International Journal of Innovative Computing, Information and Control
Volume **19**, Number **3**, June **2023**

ICIC International ©2023 ISSN 1349-4198
pp. **861–876**

EMOTIONALLY AWARE CHATBOT FOR RESPONDING TO INDONESIAN PRODUCT REVIEWS

Rhio Sutoyo[1,*], Harco Leslie Hendric Spits Warnars[1]
Sani Muhamad Isa[2] and Widodo Budiharto[3]

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

Received July 2022; revised November 2022

Abstract. *Product reviews represent the brand image of online stores. Shop owners need to respond to product reviews with negative emotions. The high velocity of product review data and shallow understanding of current auto-reply features are problems to be answered. Nevertheless, some existing studies often ignore the emotions located in product reviews. This research proposes a model to build an emotionally aware chatbot for responding to Indonesian product reviews. Six emotion recognition models are examined using two pre-trained models and a multilayer perceptron of a feedforward artificial neural network. Hyper-parameter tuning is performed on all models to find the optimum performance. This work proposes a product review dataset annotated with Shaver's emotion model and several predefined response templates for each emotion to enable answer variations. The reliability of agreement between annotators for the proposed product review dataset is substantial (κ = 0.66). The experimental results show a 1.51x performance improvement in the transfer learning model. The best testing F_1 score of the emotion recognition model is 80% from IndoBERT$_{BASE}$ + phase 2 with transfer learning. The chatbot web app was implemented using PHP scripting language and CSS library from STISLA. It was deployed with Amazon AWS cloud computing service.*
Keywords: Emotion recognition, Chatbot, Affective computing, Product reviews, Natural language processing

1. **Introduction.** Humans express emotions to show their sentiments towards certain things, such as describing the shopping experience in an online marketplace [1, 2], showing support for a preferred presidential candidate on Twitter [3, 4], or to convey their emotional intentions [5]. Previous studies classify emotions into five basic emotions based on Shaver's theory: love, happiness, anger, fear, and sadness [6]. Emotions need to be captured and understood by the sellers to understand the social signals of a given product review [7]. For instance, a product review like "*Produk tidak sampai tapi dibilang sudah sampai. Kecewa banget.*" (The product did not arrive but it said it had arrived. Very disappointed.) contains the emotion of "sadness" because the buyer expresses an awful experience with the service.

In Indonesia, the COVID-19 pandemic has developed online shopping habits in e-commerce sectors [8]. Product reviews from customers are an essential element that can influence potential buyers in the online marketplace [9]. Product review represents

DOI: 10.24507/ijicic.19.03.861

the reputation and brand image of an online store. Appropriate responses from sellers build a positive shopping experience for shoppers and potential buyers. Thus, it is crucial to identify and respond to product reviews immediately, particularly negative reviews.

Each month, online shop owners might get hundreds or thousands of customer product reviews. This condition makes it quite hard for sellers to identify the emotions in the product review and reply to the product reviews accordingly. In the current approach, online shop owners respond to product reviews independently or via the auto-reply feature. Tokopedia is one of the biggest online marketplaces in Indonesia. In Tokopedia's auto-reply feature[1], product reviews with one to three stars cannot be replied to automatically because they are considered bad reviews. The system can reply to product reviews with four to five stars with the auto-reply feature because they are considered good reviews.

The auto-reply feature has disadvantages: 1) It can only understand the product ratings, and 2) it cannot empathize with the emotions in the product reviews. Product reviews with four stars may seem incredible for online shop owners, but there is one star of disappointment from the customers. If the sellers choose to use the feature, the seller might miss the disappointed remark from the customers. An example of this case is shown in Figure 1. The seller's information is redacted from the image for privacy purposes. Although the customer is satisfied with the transaction, they complain about the product's packaging.

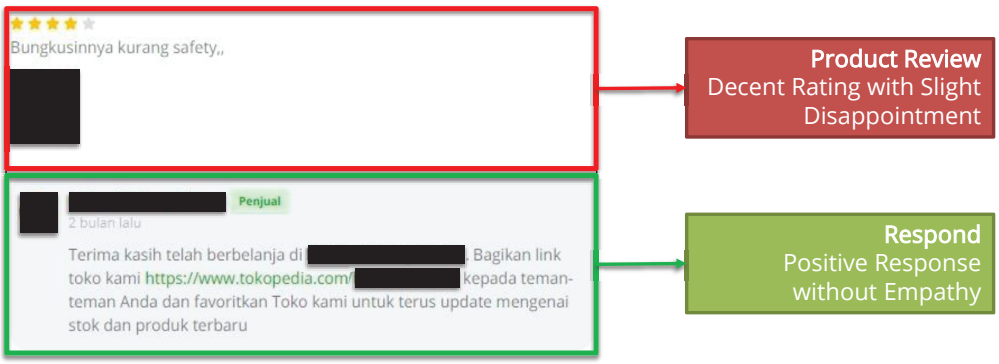


FIGURE 1. The example of product reviews with slight disappointment

A chatbot is a software application built by programmers to allow computers to communicate with humans. Currently, industries use it for simple automatic tasks such as retrieving FAQs in the QA database, answering specific questions, and product review answering engine for customer service [10, 11]. The banking industry actively promotes using chatbots amid the COVID-19 pandemic [12]. In the universities, researchers explore and build a voice chatbot application for a FAQ system [13, 14, 15]. Furthermore, chatbots can automatically identify and reply to product reviews according to the emotions in the product reviews. Implementing chatbots in the online marketplace for responding to product reviews can increase customer satisfaction and decrease company operational expenses [16].

Motivated by the research done by [11], this work proposes a model to build an emotionally aware chatbot for responding to Indonesian product reviews by utilizing the Indonesian emotion recognition model. First, the Indonesian emotion recognition model automatically captures the emotions emphasized in the product reviews written in Bahasa. Then, the emotionally aware chatbot returns a thankful reply for positive product reviews

[1]https://seller.tokopedia.com/edu/fitur-ulasan-produk

(i.e., happiness, love) and a regretful reply to mitigate negative product reviews (i.e., anger, fear, sadness).

For system evaluation, this study evaluates the emotionally aware chatbot at the emotion level [17]. Six configurations of the emotion recognition model are proposed and compared by its F_1 score result. The experiment results have shown that the transfer learning method on a pre-trained model with relatively small training data increases the model performance by 1.51x. The best testing F_1 score achieved was 80% using IndoBERT$_{\text{BASE}}$ + phase 2 with transfer learning. Furthermore, emotionally aware chatbots have demonstrated that such a system can help shop owners automatically respond to product reviews.

The rest of the paper is structured as follows. Related works are described in Section 2. The proposed model of this work is presented in Section 3. Section 4 introduces the details of experimental settings in our work. The results and discussions are given in Section 5. Lastly, Section 6 concludes the paper and explains future works.

