scenario

C	M	Variation	Base model	Architecture	BS	ML
1	A	BERT+CNN	indobert-base-uncased	2L CNN (128,64)	16	128
2	A	BERT+CNN	indobert-base-uncased	2L CNN (64,32)	32	256
3	A	BERT+CNN	indobertweet-base-uncased	2L CNN (128,64)	16	128
4	A	BERT+CNN	indobertweet-base-uncased	2L CNN (64,32)	32	256
5	B	BERT+BiLSTM	indobert-base-uncased	1L LSTM (128)	32	256
6	B	BERT+BiLSTM	indobert-base-uncased	1L LSTM (256)	32	256
7	B	BERT+BiLSTM	indobertweet-base-uncased	1L LSTM (128)	32	256
8	B	BERT+BiLSTM	indobertweet-base-uncased	1L LSTM (256)	32	256

Table 7. Hyperparameter tuning result

Combination	Accuracy	Precision	Recall	F1-Score
1	66.67%	63.33%	62.55%	62.69%
2	58.56%	57.75%	54.02%	53.47%
3	**69.26%**	**65.96%**	**65.15%**	**65.49%**
4	68.52%	64.94%	63.82%	63.94%
5	65.74%	62.34%	61.36%	61.52%
6	63.79%	59.78%	58.85%	59.00%
7	66.76%	63.13%	62.19%	62.53%
8	65.28%	61.58%	60.52%	60.86%

(Model A and Combination 3) has increased performance and achieved the highest F1-Score and accuracy compared to other models in scenarios. The model achieved 69.26%, 65.96%, 65.15%, and 65.49% for accuracy, precision, recall, and F1-Score, respectively. The combination of BERT models as the features extraction and CNN as the features extraction (2 layers of 128 and 64) and classification layer improved the performances of the trained model. Moreover, the combination of BERT with indobert base pretrained model with two layers of CNN (64,32) (Model A and Combination 2) achieved the lowest performances (58.56%, 57.75%, 54.02% and 53.47% for the accuracy, precision, recall, and F1-Score respectively).

This study also applies the data augmentation and the weighting approach as a treatment for unbalanced datasets to achieve the best performance BERT model in recognizing emotions from a product review. From the scenario, in Table 6, the experiment is continued using the dataset with augmentation and the weighted approach with the same parameter for all combinations: learning rate 5E-05 and Adam optimizer. Table 8 shows the result of the training with data augmentation. The augmentation technique did not improve the models' performance. On the contrary, it can be concluded that all models got a performance drop from the data augmentation. The best model was achieved by the one trained with BERT based model with an indobert tweet pre-trained model and two layers (64 and 32) of CNN (combination 4 and Model A). The model achieved 66.11%, 64.60%, 66.05%, and 64.01% for accuracy, precision, recall, and F1-Score, respectively. Moreover, Table 9 illustrates the result of the weighted approach. In general, the weighted model approach did not significantly improve the models' performances. Model A, Combination 2 (BERT-based model with indobert pre-trained model and two layers CNN (62 and 32)) suffers from a significant performance drop (from the best accuracy score of 58.56% to 14.72%). However, the weighted model approach improves the stability of precision, recall, and hence the F1-Score of almost all models, particularly the best model

Table 8. Data augmentation result

Combination	Accuracy	Precision	Recall	F1-Score
1	58.15%	58.15%	59.55%	56.29%
2	62.78%	60.29%	60.84%	59.02%
3	66.02%	64.07%	65.79%	63.92%
4	**66.11%**	**64.60%**	**66.05%**	**64.01%**
5	55.74%	58.01%	58.42%	54.45%
6	58.61%	58.72%	60.11%	56.87%
7	62.96%	61.06%	63.00%	60.43%
8	60.28%	58.11%	59.74%	57.67%

Table 9. Weighted approach tuning result

Combination	Accuracy	Precision	Recall	F1-Score
1	62.87%	60.08%	58.51%	57.35%
2	14.72%	82.94%	20.00%	5.13%
3	**69.17%**	**66.19%**	**66.97%**	**66.13%**
4	68.06%	64.16%	63.07%	63.24%
5	65.18%	61.21%	60.57%	60.50%
6	65.28%	61.19%	60.24%	60.39%
7	66.20%	62.54%	62.17%	62.24%
8	66.76%	62.99%	62.45%	62.62%

(combination 3). The model achieved 69.17%, 66.19%, 66.97%, and 66.13% for accuracy, precision, recall, and F1-Score, respectively.

