Bulletin of Electrical Engineering and Informatics
Vol. 13, No. 5, October 2024, pp. 3579~3589
ISSN: 2302-9285, DOI: 10.11591/eei.v13i5.8032

ABSA of Indonesian customer reviews using IndoBERT: single-sentence and sentence-pair classification approaches

Evi Yulianti, Nuzulul Khairu Nissa
Department of Computer Science, Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia

Article Info

Article history:

Received Dec 15, 2023
Revised Mar 8, 2024
Accepted Mar 29, 2024

Keywords:

Aspect-based sentiment analysis
Customer review
Indobert
Sentence-pair classification
Single-sentence classification
Transformer

ABSTRACT

Aspect-based sentiment analysis (ABSA) task is important to identify user satisfaction from customer reviews by recognizing the sentiments of all aspects discussed in the reviews. This work investigates a novel study on the effectiveness and efficiency of three IndoBERT-based models for solving the ABSA task in Indonesian language. IndoBERT is a state-of-the-art transformer-based model, i.e., bidirectional encoder representations from transformers (BERT), that was pre-trained on Indonesian language. Our first model utilizes IndoBERT in a feature-based mode, paired with the convolutional neural network (CNN) and machine learning models, for single-sentence classification. Next, our second model is obtained by fine-tuning the IndoBERT model for a typical single-sentence classification to build an end-to-end model. At last, our third model also adopts a fine-tuning approach to use IndoBERT, but for sentence-pair classification by utilizing auxiliary sentences. Our results demonstrate that the third model, the fine-tuned IndoBERT for sentence-pair classification, gains the highest effectiveness. It demonstrates significant improvement over deep learning baselines (Word2Vec-CNN-XGBoost) by 23.6% and transformer-based baselines (mBERT-aux-NLIB) by 2.2% in terms of F-1 score. When considering both effectiveness and efficiency, the results show that the best-performing model is our second model, the fine-tuned IndoBERT for single-sentence classification.

Corresponding Author:

Evi Yulianti
Department of Computer Science, Faculty of Computer Science, Universitas Indonesia
Jl. Prof. DR. Sudjono D. Pusponegoro, Kecamatan Beji, Depok, Jawa Barat 16425, Indonesia
Email: evi.y@cs.ui.ac.id

1. INTRODUCTION

Customer reviews are opinions of customers regarding their experience using a particular product or service. A valuable customer review can help companies to measure satisfaction and improve their products or services. Reviews from customers can contain information about several categories or aspects of a product or service, and each aspect can have different sentiment or polarity. Analyzing the sentiments based on aspects across all reviews is a challenging and time-consuming task. An automatic system that can effectively and efficiently predicts the sentiment of aspects contained in the review is highly needed.

Aspect-based sentiment analysis (ABSA), is a task that aims to extract sentiment information based on the aspect categories contained in a text [1]. According to Pontiki *et al.* [2], there are four general tasks in ABSA: aspect term extraction, aspect term polarity detection, aspect category detection, and aspect category polarity detection. If our goal is to classify sentiments based on each aspect in a review text, then the task is aspect category polarity detection (which is also the main task of ABSA). Some methods have been explored

in previous work to solve the ABSA task, including rule-based [3], lexicon-based [4], conventional machine learning [5]–[8], and deep learning [9]–[11] models.

Piryani *et al.* [3] devised a linguistic rule-based approach that identifies the aspects from movie reviews, locates opinions about that aspect, and computes the sentiment polarity of those opinions. Next, Mowlaei *et al.* [4] applied a lexicon generation method using genetic algorithm for aspect-level sentiment analysis. Rule-based and lexicon-based methods are considered less flexible because they rely heavily on predefined rules. It relies on the assumption that the text's semantic orientation is strictly related to the polarity of words and phrases that occur in it [12]. This limitation is overcome with machine learning and neural network models, because they can automatically learn the pattern from the given dataset without human-defined rules.

Bachtiar *et al.* [5] used machine learning models, such as support vector machine (SVM) and Naive Bayes, to classify the sentiment based on aspects of the guest house reviews dataset, and reported a superior performance of the SVM model. Later, Azhar *et al.* [9] and Amalia and Winarko [13] utilized convolutional neural network (CNN) deep learning model as feature extractor in their ABSA systems. While Azhar combined CNN feature extraction with Word2Vec embedding and machine learning classifiers, Amalia and Winarko [13] combined the CNN module with IndoBERT embedding. CNN is utilized because of its effectiveness and capability to capture local correlations to model sentences. CNN emphasizes features at different sentence positions through convolutional filters and pooling.

Recently, pre-trained language models such as bidirectional encoder representations from transformers (BERT) [14] has gained popularity in various NLP tasks, including ABSA, due to its remarkable performance. The BERT model can be an option to improve the semantic representations of text because this model produces contextualized embedding, which can better capture contextual relationships between words. BERT employs the transformer [15] architecture, which has a bidirectional pre-training mechanism that can learn text from both directions. With this mechanism, the embeddings produced by BERT have been shown to be more accurate for various tasks [14], [16], [17]. BERT leverages pre-trained knowledge and can be adapted to specific domains. Therefore, it is potential to use BERT to improve the accuracy of an ABSA system. Table 1 summarizes related research in ABSA task using BERT pre-trained language model. Several studies have utilized BERT using feature-based approach [13], [18]–[21], and fine-tuning approach with a typical single-sentence type input [22]–[24] or sentence-pair input [25]–[32].