2. Related Works.

2.1. Emotion recognition.

Emotions are part of human nature. They generally convey the meaning of the sentences. In written text, we use words (e.g., lousy, incredible, stupid), punctuation (e.g., exclamation mark, question mark), and letter case (e.g., UPPERCASE) to emphasize emotions. Thus, recognizing and understanding emotions is vital to developing an auto-reply system. Emotions can be recognized from written text using natural language processing techniques, i.e., text mining [18]. The general pipeline of emotion recognition is dataset pre-processing, training, and evaluation.

Several works of literature have discussed emotion recognition in the Indonesian language. In the early work, the researcher explores keyword-based model, and machine learning [19, 20, 21]. Then, Saputri et al. [22] achieved a 69.73% F_1 score in their dataset by implementing a hybrid approach, i.e., basic features, lexicon, POS tag, and orthographic. Later, the transformer-based model, BERT, increases the capability of computers to understand natural language in the form of written text [23]. BERT is a self-supervised pre-trained model for natural language processing tasks. It allows researchers to fine-tune and transfer learning to learn a specific task (e.g., emotion recognition). Since then, several researchers have built variants of the BERT models. For instance, the RoBERTa [24] modifies pre-training steps and adds more data points. Then another model is IndoBERT [25] which builds upon four billion words from Indonesian pre-processed text data.

2.2. Emotionally aware chatbot.

Understanding emotions is an important element for chatbots to improve user engagement [26] and positive mood [27]. An emotionally aware chatbot (EAC) can understand emotions in natural language. The EAC becomes affective computing by incorporating emotion classifiers in its architecture.

Since 2019, Tokopedia has utilized chatbots to serve customer complaints. The goal of this feature is to give responses to customers as quickly as possible. The chatbot is implemented in Tokopedia's Care system[2]. Moreover, Shopee also has its chatbot to help customer[3]. It offers various information for the customers, such as delivery status, ShopeePay, and payment status. Tokopedia and Shopee have neither implemented EAC nor emotion recognition in their product review system.

Several works have discussed chatbot technology and online reviews [10, 28, 29]. However, as far as our findings, there is still no literature focused on developing emotionally aware chatbots for responding to product reviews.

[2]https://www.tokopedia.com/help/article/dimana-saya-dapat-menghubungi-tokopedia
[3]https://chatbot.shopee.co.id

2.3. **Product review datasets.** Datasets are a critical element in machine learning and deep learning training. The Amazon Review Data [30], the Tokopedia Product Online Reviews [31], and the Indonesian E-Commerce Product Reviews [32] are datasets consisting of product reviews. The Amazon Review Data contains 233.1 million product reviews from Amazon [30]. The dataset has a great range of emotions and is suitable for training. However, it has not been annotated with a sentiment label or an emotions model. The language of the dataset is English. Tokopedia Product Online Reviews contains 40,067 product review data [31]. The language of the dataset is Indonesian. It has been annotated with sentiments (i.e., positive, negative) via lexicon-based scoring. Nevertheless, it is not publicly available. Indonesian E-Commerce Product Reviews contains 89.5 million product reviews across 18 product categories from Tokopedia [32]. The category-specific is classified via a semi-supervised technique. However, the dataset is not annotated with a sentiment label or an emotions model and is not publicly available.

3. **Proposed Model.** This work proposes a model to build an emotionally aware chatbot for responding to Indonesian product reviews (see Figure 2). There are two primary components: the Indonesian Emotion Recognition Model and the Emotionally Aware Chatbot for Responding to Indonesian Product Reviews.

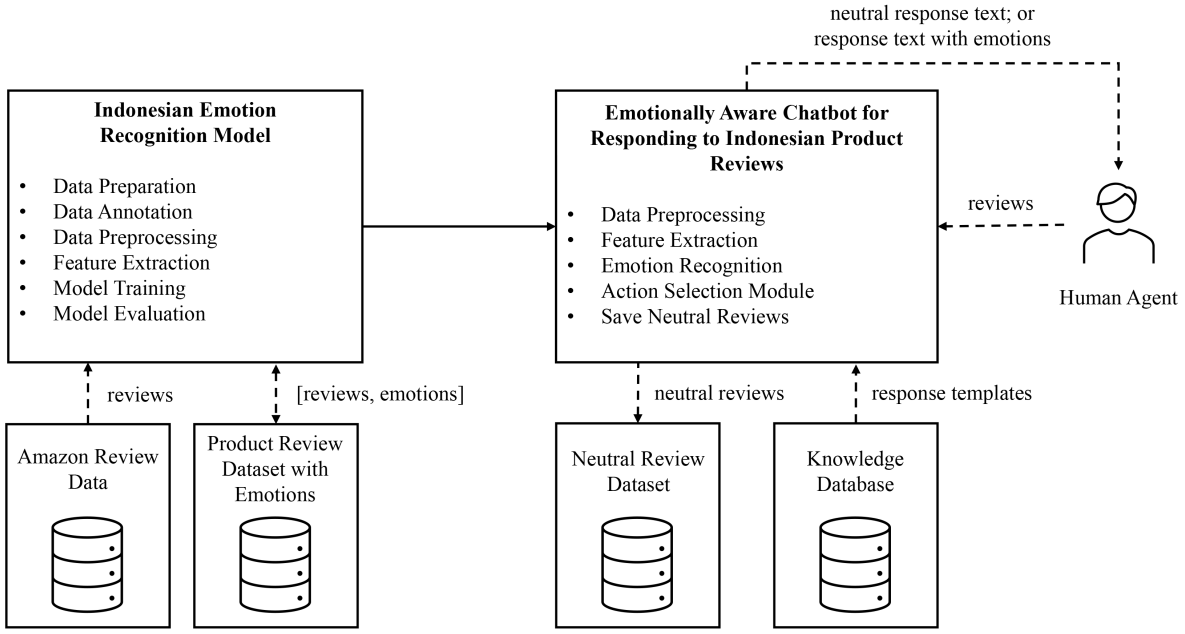


FIGURE 2. Emotionally aware chatbot for responding to Indonesian product reviews

3.1. **Indonesian emotion recognition model.** The Indonesian emotion recognition model has several steps. The data preparation step selects product reviews with an emotional expression from the Amazon Review Data. Three popular categories were chosen, that is, industrial & scientific, cell phones & accessories, and office products. Then, thirty product reviews are carefully selected for each emotion using emotions annotation criteria from an expert in clinical psychology. All product reviews are translated and validated into Indonesian by two human translators.

Next, each product review is annotated by three annotators with a single label of emotion, i.e., anger, fear, sadness, happiness, and love. The annotation process also follows the annotation guide from the expert in clinical psychology. Then, the inter-annotator

agreement (IAA) was used to measure the level of agreement between annotators. The dataset that has been translated and annotated is then stored as a product review dataset with emotions.