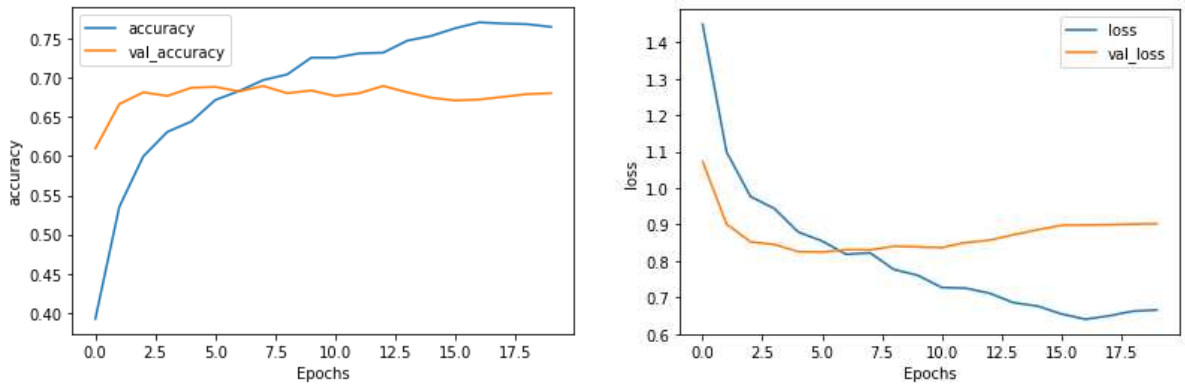
Table 10 demonstrates the overview of the best models for each approach (i.e., Layer Exploration, Hyper-parameter Tuning, Data Augmentation, and Weighted Class). Combination 3 of Model A provides the best model among all the approaches. The Hyper-parameter Tuning approach achieved the best accuracy; however, the data augmentation approach achieved the best precision, recall, and F1-Score. Although the Weighted Class approach did not significantly improve the accuracy, it provides a balanced score for the model's precision, recall, and F1-Score. In overview, Model A, BERT, with additional CNN layers, got the best performance in recognizing emotion from Indonesian product review datasets. The highest F1-Score and accuracy are achieved by using combination number 3 in Table 6. The model architecture employs two features extraction layers of CNN with 128 and 64 dense units.

TABLE 10. The best results

Approach	C	Accuracy	Precision	Recall	F1-Score
Layer Exploration	5	68.89%	65.72%	65.43%	65.49%
Hyper-parameter Tuning	3	**69.26%**	65.96%	65.15%	65.49%
Data Augmentation	4	66.11%	64.60%	66.05%	64.01%
Weighted Class	3	69.17%	**66.19%**	**66.97%**	**66.13%**

Figures 7(a) and 7(b) show the history of accuracy and loss during the training and validation phase of the model trained with the Hyper-parameter Tuning approach. Figure 9(a) shows the confusion matrix from the testing phase of the model trained with the Hyper-parameter Tuning approach. Based on the confusion matrix, it can be seen that the model has difficulty recognizing the emotions of fear and anger. This is indicated by the small number of predictions by the model for the labels fear and anger, while the emotions that should be fear and anger are more often predicted as sadness by the model. On the other hand, the model can recognize the emotions of happy, sadness, and love well. However, the model still often predicts some love emotions as happy. This can be caused by the similarity in the expression of someone who feels happy or love when writing product reviews.

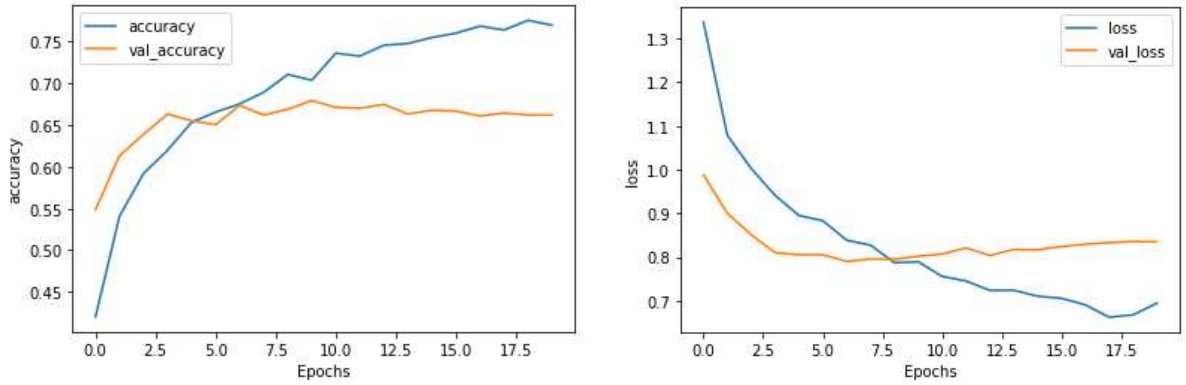
Moreover, Figures 8(a) and 8(b) show the history of accuracy and loss during the training and validation phase of the model trained with the Weighted Class approach.