Table 1. Some previous work on ABSA task using BERT-based models

Work	BERT strategy	Classification type	Model	Language
Amalia and Winarko [13]	Feature-based	Single-sentence	IndoBERT-CNN	Indonesian
Wahyudi and Sibaroni [18]	Feature-based	Single-sentence	IndoBERT-LSTM	Indonesian
Sirisha and Chandana [19]	Feature-based	Single-sentence	RoBERTa-LSTM	English
Mewada and Dewang [20]	Feature-based	Single-sentence	BERT-XGBoost	English
Nuha and Lin [21]	Feature-based	Single-sentence	BERT-MLP	English
Bahri and Suadaa [22]	Fine-tuning	Single-sentence	IndoBERT	Indonesia
Said and Manik [23]	Fine-tuning	Single-sentence	IndoBERT, RoBERTa	Indonesian
Tiwari *et al.* [24]	Fine-tuning	Single-sentence	BERT, DeBERTa	English
Azhar and Khodra [25]	Fine-tuning	Sentence-pair	mBERT (NLI-B)	Indonesian
Wu and Ong [26]	Fine-tuning	Sentence-pair	BERT (NLI-M)	English
Li *et al.* [27]	Fine-tuning	Sentence-pair	BERT (NLI-M, QA-M)	English
Sun *et al.* [28]	Fine-tuning	Sentence-pair	BERT (NLI-M, NLI-B, QA-M, QA-B)	English
Pathak *et al.* [29]	Fine-tuning	Sentence-pair	mBERT (NLI-M, NLI-B, QA-M, QA-B)	India
Ours	Feature-based	Single-sentence	IndoBERT-CNN-XGBoost/RndForest/NB	Indonesian
	Fine-tuning	Single-sentence	IndoBERT	
	Fine-tuning	Sentence-pair	IndoBERT (NLI-M, NLI-B)	

In this work, we want to examine the effectiveness and efficiency of three variants of models utilizing IndoBERT [33], [34], a BERT-based model pretrained using a large corpus of Indonesian language, to solve the ABSA task on Indonesian customer reviews. This is a novel study to compare the effectiveness and efficiency of three different approaches utilizing IndoBERT: i) feature-based approach with a typical sentence classification; ii) fine-tuning approach with a typical sentence classification; and iii) fine-tuning approach using auxiliary sentence with a sentence-pair classification. This study has not been conducted in any previous work on ABSA. Our first and second models are also our novel methods that respectively depart from the work in [9] and [25], and therefore aimed to improve the methods used in these prior work. More detailed explanation about a contrast between our models and previous works' models are explained in the following subsections.

Our first model attempts to enhance Azhar *et al.*'s method [9] by improving the way to represent a text more accurately by changing Word2Vec into IndoBERT embedding to tackle the drawback of Word2Vec as a static embedding. Azhar *et al.* [9] proposed a model that combines CNN and XGBoost to perform ABSA task using Word2Vec as the text embedding. Our first model utilizes IndoBERT in a feature-based mode, replacing

the Word2Vec component of Azhar *et al.*'s [9] method. This model has been studied in [17] for aspect category detection, while in this work it is used for aspect category polarity detection.

Our second model is obtained by fine-tuning the IndoBERT model using a typical single-sentence classification to perform ABSA task. While a few recent works on ABSA task in Indonesia have investigated this approach [22], [23], but they used presidential election tweets and tourist destination reviews dataset. Similar to [17], this work also employs a hotel reviews dataset, but our subtask of ABSA are different to them, as mentioned in the previous paragraph.

Our third model attempts to improve a recent work by Azhar and Khodra [25] that utilized sentence-pair classification in the fine-tuning process. In this work, the mBERT language model of Azhar and Khodra's method is replaced with IndoBERT. Here, we hypothesize that IndoBERT could better capture the semantic meanings of Indonesian text which could potentially increase the accuracy of ABSA system since IndoBERT was pre-trained on a large and varied Indonesian collections. Note that none of the previous work has investigated the use of IndoBERT model for sentence-pair classification to solve ABSA task. In addition, while Azhar and Khodra [25] only examined their model for binary classification, we investigate both types of sentence-pair classification tasks: binary and multiclass classification.

In summary, the contribution of this study is three-folds. First, this work performs a novel comparative evaluation on the effectiveness and efficiency of three different models utilizing IndoBERT for ABSA task. Second, this study enhances the feature-based method used in [9] by replacing the Word2Vec with BERT embedding. Third, this study enhance the fine-tuning method using auxiliary sentences used in [25] by replacing the mBERT with IndoBERT language model, and also investigates the multiclass classification in addition to the binary classification that was experimented in that previous work.

The rest of the paper is organized as follows. Section 2 describes three models utilizing IndoBERT for ABSA task that are examined in this study. Sections 3 and 4 detail our experiment together with the results and analysis. Section 5 discuss our main findings in this study, the comparison between our findings and the results reported in previous work, and some possible avenues for future work. At last, section 6 concludes this study.