The data preprocessing step aims to standardize and clean the text inputs for model training. This work utilizes stop words from NLTK's library to remove low-value words. Stop words are removed because it only provides trivial meaning to the sentences. Several examples of stop words are "*adalah*" (is), "*agar*" (that), and "*akan*" (will). The model training can focus more on important words by removing stop words.

Then, the product reviews in natural language text were transformed into floating-point data using BERT. The results are contextualized word embedding for model training. For the model training, the ratio for data training is 80%. The ratio for test data and validation data is both 10%. Furthermore, this work explores different BERT models, hyperparameter tuning, and transfer learning to get the best possible performance. This work compares six configurations of the emotions model with different hyperparameter tuning and transfer learning (see Table 2).

The emotions model is evaluated using a weighted average of the precision and recall, i.e., F_1 score. The formula is shown in Equation (1). The TP is true positive, the FP is a false positive, and FN is a false negative. The test dataset is used to evaluate the model. The emotions model with the highest F_1 score is chosen for the emotionally aware chatbot.

$$F_1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{1}$$

3.2. **Emotionally aware chatbot for responding to Indonesian product reviews.** The chatbot utilizes the best emotion recognition model by accuracy to classify emotions in product reviews. Three components of our simple chatbot are intent, actions, and stories. The implementation of entities is not the scope of this work. The intent is the objective of the users. In this work, the user intends to show a specific emotion. For instance, the intent of the sentence "*Barang pecah dalam waktu dua puluh empat jam setelah saya menerima barang ini. Ini benar-benar sampah.* (The item broke within twenty-four hours after I received this item. This is really trash.)" is anger. The action defines the operation that the chatbot can perform. Our chatbot can respond to the user's emotional queries in this work. For example, *utter_happy* is an action to show a happy response for the users. The stories are the pair of identified intents and actions. For instance, a story could be about a user showing emotion of anger in a product review. Then, our chatbot responds by showing regret action. The entities are information from the product reviews (e.g., delivery, and seller). The chatbot can understand why the users are angry or happy with entities.

The human agent can input product reviews into the system via a text box on a web app. The web app is deployed using Amazon AWS. Then, the text inputs are passed into data preprocessing and feature extraction steps to prepare data inputs for the emotion recognition process. The emotion recognition process returns the predicted emotions from the given product review. There are six possible actions in our chatbot. Each action responds to a specifically identified intent. Those actions are *utter_happy* for happiness, *utter_love* for love, *utter_sorry* for sadness, *utter_soothe* for fear, *utter_regret* for anger, and *utter_neutral* for neutral.

Each intent (i.e., anger, fear, happiness, love, sadness) has five predefined response templates for answering product reviews. The predefined response templates are stored in a knowledge database. This work also prepares neutral responses to handle low-threshold confidence scores, i.e., $\leq 50\%$. If the emotions model outputs an accuracy score below the

minimum threshold, the input will be considered neutral and responded with *utter_neutral*. Then, the product review input will be saved into a neutral review dataset. The neutral product is saved in a comma-separated values (CSV) file.

4. Experimental Settings.

4.1. Datasets.

According to the findings, Indonesian product review datasets annotated with emotions are unavailable. Thus, this study decides to utilize Amazon Review Data from [30] as the source of product reviews. It contains numerous reviews that have a good range of emotions. This work process annotates the Amazon Review Data to create a product review dataset with emotions. The emotions annotation and validation process are explained in Section 4.2. The database is publicly available for research purpose[4].

Moreover, this work creates a knowledge database consisting of predefined response templates for answering product reviews. It was built by studying, modifying, and enhancing product review responses from the sellers on the Tokopedia website. To enable answer variations, this work prepares five unique response templates for each emotion and neutral ($i \in \{1, 2, 3, 4, 5, 6\}$). The dataset is also shared on the GitHub page.

4.2. Data annotations.

Initially, the Amazon Review Data [30] does not have an emotion label. This work annotates each selected product review with Shaver's emotions model [6]. Three annotators annotate the product review dataset with emotions. The 1st annotator is Ph.D., the 2nd annotator is a Master, and the 3rd annotator is a Graduate student. Each annotator was tasked to annotate each product review with a single emotion label. The annotation guide from an expert in clinical psychology was followed during the annotation process. The example of the product review annotation process from the 2nd annotator is shown in Table 1. The reviews in the first and second rows are labeled as sadness and anger, respectively.

Table 1. The illustration of the annotation process

Product review	Anger	Fear	Happiness	Love	Sadness
Saya sedikit kecewa dengan pena ini. Saya mengharapkan kualitas yang sedikit lebih baik untuk harga barang ini. (I am a little disappointed with this pen. I expected slightly better quality for the price of this item.)	0	0	0	0	1
Ini SAMPAH!!! Produk tidak akurat dan rusak setelah beberapa hari. (This is GARBAGE!!! The product is inaccurate and breaks after a few days.)	1	0	0	0	0

The reliability of agreement between annotators was calculated using Fleiss' kappa [33]. The Fleiss' kappa formula is defined in Equation (2). The $(P_o - P_e)$ gives the level of agreement achieved above chance. Moreover, the $(1 - P_e)$ gives the level of agreement a possible above chance. If the annotators are in complete agreement then $\kappa = 1$. However, if there is no agreement between the annotators, then $\kappa \leq 0$.

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{2}$$

Based on the annotation and calculation, the result of Fleiss' kappa for the product review dataset is 0.66. The kappa statistic interpretation shows that the annotators have

[4]https://github.com/rhiosutoyo/Indonesian-EAC

a substantial agreement [34]. Nevertheless, annotators have conflict labels when assigning emotions for love & happiness and anger & fear. Hence, two reviews are removed because the annotators do not have a majority agreement on a single emotion.

4.3. **Emotions model configuration.** As stated in Section 2, some existing studies have implemented term frequency – inverse document frequency (TF-IDF) and word embedding techniques to represent natural language text mathematically. This work explores the use of BERT and its transfer learning capability. The already-trained network will be used to perform the emotion recognition task. This work also explores various configurations of deep learning architectures. Each proposed architecture is fine-tuned to its best possible combination. The best performance architecture will be presented and implemented in the emotionally aware chatbot.

The emotions model configuration for finding the most optimum performance is presented in Table 2. All models use *categorical_crossentropy* as the loss function. The prepared product review dataset will be used to train and evaluate the model. For the baseline performance of our dataset, a multilayer perceptron (MLP) of feedforward artificial neural network is proposed as M_1. The rest of the model utilizes Indo RoBERTa and IndoBERT.