(a) Training and validation accuracy

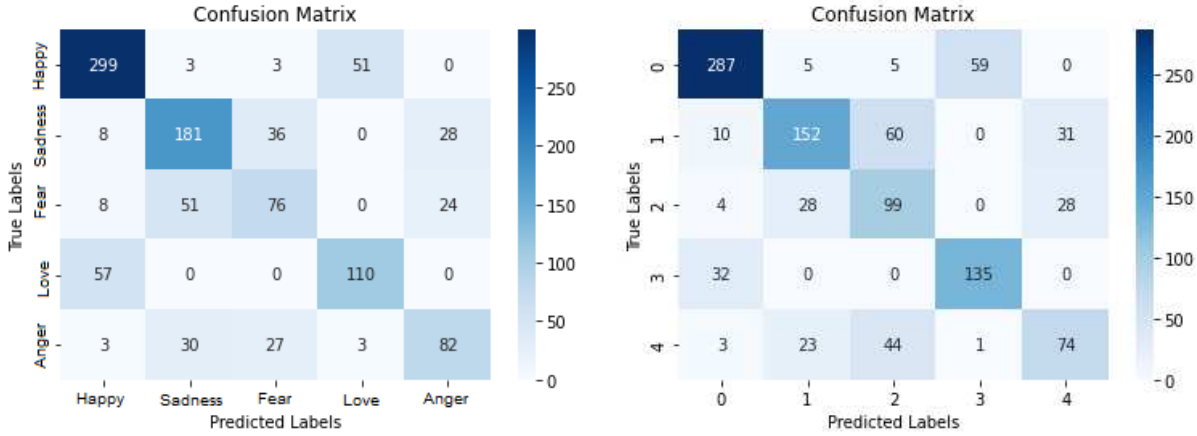
(b) Training and validation loss

FIGURE 7. Training and validation Hyper-parameter Tuning performance



(a) Training and validation accuracy (b) Training and validation loss

FIGURE 8. Training and validation weighting approach performance



(a) Hyper-parameter Tuning approach (b) Weighting approach

FIGURE 9. Confusion matrix comparison

Figure 9(b) shows the confusion matrix from the testing phase of the model trained with the Weighted Class approach. The model's accuracy and loss history graph are quite similar to the one trained with the Hyper-parameter Tuning approach. However, the number of corrected predicted of Fear (2), and Love (3) are significantly increased, while the other class is dropped. This will provide a more balanced and stable model for emotions recognition compared to the other approach (e.g., the Hyper-parameter Tuning approach). Similarly, with the Hyper-parameter Tuning approach, the model sometimes miss-predicts some Love emotions as Happy due to the similarity of expression of happiness and love in a written text (e.g., product review).

6. **Conclusion and Future Work.** A deep learning architecture for recognizing emotions from product review data in Indonesia has been explored. Various experimental scenarios have been carried out, including model exploration, layer exploration, and hyperparameter tuning. The combination of BERT and CNN layers can recognize emotions well in product review data. Unbalanced data handling scenarios are also carried out using data augmentation and weighting approaches. The experimental results show that the augmentation technique and the Weighted Class approach have not significantly increased the accuracy of the dataset used. However, the Weighted Class approach provides a more stable model with balanced precision, recall, and F1-Score. If we zoom into the detail

per class, the number of corrected predicted of Fear (2), and Love (3) are significantly increased, while the other class is dropped. This will provide a more balanced and stable model for emotions recognition compared to the other approach. However, the model sometimes miss-predicts some Love emotions as Happy due to the similarity of expression of happiness and love in a written text (e.g., product review). Moreover, the addition of the CNN layer effectively increases the F1-Score value compared to the performance of the BERT model without the additional layer. The results of this research can also be used as a deep learning performance benchmark for emotion recognition problems in the Indonesian language product review dataset. For future research, the authors suggest implementing a stacked BERT pre-trained weights model and the ensemble BERT-based model to increase emotion recognition accuracy in product review datasets.

Acknowledgment. This work is supported by Bina Nusantara University as part of "Hibah Penelitian Internasional BINUS" 2022 No: 061/VR.RTT/IV/2022. The title of the grant is "Model Klasifikasi Emosi pada Data Ulasan Produk Toko Daring di Indonesia Menggunakan Metode Machine dan Deep Learning" or "Emotions Classification Model for Online Shop Product Review in Indonesia with Machine and Deep Learning Method".