2. METHOD

In this work, three variants of models utilizing IndoBERT are proposed to perform ABSA task in Indonesian reviews dataset. This study uses two data input formats, including single sentence and sentence pair formats. If the ABSA task is formulated as a single-sentence classification, where a sentence is a review text to be classified, then the input data is a single sentence. Meanwhile, if the ABSA task is formulated as sentence-pair classification, then the input data is in the format of sentence pairs. Here, the first sentence is the actual review text, while the second sentence is an auxiliary sentence that is generated using the aspect information.

Before the text review is processed by the models, a data pre-processing step is needed to prepare input text data for BERT model. It includes lowercasing, tokenization, and padding processes. For our third method, it also includes an additional step of generating auxiliary sentences. Detailed information about auxiliary sentences is explained later in section 3.1. After completing the data pre-processing step, the next step is adjusting the input representation to the IndoBERT input format.

2.1. Model 1: IndoBERT-CNN-machine learning (single sentence classification)

In Model 1, IndoBERT is used for text representation to generate text features from the review text. CNN is then used for feature refinement before the features are actually inputted to the classifier. Finally, the ABSA task is performed by a machine learning classifier by applying single sentence classification (because the input for the classifier is a text review, which is a single sentence). Initially, the review sentence will be fed and pre-processed using BERT Tokenizer, namely WordPiece. A classification token (*[CLS]*) is a special token that is placed in the first position of the BERT input, which is then followed by a sequence of sentence tokens (*C*). A separator token (*[SEP]*) is a special token that separates between a pair of input sentences, but since Model 1 uses single-sentence (as opposed to sentence-pair) input, then this token is placed at the end of the BERT input. The maximum token length for data input is 128. The formulation of the input sequence of tokens from the BERT model using the single sentence classification can be written (1):

$$S = ([CLS], C, [SEP]) \quad (1)$$

To get the input representation, a summation process is then carried out from token, positional, and segment embeddings. The resulting token representation will be embedded into the embedding layer of the IndoBERT model using a feature extraction strategy. The embedding used in this model follows the setting in the original paper of IndoBERT [33], which has a dimension of 768 with a vocabulary size of 30,522. Therefore, we obtain an embedding matrix size of 30,522x768. Each token is represented as an embedding

vector w_i, where $w_i \in R^d$, and d denotes the dimensions of each word embedding. Each sentence is represented as a matrix of word token representations in the sentence (2):

$$S = w_{1:m} = w_1 \oplus w_2 \oplus \ ... \ \oplus w_m \quad (2)$$

where m is the maximum length of the input sentence. So, the input representation of a sentence is the matrix of $m \times d$. The vector representation is used as a feature to train the CNN model. The next step is to replace the output layer of the CNN model (which has been trained) with a machine learning classifier, such as: RandomForest, NaiveBayes, and XGBoost. The text features derived from the CNN model are then used to train the top-level machine learning classifier to perform the ABSA task.

2.2. Model 2: fine-tuned IndoBERT (single-sentence classification)

The pre-processing steps to generate input representation for Model 2 are similar to the ones for Model 1. After the input representation is obtained in single sentence format, the next step is to input the representation result to the IndoBERT encoders for fine-tuning process. The fine-tuning of a BERT-based model is performed by adding one additional layer (output layer) after the final BERT layer and training the entire network for just a few epochs. The architecture of the IndoBERT model is the same as the BERT model, which consists of 12 encoder layers, where each encoder layer consists of a multi-head self-attention sub-layer and a fully connected sub-layer. The architecture of Model 2 using a single sentence classification approach can be seen in Figure 1.

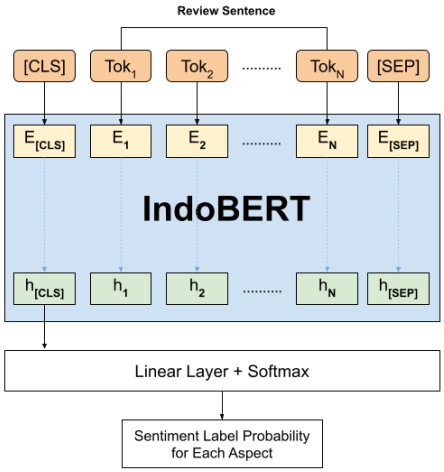

Figure 1. Fine-tuned IndoBERT model using single-sentence classification approach

In the fine-tuning process, we add an output layer (i.e., linear layer + softmax) as the final layer in the IndoBERT architecture to enable the model to perform the ABSA task. We train the network using our specific ABSA dataset which causes the initial model weights of the pre-trained IndoBERT model to be updated to more closely portray the characteristics of ABSA data (while the weights of the output layer that was just added are learned from scratch). After the fine-tuning process is complete, the final vector output for each token will be obtained. However, only the final vector output or final hidden state of the $[CLS]$ token will be used. The final hidden state of the $[CLS]$ token is considered as a fixed-dimensional pooled representation of the entire input sequence that has been processed using IndoBERT. The vector can be denoted by $h_{[CLS]} \in R^H$, where H is the size of the hidden state. Next, the $h_{[CLS]}$ vector will be fed to a linear layer with the parameter matrix $W \in R^{K \times H}$, where K denotes the number of labels. In the last step, the softmax function $P = softmax(h_{[CLS]}W^T)$ is used to calculate the probability for each sentiment label based on each aspect.