Table 2. The configuration of Indonesian emotion recognition model

Code	Model name	TL	Epoch	LR	BS	Optimizer
M_1	Multilayer Perceptron	No	200	0.01	5	SDG
M_2	Indo RoBERTa Emotion Classifier	No	7	0.00002	n/a	n/a
M_3	Indo RoBERTa Emotion Classifier	Yes	5	0.00005	4	AdamW
M_4	IndoBERT-lite$_{\text{BASE}}$ + phase 2	Yes	5	0.00002	2	AdamW
M_5	IndoBERT$_{\text{BASE}}$	Yes	5	0.0002	4	AdamW
M_6	IndoBERT$_{\text{BASE}}$ + phase 2	Yes	5	0.00005	2	AdamW

*TL: Transfer Learning, LR: Learning Rate, BS: Batch Size

5. **Results and Discussions.** This work utilizes Google Colab[5] for the experiments. The model training of the emotion recognition model was developed using Python v3.6, PyTorch v1.11.0, and Transformers v4.13.0. The emotionally aware chatbot was implemented using PHP scripting language and CSS library from STISLA[6]. For the web app deployment, this work utilizes an on-demand cloud computing service, i.e., Amazon AWS.

The word frequency analysis of the product review dataset with emotions shows popularly used words in the corpus. Figure 3 presents the word frequency analysis of top 20 common words excluding stop words. It shows the words "*membeli*" (buy), "*produk*" (product), "*barang*" (item), and "*bagus*" (good) have the highest frequency of occurrence. The frequent use of words is also illustrated as a word cloud (see Figure 4). Several negative words are visible, such as "*kecewa*" (disappointed), "*rusak*" (damaged), and "*sampah*" (rubbish).

This work explored six different hand-crafted configuration models to find the best tuning. The IndoBERT model used in this experiment is the BASE model with IndoBERT and IndoBERT-lite in both phase 1 and phase 2 pre-training setups. There are two main differences between the architecture of IndoBERT$_{\text{BASE}}$ and IndoBERT-lite$_{\text{BASE}}$ that is the number of parameters and the number of embedding sizes. IndoBERT$_{\text{BASE}}$ has a 124.5 M number of parameters and IndoBERT-lite$_{\text{BASE}}$ only has 11.7 M. Moreover, IndoBERT$_{\text{BASE}}$ has 768 embedding sizes and IndoBERT-lite$_{\text{BASE}}$ only has 128.

[5]http://colab.research.google.com/

[6]https://getstisla.com/

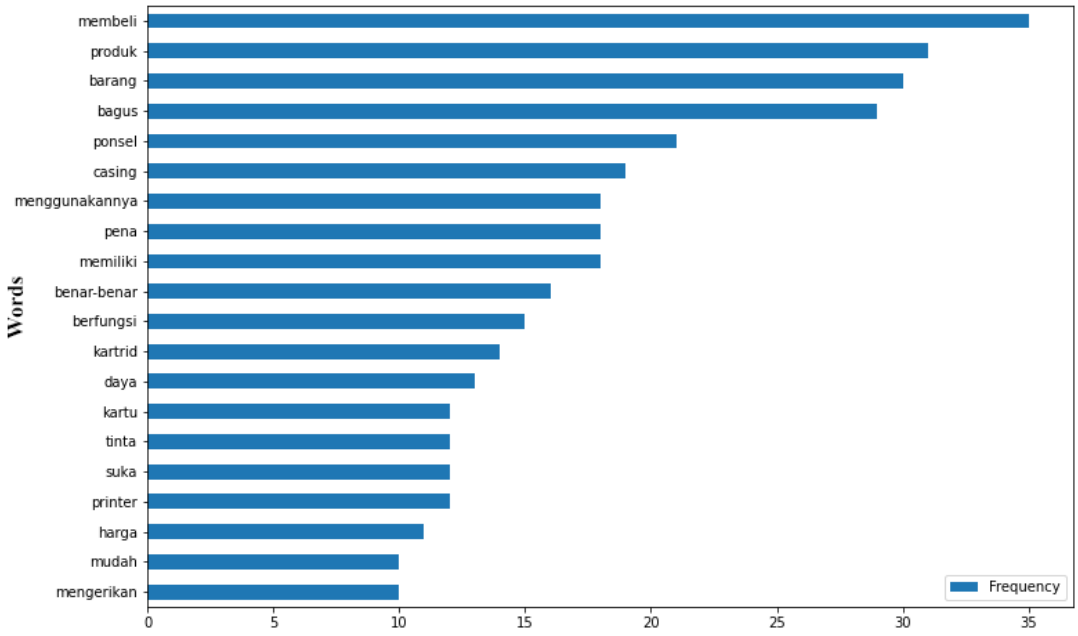


FIGURE 3. Word frequency analysis of top 20 common words (excluding stop words)

FIGURE 4. Word cloud representation of product review dataset with emotions (excluding stop words)

The IndoBERT$_{\text{BASE}}$ model was trained in two phases. In the first phase, the model of IndoBERT$_{\text{BASE}}$ and IndoBERT-lite$_{\text{BASE}}$ were trained with a maximum sequence length of 128. In the second phase, the model of IndoBERT$_{\text{BASE}}$ and IndoBERT-lite$_{\text{BASE}}$ were trained with a more considerable maximum sequence length of 512.

All configurations use the test set of product review datasets with emotions for the model evaluation. The performance is measured by its F_1 score result. The experiment results are shown in Figure 5. The model code M_1 is the baseline performance for the product review dataset with emotions. The M_1 provides 49.52% of testing F_1 score.

The F_1 score performance in the BERT model with transfer learning is significantly improved compared with that not using transfer learning. It can be seen in the result of M_2 (Indo RoBERTa Emotion Classifier without transfer learning) vs. M_3 (Indo RoBERTa Emotion Classifier with transfer learning). The M_2 model delivers 52.76% for the testing F_1 score. And the M_3 model results in an increased 1.51x of testing F_1 score from the M_2 model (i.e., 79.81%). The experiment results confirm the effectiveness of the transfer learning process with the product review dataset with emotions on the emotion recognition task.

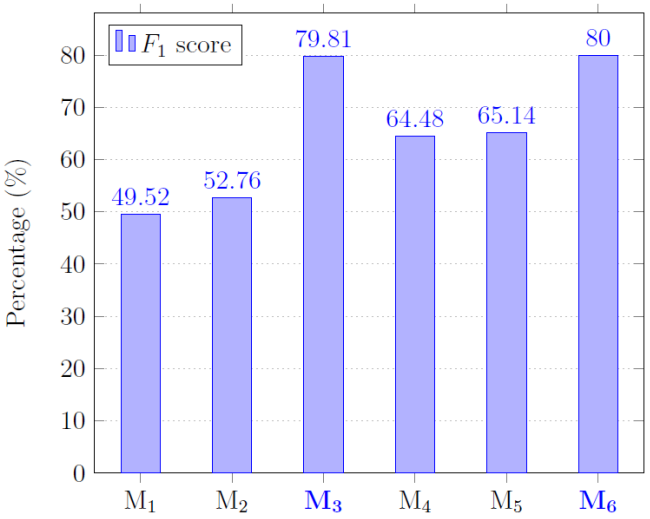


FIGURE 5. The experiment results from the emotion recognition model

The best two models from the conducted emotion recognition experiment are M_3 (Indo RoBERTa Emotion Classifier with transfer learning) and M_6 (IndoBERT$_{BASE}$ + phase 2 with transfer learning). The validation accuracy and F_1 score of those models are shown in Figure 6 and Figure 7, respectively. The model M_3 achieves high performance in the first epoch. Nevertheless, it gradually decreases in the following epochs. Then, the model reaches a plateau performance in the 5th epoch. The model M_6 performs less than M_3 in the first three epochs. Then, it performs better in the 4th epoch. Nevertheless, further training does not improve the emotion recognition model. Hence, this work stopped the training at the 5th epoch.