REFERENCES

[1] R. Y. Kim, When does online review matter to consumers? The effect of product quality information cues, *Electronic Commerce Research*, vol.21, no.4, pp.1011-1030, 2021.

[2] M. Malik and A. Hussain, Exploring the influential reviewer, review and product determinants for review helpfulness, *Artificial Intelligence Review*, vol.53, no.1, pp.407-427, 2020.

[3] M. Schreiner, T. Fischer and R. Riedl, Impact of content characteristics and emotion on behavioral engagement in social media: Literature review and research agenda, *Electronic Commerce Research*, vol.21, no.2, pp.329-345, 2021.

[4] I. W. Stats, *Asia Internet Usage Facebook Stats and Population Statistics*, https://www.Internet worldstats.com/stats3.htm, Accessed on Sep. 25, 2022.

[5] D. L. Warganegara and R. B. Hendijani, Factors that drive actual purchasing of groceries through e-commerce platforms during COVID-19 in Indonesia, *Sustainability*, vol.14, no.6, 3235, 2022.

[6] J. Globe, *Tokopedia Wins Best E-Commerce Award 2021*, https://jakartaglobe.id/special-updates/tokopedia-wins-best-ecommerce-award-2021, Accessed on Sep. 24, 2022.

[7] Tokopedia, *Situs Jual Beli Online Terlengkap, Mudah & Aman | Tokopedia*, https://www.tokopedia.com, Accessed on Sep. 8, 2022.

[8] Stephenie, B. Warsito and A. Prahutama, Sentiment analysis on Tokopedia product online reviews using random forest method, *E3S Web of Conferences*, vol.202, 16006, 2020.

[9] S. F. N. Hadju and R. Jayadi, Sentiment analysis of Indonesian e-commerce product reviews using support vector machine based term frequency inverse document frequency, *Journal of Theoretical and Applied Information Technology*, vol.99, no.17, 2021.

[10] M. S. Saputri, R. Mahendra and M. Adriani, Emotion classification on Indonesian Twitter dataset, *2018 International Conference on Asian Language Processing (IALP)*, pp.90-95, 2018.

[11] R. Sutoyo, S. Achmad, A. Chowanda, E. W. Andangsari and S. M. Isa, PRDECT-ID: Indonesian product reviews dataset for emotions classification tasks, *Data in Brief*, vol.44, 2022.

[12] X. Fang and J. Zhan, Sentiment analysis using product review data, *Journal of Big Data*, vol.2, no.1, pp.1-14, 2015.

[13] Z. Bao, W. Li, P. Yin and M. Chau, Examining the impact of review tag function on product evaluation and information perception of popular products, *Information Systems and e-Business Management*, vol.19, no.2, pp.517-539, 2021.

[14] P. Nandwani and R. Verma, A review on sentiment analysis and emotion detection from text, *Social Network Analysis and Mining*, vol.11, no.1, pp.1-19, 2021.

[15] B. Shmueli and L.-W. Ku, SocialNLP EmotionX 2019 challenge overview: Predicting emotions in spoken dialogues and chats, *arXiv Preprint*, arXiv: 1909.07734, 2019.

[16] P. R. Shaver, U. Murdaya and R. C. Fraley, Structure of the Indonesian emotion lexicon, *Asian Journal of Social Psychology*, vol.4, no.3, pp.201-224, 2001.

[17] A. Chowanda, R. Sutoyo, S. Tanachutiwat et al., Exploring text-based emotions recognition machine learning techniques on social media conversation, *Procedia Computer Science*, vol.179, pp.821-828, 2021.
[18] W. G. Parrott, *Emotions in Social Psychology: Essential Readings*, Psychology Press, 2001.
[19] Y.-H. Huang, S.-R. Lee, M.-Y. Ma, Y.-H. Chen, Y.-W. Yu and Y.-S. Chen, EmotionX-IDEA: Emotion BERT – An affectional model for conversation, *arXiv Preprint*, arXiv: 1908.06264, 2019.
[20] B. Wilie, K. Vincentio, G. I. Winata, S. Cahyawijaya, X. Li, Z. Y. Lim, S. Soleman, R. Mahendra, P. Fung, S. Bahar et al., IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding, *arXiv Preprint*, arXiv: 2009.05387, 2020.
[21] F. Koto, A. Rahimi, J. H. Lau and T. Baldwin, IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP, *arXiv Preprint*, arXiv: 2011.00677, 2020.
[22] F. Koto, J. H. Lau and T. Baldwin, IndoBERTweet: A pretrained language model for Indonesian Twitter with effective domain-specific vocabulary initialization, *arXiv preprint*, arXiv: 2109.04607, 2021.