2.3. Model 3: fine-tuned IndoBERT + auxiliary sentence (sentence-pair classification)

The pre-processing steps to generate input representation for Model 3 are also similar to the ones for Models 1 and 2. The difference is only in the input format. Model 3 accepts sentence pairs (i.e., review text and auxiliary sentence) and performs sentence-pair classification. The mechanism for constructing the auxiliary sentence is by creating pseudo-sentences and adding them to the input of the BERT model. Examples of sentence pairs input are described in section 3.1.

The architecture of Model 3 is illustrated in Figure 2. We can see that the difference between Figures 1 and 2 are only in terms of the input formats. The formulation for the input sequence of tokens using the sentence pair classification approach can be written (3).

$$S = ([CLS], C, [SEP], A, [SEP]) \quad (3)$$

where C denotes the customer review text, A denotes the auxiliary sentence, and the explanation of *[CLS]* and *[SEP]* tokens are similar to that described in section 2.1 earlier. Similar to Model 2, the resulting input representation is fed to the IndoBERT model for fine-tuning. An output layer (i.e., linear layer + softmax) is also added as the final layer in the IndoBERT architecture. While Model 2 applies a single-sentence classification, Model 3 applies a sentence-pair classification.

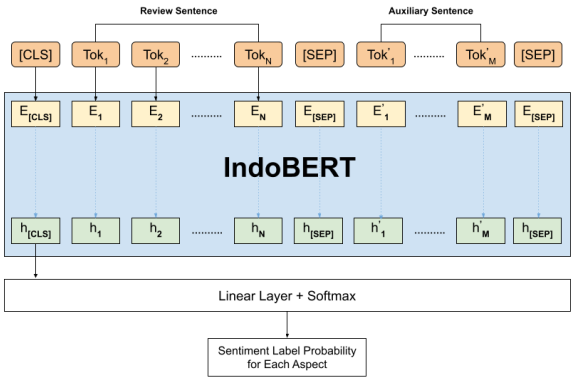

Figure 2. Fine-tuned IndoBERT model using sentence-pair classification approach

3. EXPERIMENT

3.1. Dataset

We evaluate our models on Indonesian hotel customer reviews dataset [9], [25]. This dataset was collected from the aggregator platform of AiryRooms. This dataset contains 2,854 reviews, divided as follows: 2,283 for training data, 286 for testing data, and 285 for validation data. A review text has sentiment labels for each aspect. Table 2 presents the labels distribution in our dataset.

Table 2. The statistics of our dataset

Aspect	Train data			Validation data			Test data		
	Neg	Neut	Pos	Neg	Neut	Pos	Neg	Neut	Pos
AC	419	1,821	52	52	226	7	45	233	8
Air panas (hot water)	336	1,930	26	48	230	7	38	245	3
Bau (smell)	360	1,920	12	37	247	1	41	244	1
General	30	2,028	234	2	253	30	5	243	38
Kebersihan (cleanliness)	729	1,352	211	104	157	24	96	158	32
Linen	609	1,617	66	82	190	13	80	197	9
Service	387	1,655	250	42	211	32	43	207	36
Sunrise meal (breakfast)	101	2,116	75	15	261	9	14	264	8
Tv	196	2,083	13	28	254	3	26	257	3
Wifi	330	1,937	25	39	244	2	35	250	1

For Model 3, we use two types of auxiliary sentences, i.e., natural language inference-multiclass (NLI-M) and natural language inference-binary (NLI-B), which determine the type of sentence-pair classification task to be conducted. In NLI-M type, the task is multiclass classification, in which each sentence pair of <*review, auxiliary sentence*> will be classified into one of the sentiment labels: positive, neutral, or negative. On the other hand in NLI-B type, the task is binary classification, in which each sentence pair input will be classified into 0 or 1.

The mechanism for generating the NLI-M type includes constructing simple pseudo-sentences which consist of aspect categories. Here, one review text will be transformed into *N* pairs of <*review, auxiliary sentence*> data, where *N* denotes the number of aspect categories in our dataset. For example, there are 10 aspect categories in the hotel reviews dataset. Therefore to generate the sentence pairs data with NLI-M type, each review text will be transformed into 10 pairs of <*review, auxiliary sentence*> data. The examples of sentence pairs data for NLI-M type, together with the sentiment labels, can be seen in Table 3.

Table 3. Examples of <*review, auxiliary sentence*> pair with NLI-M type for a particular review

Review	Auxiliary sentence	Sentiment
tempat nya nyaman bersih, cuma AC nya kurang dingin (the place is comfortable and clean, but the AC is not cold enough)	ac	neg
tempat nya nyaman bersih, cuma AC nya kurang dingin (the place is comfortable and clean, but the AC is not cold enough)	hot_water	neut

The mechanism for generating the NLI-B type includes constructing pseudo-sentences that consist of the concatenation of aspect categories and their polarities. Here, one review text will be transformed into $N \times K$ pairs of <*review, auxiliary sentence*> data, where N denotes the number of aspect categories in our dataset and K denotes the number of sentiment labels. For example, there are 10 aspect categories and 3 sentiment labels (i.e., positive, neutral, negative) in our hotel reviews dataset. Therefore to generate the sentence pairs data with NLI-B type, each review text will be transformed into 10×3=30 pairs of<*review, auxsentence*> data. The examples of sentence pairs data for NLI-B type, together with the sentiment labels, can be seen in Table 4.