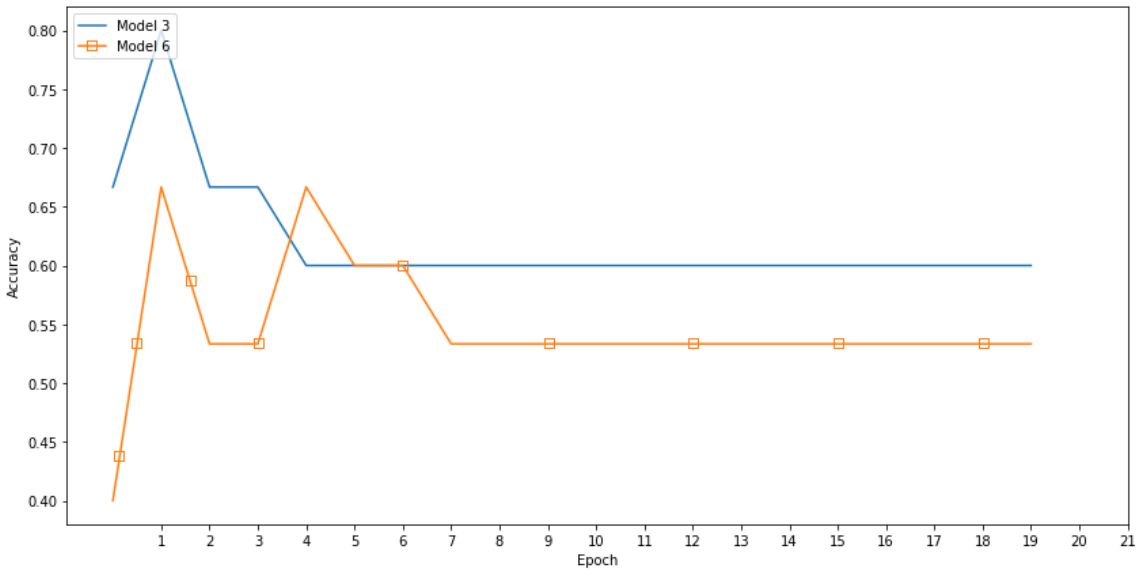


FIGURE 6. Validation accuracy of M_3 vs. M_6

The best F_1 score was achieved by the M_6 which utilizes IndoBERT$_{BASE}$ + phase 2 with transfer learning. It outperforms other IndoBERT models, namely M_4 and M_5. Hence, the M_6 model is selected to classify the emotion of product review input for the chatbot. This research trained the M_6 model with five epochs, two batch sizes, and an AdamW optimizer. It delivers 80% testing F_1 score. Moreover, the model results in 73.33%

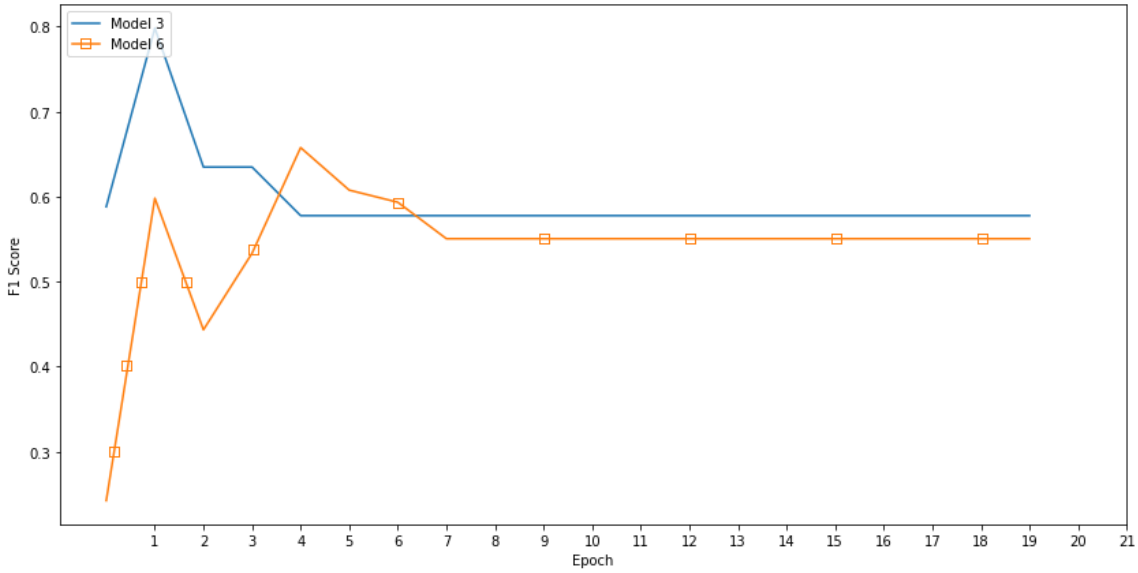


FIGURE 7. Validation F_1 score of M_3 vs. M_6

validation accuracy and 71.48% validation F_1 score. The model M_6 provides 0.004 training loss and 1.4042 validation loss.

The performance of previous work in Indonesian emotion recognition tasks is shown in Table 3. The first work from Saputri et al. [22] utilizes a hybrid approach of basic features, lexicon, POS tag, and orthographic. The second work from Wilie et al. [25] utilizes $\text{IndoBERT}_{\text{LARGE}}$ + phase 2. Finally, this work utilizes $\text{IndoBERT}_{\text{BASE}}$ + phase 2 with transfer learning and fine-tuning. The result of this work is comparable to other previous models.

TABLE 3. Performance of existing studies on Indonesian emotion recognition task

No	Dataset name	F_1 score
1	Indonesian Twitter Emotion Dataset [22]	69.73%
2	EmoT [25]	79.47%
3	Product Review Dataset with Emotions	80.00%

Figure 8 shows the confusion matrix for each class in the best model, i.e., M_6. The model can correctly classify happiness and love emotions from the product review dataset. Nevertheless, the result shows that the model has difficulty detecting anger, fear, and sadness. The model mistakes anger for fear, sadness for anger, and fear for sadness.