Author Biography

Andry Chowanda earned his Bachelor's degree in Computer Science from Bina Nusantara University (Indonesia, 2009), a Master's degree in Business Management from BINUS Business School (Indonesia, 2011), and a Ph.D. degree in Computer Science from Nottingham University (England, 2017). He is now a Computer Science Lecturer at Bina Nusantara University. His research is in agent architecture and machine (and deep) learning. His work is mainly on how to model an agent that can sense and perceive the environment based on the perceived data and build a social relationship with the user over time. In addition, he is also interested in serious game and gamification design.

Rhio Sutoyo received his bachelor's degree in Computer Science from Bina Nusantara University, Indonesia. He graduated with a master's degree in Software Systems Engineering from the King Mongkut's University of Technology North Bangkok (KMUTNB), Thailand. During his master's degree, he took the student exchange program at Information Systems & Databases (i5) with RWTH Aachen University, Germany. Then, he continues his doctoral degree in Computer Science at Bina Nusantara University, Indonesia. He is currently a Computer Science Lecturer at Bina Nusantara University. His research interests include sentiment analysis, computational linguistics, and natural language processing. He has over a decade of experience in teaching, research, and community service in computer science. He has also received several awards and honors, including the Rector's Best Teaching Award and the 1st Winner of BINUS's Innovation Award 2023.

Said Achmad is an academic in computer science. Having experience as a software engineer, he started his academic career in 2022 by becoming a lecturer at the Computer Science Department at Bina Nusantara University. Having bachelor's degree from Lampung University at 2018 and master's degree from Bina Nusantara University at 2022 in Computer Science, he has taught various courses related to computer science, such as database, data mining, and data analytics. With a passion for computer science, he has also written several scientific publications presented at international seminars and journals. His research interests include artificial intelligence, especially natural language processing, and the application of artificial intelligence for spatial data processing. In addition, he is also active in community service activities and various self-development.

Esther Widhi Andangsari is the Head of the Psychology Department at Bina Nusantara University and has been a Faculty Member at the same university since 2007. She earned her bachelor's and master's degrees in Psychology from the University of Indonesia. She continued to pursue her doctoral degree in Psychology from Universitas Padjadjaran, Indonesia. Her research interests include emotion, personality, and human behavior related to digital technology use. She and her colleague have published numerous papers about this, such as personality prediction based on Twitter in Bahasa Indonesia, emotion recognition with digital technology, and many more. She also expresses her concern about enhancing mental health quality with her research and community service. Based on her study, she constructed her psychological measurement to detect problematic Internet use named Indonesia Problematic Internet Use Scale (IPIUS). Her dedication as a lecturer is recognized by achieving the Community Services Award (2021), Best Teaching Award (2022), and 2nd Best Lecturer Award Faculty Member Structural (2023) from Bina Nusantara University.

Sani Muhamad Isa is the Director of the Graduate Program at Bina Nusantara University. He received a bachelor's degree in Mathematics from Padjadjaran University. Then, he continues his master's degree and a Ph.D. degree in Computer Science from the University of Indonesia. During his doctoral degree, he participated in the Ph.D. Sandwich Program in computer science with Michigan State University, USA. His research interests encompass signal processing, biomedical engineering, data mining, remote sensing, and machine vision. He has authored over 100 academic papers and secured various research grants, including the prestigious HIBAH Penelitian Unggulan BINUS and HIBAH Penelitian Dasar Unggulan Perguruan Tinggi. He also has intellectual properties in ECG Compression 12-lead and Integrated Early Diagnosis and Monitoring System for Cardiac Disease (E-Cardio). Apart from his research experiences, he has worked with LEMIGAS, i.e., a research center for oil and gas technology development, designing a data integration system. He has more than five years of experience as a facilitator in the financial industry and a cloud computing certification from Alibaba Cloud.

Tin-Kai Chen has a first honor B.Sc. diploma of Electrical Engineering from Lan-Yang Institute of Technology (Taiwan, 1986), an M.Sc. in Manufacturing System Engineering from the University of Warwick (UK, 1994), and an MPhil/PhD research degree in Gesture Interface Design and Ergonomics from De Montfort University, Leicester (UK, 2009). Now he is an Assistant Professor of the Department of Comic Art, Tainan University of Technology, Taiwan. His research encompasses (1) design research, (2) ergonomics, (3) e-sensor, and (4) aging study.