Table 4. Examples of <*review, auxiliary sentence*> pair with NLI-B type for a particular review

Review	Auxiliary sentence	Sentiment
tempat nya nyaman bersih, cuma AC nya kurang dingin (the place is comfortable and clean, but the AC is not cold enough)	ac-positive	0
tempat nya nyaman bersih, cuma AC nya kurang dingin (the place is comfortable and clean, but the AC is not cold enough)	ac-negative	1
tempat nya nyaman bersih, cuma AC nya kurang dingin (the place is comfortable and clean, but the AC is not cold enough)	ac-neutral	0
tempat nya nyaman bersih, cuma AC nya kurang dingin (the place is comfortable and clean, but the AC is not cold enough)	ac-positive	0
tempat nya nyaman bersih, cuma AC nya kurang dingin (the place is comfortable and clean, but the AC is not cold enough)	ac-negative	0
*tempat nya nyaman bersih, cuma AC nya kurang dingin (*the place is comfortable and clean, but the AC is not cold enough)	ac-neutral	1

3.2. Setup

Our systems' architecture are developed using Python 3.7 programming language and the Pytorch library. Our experiments are run using a 1-core NVidia Tesla T4 GPU. For Model 1, IndoBERT-CNN-machine learning, the variants of machine learning classifiers used are random forest, Naive Bayes, and XGBoost. The other parameters follow the parameters used in [25]: learning rate of 1e-3, number of CNN model filters of 128, dimensions of CNN of 400, maximum input length CNN of 180, window size CNN of [3]-[5], loss function of cross entropy loss, optimizer of Adam and dropout rate of 0.5. For Models 2 and 3, the parameters used follow the recommendation parameters from the study in [14], [25], such as the learning rate=2e-5, epoch=4, batch size=32, maximum sequence length=128, loss function=Cross Entropy Loss, optimizer=Adam, and dropout rate=0.1.

3.3. Baseline systems

We employ several baseline methods that come from previous work to test the effectiveness of our models. They include two methods using feature-based approach, two methods using fine-tuning approach with a typical single-sentence input, and two methods using fine-tuning approach with a sentence-pair input. The two baselines that use feature-based approach are Word2Vec-CNN-XGBoost [9] and IndoBERT-CNN [13]. The two baselines that use fine-tuning approach with single-sentence classification are mBERT and distilBERT. At last, the two baselines that use fine-tuning approach with sentence-pair classification are mBERT-aux-NLIB [25] and mBERT-aux-NLIM [29]. Note that our main baselines in this study are the method from Azhar *et al.* [9] which is Word2Vec-CNN-XGBoost, and the method from Azhar and Khodra [25] which is mBERT-aux-NLIB, because recall that our novel methods are built upon these previous methods that are enhanced using IndoBERT in this study.

3.4. Evaluation

We use classification accuracy and F1-score to measure the performance of our model. They are common metrics that were used in many previous works on ABSA task. Accuracy is the ratio of correctly predicted labels to the total number of the actual testing dataset. F1-Score is a harmonic mean of precision and recall.

4. RESULTS

This section presents the results and discussion of our experiments using our models described previously in section 2 and a range of baselines described in section 3.3. We denote the version of IndoBERT that was pre-trained by Wilie *et al.* [33] using approximately 4B words as *IndoBERT*^{W}, while the version pre-trained by Koto *et al.* [34] using around 220M words as *IndoBERT*^{K}. Table 5 shows the performance and efficiency results of our proposed models on hotel review dataset, and Table 6 describes the results of the statistical significance test for these results. The significance test is measured based on the McNemarTest (p<0.05). Symbols a, b, c, d, e, and f respectively denote Word2Vec-CNN-XGBoost [9], IndoBERT^{W}-CNN [13], mBERT-aux-NLIM [29], mBERT-aux-NLIB [25], mBERT, and DistilBERT baselines.

Table 5. The effectiveness and efficiency of our models

Type	Model	Accuracy	F-1 score	Testing time (s)
Baseline	Word2Vec-CNN-XGBoost [9]	0.9273	0.7436	0.06
	IndoBERT^{W}-CNN [13]	0.9556	0.8906	2.72
	mBERT-aux-NLIM [29]	0.9510	0.8965	78.03
	mBERT-aux-NLIB [25]	0.9587	0.8995	120.02
	mBERT	0.9276	0.7207	3.89
	distilBERT	0.8885	0.6931	3.50
Model 1	IndoBERT^{W}-CNN-XGBoost	0.9367	0.8008	**0.03**
	IndoBERT^{W}-CNN-RandomForest	0.9402	0.8138	0.16
	IndoBERT^{W}-CNN-NaiveBayes	0.9294	0.7763	0.05
Model 2	IndoBERT^{W}	0.9580	0.8840	2.81
	IndoBERT^{K}	0.9238	0.7337	3.78
Model 3	IndoBERT^{W}-aux-NLIM	0.9612	0.9129	51.23
	IndoBERT^{K}-aux-NLIM	0.9486	0.8935	66.02
	IndoBERT^{W}-aux-NLIB	**0.9636**	0.9111	87.04
	IndoBERT^{K}-aux-NLIB	0.9633	**0.9194**	108.03