The emotions model is important for the chatbot to return appropriate answers. The best emotion recognition model, M_6 ($\text{IndoBERT}_{\text{BASE}}$ + phase 2), was used to classify the emotion of product review input. The emotionally aware chatbot for responding to Indonesian product reviews is defined as follows.

Given a set of emotions-response (ER) templates $ER = \{e_i, \langle r_{i1}, \ldots, r_{i5}\rangle\}_{i=1}^{n}$, where $n = 6$ and a product review (PR) from human agents, the emotion recognition returns a single emotion e_j which has the highest prediction score confidence ($j \in \{1, 2, 3, 4, 5\}$). If the confidence score is less than or equal to a predefined threshold, the model returns neutral emotion, i.e., $j = 6$. The minimum confidence score for emotion recognition in our model is 51%. Then, the action selection module returns the corresponding response templates r_{jk}, where k is randomized to enable answer variations ($k \in \{1, 2, 3, 4, 5\}$). The

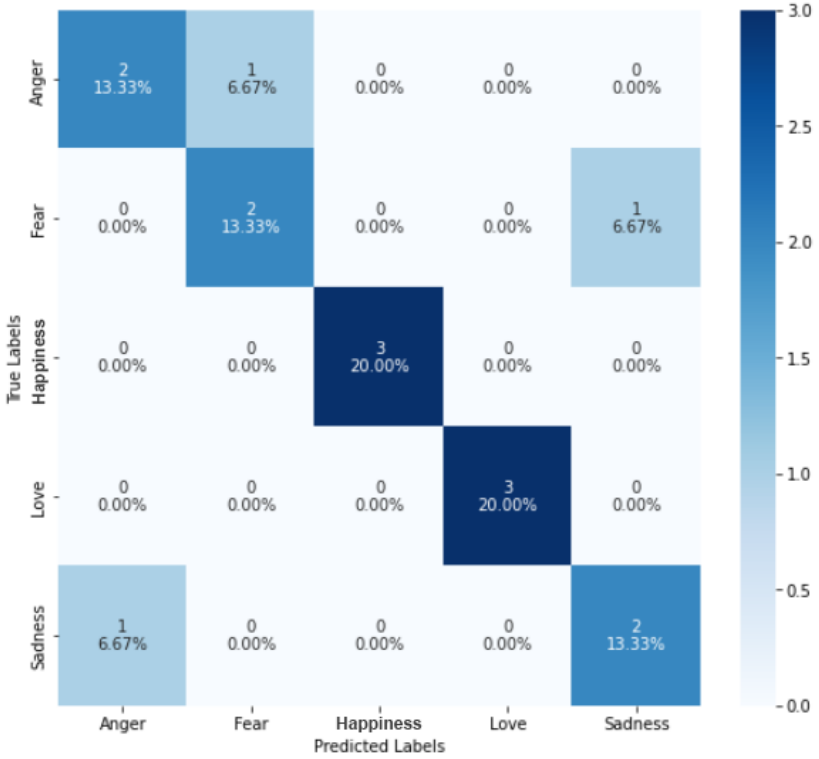


FIGURE 8. Confusion matrix of M_6 model

proposed model output is presented as follows. The emotion emphasized in the review example below is anger. Thus, the model returns *utter_regret* action to mitigate the angry sentiment from the customers.

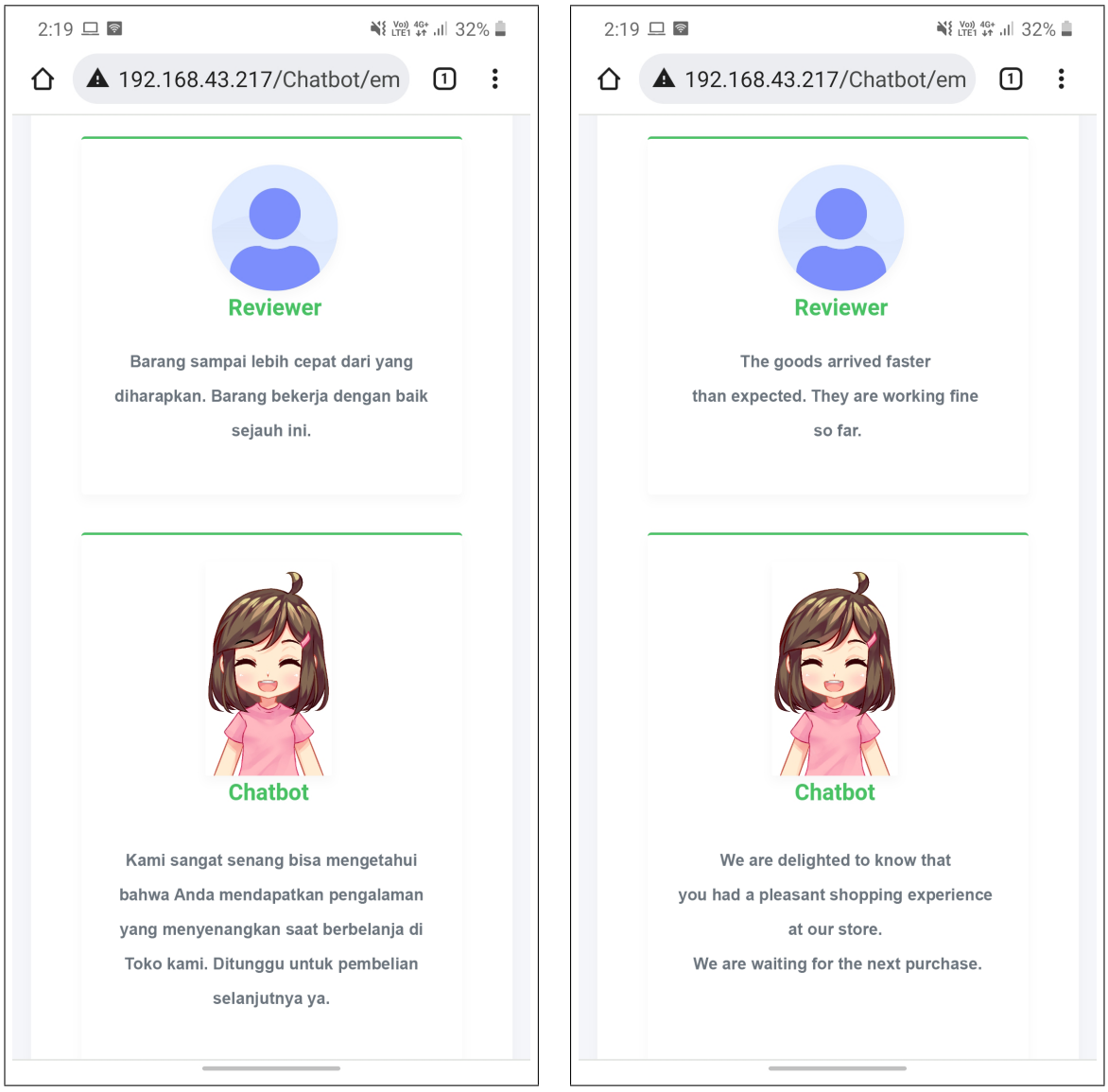
PR: *Sarung tangan ini jelek sekali.* (These gloves are so ugly.)