Table 6. The significant differences of our models against baseline systems

Type	Model	Accuracy is significantly better against	F-1 Score is significantly better against
Model 1	IndoBERT^{W}-CNN-XGBoost	a,e,f	a,e,f
	IndoBERT^{W}-CNN-RandomForest	a,e,f	a,e,f
	IndoBERT^{W}-CNN-NaiveBayes	f	f
Model 2	IndoBERT^{W}	a,b,e,f	a,e,f
	IndoBERT^{K}	f	f
Model 3	IndoBERT^{W}-aux-NLIM	a,b,e,f	a,b,e,f
	IndoBERT^{K}-aux-NLIM	a,e,f	a,b,e,f
	IndoBERT^{W}-aux-NLIB	a,b,c,e,f	a,b,c,e,f
	IndoBERT^{K}-aux-NLIB	a,b,c,d,e,f	a,b,c,d,e,f

It can be seen that all combinations in Model 1 can show significant improvement over the baseline model from Azhar *et al.* [9] that uses Word2Vec as text embedding (Word2Vec-CNN-XGBoost). More specifically, *IndoBERT*^{W}*-CNN-RandomForest* can improve the accuracy and F-1 scores of the *Word2Vec-CNN-XGBoost* baseline by 1.39% and 9.44%, respectively. Next, Model 2 (Fine-tuned IndoBERT using single-sentence classification) is shown to further improve the effectiveness of Model 1. The range of improvement is 1.9-3.1% for accuracy and 8.6-13.9% for F-1 scores. The *IndoBERT*^{W} model outperforms the *Word2Vec-CNN-XGBoost* [9] significantly, and it performs comparably well with the *IndoBERT*^{W}*-CNN* [13], *mBERT-aux-NLIB* [9], and *mBERT-aux-NLIM* [29] baseline models. Then, Model 3 (Fine-tuned Indo-BERT using sentence pair classification) in NLIB version also denotes an improved performance over Model 2. The *IndoBERT*^{W}*-aux-NLIB* model is demonstrated to be the most effective model, outperforming all baseline models. It improves the F-1 scores of the *Word2Vec-CNN-XGBoost* [9], *IndoBERT-CNN* [13], *mBERT-aux-NLIB* [25], and *mBERT-aux-NLIM* [29] baseline models respectively by 22.5%, 2.3%, 1.3%, and 1.6%. in terms of F-1 scores. The NLIB version also appears to be more effective than the NLIM version.

Next, the efficiency analysis is performed based on execution time on the testing data. Based on the results presented in Table 4, it is found that the longest testing time occurs on the *mBERT-aux-NLIB* model [25], and the shortest execution time (testing) is generally found in the models based on machine learning. The models using auxiliary sentences (for sentence-pair classification) consistently require significantly higher execution times compared to those without auxiliary sentences.

4.1. Confusion matrix analysis

This analysis is conducted to better understand the performance of our model for each sentiment class. We choose to analyze the confusion matrix of the *IndoBERT*^{W}*-aux-NLIB* model (see Figure 3), since it is the best-performing model according to our experiment results presented in the previous section.

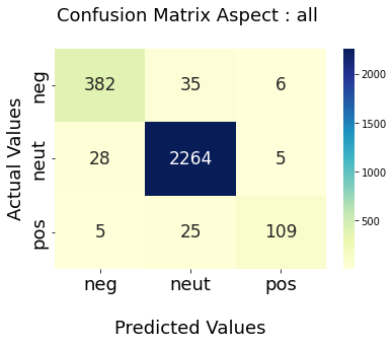

Figure 3. Confusion matrix of IndoBERT^{W}-aux-NLIB model

For each sentiment, the number of true positive (TP) cases is greater than the number of false positive (FP) and false negative (FN) cases. For example, in neutral sentiment, the total TP is 2,264, the total FP is 60, and the total FN is 33. This indicates that in most cases, the model can classify correctly the sentiments for all aspects. From the figure, it also appears that the model is very good at classifying 'neutral' sentiments. It could be because the amount of 'neutral' sentiment in the training data is more considerable when compared to other sentiment labels, therefore it is easier for the model to identify neutral sentiments.

We analyze the confusion matrix of the *IndoBERT^{W}-aux-NLIB* model for each aspect, and found that the model shows the best performance in predicting the sentiments of the 'wifi' aspect. This can happen because customers will explicitly mention 'wifi' when they want to comment about the wifi condition in the hotel. As a result, a text review discusses the 'wifi' aspect tends to contain 'wifi' term that is followed with a direct simple adjective term describing the wifi. This makes the model easier to identify the wifi aspect and its sentiment from an input review, resulting in more accurate predictions. For example: "The hotel is good, the food is good enough, the wifi is uneven." In this review example, the label 'wifi' is mentioned explicitly. For the sake of brevity, the confusion matrix of the model for each aspect is not displayed in the paper.

5. DISCUSSION

The first finding in this study is that our first model using IndoBERT in feature-based approach, i.e., *IndoBERT^{W}-CNN-XGBoost*, *IndoBERT^{W}-CNN-RandomForest*, *IndoBERT^{W}-CNN-NaiveBayes*, can improve the baseline system *Word2Vec-CNN-XGBoost* from Azhar *et al.* [9]. It demonstrates that the text representation generated by IndoBERT is proven to be more accurate than Word2Vec. This is because Word2Vec produces a static representation of words without taking into consideration the context provided by the surrounding words. Therefore, the same words with different meanings will be assigned the same representation. On the other hand, IndoBERT can produce contextualized text representation, which can better capture the semantics of words. This finding is consistent to the result obtained by Yulianti *et al.* [16] who found that IndoBERT is more effective than Word2Vec, but for text summarization using TextRank and unweighted word embedding.

The superiority of our second model, fine-tuned IndoBERT using single-sentence classification, over our first model is the second finding int this work. This case can be explained because, in the fine-tuning process, the model weights will be updated based on our specific data. It enables the model to learn the new data patterns and new specific tasks, so the model performance results will increase. This finding confirms the results in [14], [17] that reported the use of IndoBERT using fine-tuning approach gains higher performance than feature-based approach. Our experimental results also point out that a pre-trained IndoBERT version from Wilie *et al.* [33] is more effective than that from Koto *et al.* [34] as the former included much higher number of Indonesian corpus in the pre-training process. The third finding from our results is the outperformance of our third model, fine-tuned IndoBERT using sentence-pair classification, against our second model and the superiority of NLIB version over NLIM version in our third model. The good performance of Model 3 is influenced by the mechanism for adding auxiliary sentences to the input data [28] that enables the model to be benefited from the next sentence prediction (NSP) task performed in the BERT pre-training process.

The efficiency resuls of all of our models become the fourth finding in this work. Although Model 1 is the least effective, but it is shown to be the most efficient. On the other hand, although Model 3 is the most effective, but it is shown to be the least efficient. Model 3 requires higher execution time because the number of data used on the model with auxiliary sentences increased 10 times higher for the NLIM version and 30

times higher for the NLIB version. Recall the explanation in section 4.3 that for NLIM model, one review text in our dataset is transformed into 10 pairs of <*review, auxiliary sentence*> input. While for NLIB model type, one review text in our dataset is transformed into 30 pairs of <*review, auxiliary entence*> input. Therefore, when considering both effectiveness and efficiency, we argue that Model 2 is preferred over Models 1 and 2. This is our novel finding because none of previous work has compared the three different approaches utilizing BERT-based model for ABSA task.

Table 7 shows a comparison of the prediction results from a review text in the hotel review dataset. It can be seen from the review sentence "Kamar luas, bersih, dan nyaman. Ada air panas. Minus TV, tidak ada siaran" (*Rooms are spacious, clean and comfortable. There is hot water. Minus TV, there are no TV channels*.). The *IndoBERT*^{W}*-aux-NLIB* model succeeded in correctly predicting the positive sentiment for the aspect "hot_water" and the negative sentiment for the aspect "tv". If we look at the "hot_water" aspect, the six baseline models produced incorrect sentiment prediction (i.e., negative sentiment for aspect "tv"). Based on these results, it appears that the *IndoBERT*^{W}*-aux-NLIB* model can better understand the semantic and contextual meaning of the sub-sentence "..*There is hot water*.." and the sub-sentence "*Minus TV, there are no TV channels*" in the review sentence.

Table 7. The prediction results for a customer review "*Kamar luas, bersih, dan nyaman. Ada air panas. Minus TV, tidak ada siaran*" (Rooms are spacious, clean and comfortable. There is hot water. Minus TV, there are no. TV channles)

System	AC	hot_water	smell	general	cleanliness	linen	service	sunrise_meal	tv	wifi
Ground truth	neut	pos	neut	neut	pos	neut	neut	neut	neg	neut
Word2Vec-CNN-XGBoost	neut	neg	neut	neut	pos	neut	neut	neut	neg	neut
IndoBERT^{W}-CNN	neut	neg	neut	neut	pos	neut	neut	neut	neg	neut
mBERT-aux-NLIB	neut	neg	neut	neut	pos	neut	neut	neut	neg	neut
mBERT-aux-NLIM	neut	neg	neut	neut	pos	neut	neut	neut	neg	neut
mBERT	neut	neg	neut	neut	pos	neut	neut	neut	neg	neut
distilBERT	neut	neg	neut	neut	pos	neut	neut	neut	neut	neut
IndoBERT^{W}-aux-NLIB	neut	neg	neut	neut	pos	neut	neut	neut	neg	neut

One of the reasons why the *IndoBERT*^{W}*-aux-NLIB* model gives more accurate results is because this model uses the Transformer architecture, where there is a multi-head attention mechanism that aims to make the model focus on the relevant part of the input [15]. In addition, the pre-training process of the IndoBERT^{W} model is bidirectional (learning text from both directions). With this mechanism, the representation of words produced by *IndoBERT*^{W} will be more accurate and better understand the contextual relationships and meanings between words [14], [16]. Furthermore, the *IndoBERT*^{W}*-aux-NLIB* model has a mechanism for adding an input component of an NLIB-type auxiliary sentence. The mechanism of adding learned embedding (segment embedding) at the reformatting stage of the input model enables the model to distinguish which tokens include review sentences and which tokens include auxiliary sentences. Also, it stores relevant information from specific downstream tasks to be completed [17]. This confirms the finding in [28] that the mechanism for adding auxiliary sentences for sentence-pair classification can significantly improve the performance results of the ABSA task.