R: *Kami menyesal hal tersebut terjadi kepada Anda, kami akan lebih teliti lagi kedepannya sehingga hal ini tidak terulang kembali.* (We are sorry that this happened to you, we will be more careful in the future so that this does not happen again.)

The chatbot avatar is presented in Figure 9. There are four types of emotional expression: neutral, happiness, anger, and sadness (from left to right). The avatar is utilized in the chatbot interaction system to show emotions for an action. For instance, the *utter_sorry* action uses the sadness avatar, and the *utter_neutral* action uses the neutral avatar.

FIGURE 9. The chatbot avatar with the emotional expression

The chatbot interaction system has three elements: the product review from a human agent, the emotion of the chatbot represented by an avatar, and the action response from the system. The interaction system is illustrated in Figure 10.



(a) Indonesian version (b) English version

FIGURE 10. The chatbot interaction system

Moreover, the system saves the interaction logs in a custom-structured JSON file (see Listing 1). The logs were utilized to track records, monitor the interaction time, and help troubleshoot inconsistent results. The JSON file stores several main attributes, which are

1) participantId: the order number of the participant
2) startTime: the time when the interaction begins
3) endTime: the time when the conversation ends
4) interaction: the detailed interaction data for the current session
5) interactionId: the order number of the interaction made by the participant
6) input: the input sentence from the participant for the chatbot
7) classifiedEmotions: the emotion returned by the emotion recognition model
8) confidenceScore: the confidence score for each label of emotion

```
{
  "participantId": 1,
  "startTime": "2022-06-24 10:02:43",
  "endTime": "2022-06-24 10:03:25",
  "interaction": [{
     "interactionId": 1,
     "input": "Headsetnya bagus!",
     "classifiedEmotions": "happiness",
     "confidenceScore": [{
         "anger": 0.002,
         "fear": 0.001,
         "happiness": 0.99,
         "love": 0.003,
         "sadness": 0.005
     }]
  }]
}
```

LISTING 1. The JSON file structure to store interaction logs

6. **Conclusions and Future Work.** This research proposed an emotionally aware chatbot for responding to Indonesian product reviews. Furthermore, this research created a Product Review Dataset with Emotions and Knowledge Database consisting of predefined response templates for answering product reviews. Both items are shared publicly on the previously mentioned GitHub page for research purposes. The kappa statistic interpretation for the product review dataset shows that the annotators have a substantial agreement ($\kappa = 0.66$).

The authors did the emotionally aware chatbot's evaluation at the emotional level by using the F_1 score. The best performance of the emotions model results from exploring six configurations of deep learning. Moreover, this work also compared the result to two other existing studies on the Indonesian emotion recognition task. Although the proposed dataset for transfer learning is considerably small, the results have pointed out that reusing a pre-trained model as a base point is very effective on emotion recognition tasks. The F_1 score was increased by 1.51x compared with the model that utilizes transfer learning. The best model is achieved by M_6 (IndoBERT$_{\text{BASE}}$ + phase 2) with 80% of testing F_1 score and the worst performance model is achieved by M_1 (Multilayer Perceptron) with 49.52% of testing F_1 score. The online marketplace could implement the emotions model in this study in marketplaces to improve its product reviews feature. It allows the shop owners to capture product reviews that might only contain slight disappointment.

The emotionally aware chatbot utilizes Indonesian emotion recognition model to understand the given product reviews from human agents. Emphasized emotions in product reviews are treated as the users' objectives. The chatbot that performs actions and responds based on the users' objective created in this work can be implemented to help shop owners automatically respond to product reviews with negative sentiments (i.e., fear, anger, sadness) and positive sentiments (i.e., happiness, love).

The effect of the emotions model on chatbots' believability element is an interesting take for the future work of this study. Adding affective elements, various response templates, and chatbot avatars in a chatbot system may increase its believability element.

Acknowledgment. The author would like to thank Ms. Agnes Kurniati for the chatbot avatar.

REFERENCES

[1] A. Zablocki, K. Makri and M. J. Houston, Emotions within online reviews and their influence on product attitudes in Austria, USA and Thailand, *Journal of Interactive Marketing*, vol.46, pp.20-39, 2019.

[2] M. M. D. Putra, W. F. Al Maki and A. Romadhony, Sentiment analysis on marketplace review using hybrid lexicon and SVM method, *2021 9th International Conference on Information and Communication Technology (ICoICT)*, pp.66-70, 2021.

[3] W. Budiharto and M. Meiliana, Prediction and analysis of Indonesia presidential election from Twitter using sentiment analysis, *Journal of Big Data*, vol.5, no.1, pp.1-10, 2018.

[4] R. H. Ali, G. Pinto, E. Lawrie and E. J. Linstead, A large-scale sentiment analysis of tweets pertaining to the 2020 US presidential election, *Journal of Big Data*, vol.9, no.1, pp.1-12, 2022.

[5] X. Wang, X. Hao and K. Wang, Facial expression recognition based on multi-branch adaptive squeeze and excitation residual network, *International Journal of Innovative Computing, Information and Control*, vol.17, no.3, pp.735-751, 2021.

[6] 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.

[7] A. Chowanda, Separable convolutional neural networks for facial expressions recognition, *Journal of Big Data*, vol.8, no.1, pp.1-17, 2021.

[8] D. N. Larasati, U. Bustaman and S. Pramana, Online marketplace data to figure COVID-19 impact on micro and small retailers in Indonesia, *Indonesian Journal of Statistics and Its Applications*, vol.5, no.2, pp.333-342, 2021.

[9] R. Bose, R. K. Dey, S. Roy and D. Sarddar, Sentiment analysis on online product reviews, *Information and Communication Technology for Sustainable Development*, pp.559-569, 2020.

[10] L. Cui, S. Huang, F. Wei, C. Tan, C. Duan and M. Zhou, SuperAgent: A customer service chatbot for e-commerce websites, *Proc. of ACL 2017, System Demonstrations*, pp.97-102, 2017.

[11] A. Janssen, D. Rodríguez Cardona and M. H. Breitner, More than FAQ! Chatbot taxonomy for business-to-business customer services, in *Chatbot Research and Design. CONVERSATIONS 2020. Lecture Notes in Computer Science*, Springer, 2021.

[12] J. A. Mulyono et al., Evaluation of customer satisfaction on Indonesian banking chatbot services during the COVID-19 pandemic, *CommIT (Communication and Information Technology) Journal*, vol.16, no.1, pp.69-85, 2022.

[13] A. Huddar, C. Bysani, C. Suchak, U. D. Kolekar and K. Upadhyaya, Dexter the college FAQ chatbot, *2020 International Conference on Convergence to Digital World – Quo Vadis (ICCDW)*, pp.1-5, 2020.

[14] M. Hendronoto and A. Wicaksana, Implementation of ALBERT for text mining on Jacob voice chatbot, *ICIC Express Letters*, vol.15, no.10, pp.1029-1036, 2021.

[15] B. Richardson and A. Wicaksana, Comparison of IndoBERT-lite and RoBERTa in text mining for Indonesian language question answering application, *International Journal of Innovative Computing, Information, and Control*, vol.18, no.6, pp.1719-1734, 2022.

[16] A. Wiliam, S. Sasmoko, H. Prabowo et al., Analysis of e-service chatbot and satisfaction of banking customers in Indonesia, *Asia Proceedings of Social Sciences*, vol.4, no.3, pp.72-75, 2019.

[17] H. Zhou, M. Huang, T. Zhang, X. Zhu and B. Liu, Emotional chatting machine: Emotional conversation generation with internal and external memory, *Proceedings of the AAAI Conference on Artificial Intelligence*, vol.32, no.1, 2018.

[18] B. Siswanto, F. L. Gaol, B. Soewito and H. L. H. S. Warnars, Sentiment analysis of big cities on the island of Java in Indonesia from Twitter data as a recommender system, *2021 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, pp.124-128, 2021.

[19] K. E. Purnama et al., Classification of emotions in Indonesian texts using k-NN method, *International Journal of Information and Electronics Engineering*, vol.2, no.6, pp.899-903, 2012.

[20] N. A. S. Winarsih, C. Supriyanto et al., Evaluation of classification methods for Indonesian text emotion detection, *2016 International Seminar on Application for Technology of Information and Communication (ISemantic)*, pp.130-133, 2016.

[21] R. M. Cahyaningtyas, R. Kusumaningrum, D. E. Riyanto et al., Emotion detection of tweets in Indonesian language using LDA and expression symbol conversion, *2017 1st International Conference on Informatics and Computational Sciences (ICICoS)*, pp.253-258, 2017.

[22] 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.

[23] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, MN, pp.4171-4186, https://aclanthology.org/N19-1423, 2019.

[24] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, *arXiv Preprint*, arXiv: 1907.11692, 2019.

[25] 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, *Proc. 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*, 2020.

[26] E. W. Pamungkas, Emotionally-aware chatbots: A survey, *arXiv Preprint*, arXiv: 1906.09774, 2019.

[27] A. Ghandeharioun, D. McDuff, M. Czerwinski and K. Rowan, Towards understanding emotional intelligence for behavior change chatbots, *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp.8-14, 2019.

[28] A. V. D. Sano, T. D. Imanuel, M. I. Calista, H. Nindito and A. R. Condrobimo, The application of AGNES algorithm to optimize knowledge base for tourism chatbot, *2018 International Conference on Information Management and Technology (ICIMTech)*, pp.65-68, 2018.

[29] E. Adamopoulou and L. Moussiades, An overview of chatbot technology, *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pp.373-383, 2020.

[30] J. Ni, J. Li and J. McAuley, Justifying recommendations using distantly-labeled reviews and fine-grained aspects, *Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLPIJCNLP)*, pp.188-197, 2019.

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

[32] M. Sun, M. S. Leo, E. Munawwar, P. C. Condylis, S.-y. Kong, S. P. Lee, A. Hidayat and M. D. Kerianto, Semi-supervised category-specific review tagging on Indonesian e-commerce product reviews, *Proc. of the 3rd Workshop on e-Commerce and NLP*, pp.59-63, 2020.

[33] R. Falotico and P. Quatto, Fleiss' kappa statistic without paradoxes, *Quality & Quantity*, vol.49, no.2, pp.463-470, 2015.

[34] M. L. McHugh, Interrater reliability: The kappa statistic, *Biochemia Medica*, vol.22, no.3, pp.276-282, 2012.

Author Biography

Rhio Sutoyo received his bachelor's degree in Computer Science from Bina Nusantara University. 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 study, Rhio took the student exchange program at Information Systems & Databases (i5) with RWTH Aachen University, Germany. He is pursuing his doctoral degree in Computer Science at Bina Nusantara University. His research interests include sentiment analysis, computational linguistics, and natural language processing.

Harco Leslie Hendric Spits Warnars is Head of Concentration of Information Systems at Doctor of Computer Science (DCS) Bina Nusantara University and supervisor of some Ph.D. students in Computer Science. He did bachelor's degree in Computer Science in the Information Systems field from STMIK Budi Luhur, Jakarta Selatan, Indonesia with the title S.Kom (Sarjana Komputer) between 1991-1995 with a bachelor thesis topic about information systems using Budi Luhur scholarship. Between 2004-2006 he continued his master's degree in Computer Science with a major field Information Technology at the University of Indonesia, with a degree titled M.T.I. (Magister Teknologi Informasi) with a master thesis topic about datawarehouse which was funded by Budi Luhur university. His Ph.D. Computer Science was done at the Manchester Metropolitan University, Manchester, United Kingdom, with a Ph.D. thesis topic about data mining between 2008-2012.

Sani Muhamad Isa is the Director Graduate Program at Bina Nusantara University. He received a bachelor's degree in Mathematics from Padjadjaran University. Then, he continued his master's degree and a Ph.D. in Computer Science from the University of Indonesia. Sani took the Ph.D. Sandwich Program in computer science with Michigan State University, USA. His research interests include signal processing, biomedical engineering, data mining, remote sensing, and machine vision. Sani has published over 100 academic papers and achieved various research grants, such as Penelitian Desentralisasi (PTUPT) and Penelitian Kompetitif Nasional (PSN Institusi). 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. Sani has more than five years of experience as a facilitator in the financial industry and a cloud computing certification from Alibaba Cloud.

Widodo Budiharto received a bachelor's degree in Physics from the University of Indonesia, Jakarta, Indonesia, a master's degree in Information Technology from STT Benarif, Jakarta, Indonesia, and a Ph.D. degree in Electrical Engineering from the Institute of Technology Sepuluh Nopember, Surabaya, Indonesia. He took the Ph.D. Sandwich Program in robotics with Kumamoto University, Japan, and conducted Postdoctoral Researcher work in robotics and artificial intelligence with Hosei University, Japan. He worked as a Visiting Professor with the Erasmus Mundus French Indonesian Consortium (FICEM), France, Hosei University, Japan, and the Erasmus Mundus Scholar with the EU Universite de Bourgogne, France, in 2017, in 2016, in 2007, respectively. He is currently a professor of artificial intelligence at the School of Computer Science, Bina Nusantara University, Jakarta, Indonesia. His research interests include intelligent systems, data science, robot vision, and computational intelligence.