6. CONCLUSION

ABSA is a task that aims to extract sentiment information based on the aspect categories contained in a text. In this work, three models utilizing IndoBERT are proposed to solve the ABSA task in Indonesian hotel reviews dataset using single-sentence and sentence-pair classifications. The three proposed strategies are: i) IndoBERT is utilized in feature-based mode, paired with the CNN and machine learning models, to perform single-sentence classification; ii) IndoBERT is used with fine-tuned strategy using single sentence classification approach; and iii) IndoBERT is used with a fine-tuned strategy using sentence-pair classification by adding auxiliary sentences of the NLI-B and NLI-M types.

Our results demonstrate that our third model, i.e., fine-tuned IndoBERT using auxiliary sentences with binary classification approach NLI-B, is the most effective among all of our proposed models. It leads to significant improvement over the state-of-the-art baseline methods based on deep learning and transformer. More specifically, it outperforms the F-1 scores of these baseline methods by 2.2%-23.6%. However, the addition of the auxiliary sentence can significantly reduce the efficiency of the model. The execution time of our third proposed model with a binary classification approach is more than 2000 times longer compared to our first proposed model, and more than 30 times longer compared to our second proposed model. It happened because the number of data used in the fine-tuned IndoBERT models using auxiliary sentences with binary and multiclass classification respectively increased 30 and 10 times higher than the original data.

To sum up, when considering the effectiveness only, the third proposed model is the most superior model according to the accuracy and F-1 scores. Our first proposed model is shown to be the most efficient, but it gains the least efficiency among all of our proposed models. When considering both effectiveness and efficiency, we conclude that the second proposed model is the best-performing model because it can maintain a good balance between accurate prediction results and fast prediction time.

In future, there are some possible avenues to be conducted to extend this study. Other pre-trained language models can be exploited, such as RoBERTa and XLM-RoBERTa. In the original paper, RoBERTa and XLM-RoBERTa was reported to outperform BERT in various tasks. Therefore, it might worth investigating the performance of this model in Indonesian ABSA task. In addition, testing our models on other dataset could also be explored to examine the robustness of the models. At last, future work on ABSA task are suggested to use the fine-tuning approach using single-sentence classification as opposed to the feature-based approach or the fine-tuning approach using sentence-pair classification, in utilizing the BERT-based models for a real-time system. This is based on the insight gained from this study that the former approach could achieve a good balance between effectiveness and efficiency.

ACKNOWLEDGEMENTS

This research was funded by the Directorate of Research and Development, Universitas Indonesia, under Hibah PUTI Pascasarjana 2023 (Grant No. NKB-020/UN2.RST/HKP.05. 00/2023).

REFERENCES

[1] B. Liu, "Sentiment Analysis: Mining Opinions, Sentiments, and Emotions," *Cambridge University Press*, 2nd Edition, 2020.
[2] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, "SemEval-2014 Task 4: Aspect Based Sentiment Analysis," in *Proceedings of the 8th International Workshop on Semantic Evaluation*, 2014, pp. 27–35, doi: 10.3115/v1/S14-2004.
[3] R. Piryani, V. Gupta, V. K. Singh, and U. Ghose, "A Linguistic Rule-Based Approach for Aspect-Level Sentiment Analysis of Movie Reviews," *Advances in Computer and Computational Sciences*, pp. 201–209, 2017, doi: 10.1007/978-981-10-3770-2_19.
[4] M. E. Mowlaei, M. S. Abadeh, and H. Keshavarz, "Lexicon Generation Using Genetic Algorithm for Aspect-Based Sentiment Analysis," in *22nd International Conference on Intelligent Engineering Systems (INES)*, 2018, pp. 133–138, doi: 10.1109/INES.2018.8523902.
[5] F. A. Bachtiar, W. Paulina, and A. N. Rusydi, "Text Mining for Aspect Based Sentiment Analysis on Customer Review : A Case Study in the Hotel Industry," in *International Workshop on Innovations in Information and Communication Science and Technology*, 2020, pp. 105–112.
[6] D. Arianto and I. Budi, "Aspect-based Sentiment Analysis on Indonesia's Tourism Destinations Based on Google Maps User Code-Mixed Reviews (Study Case: Borobudur and Prambanan Temples)," in *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pp. 359–367, Oct. 2020.
[7] A. Suciati and I. Budi, "Aspect-Based Sentiment Analysis and Emotion Detection for Code-Mixed Review," *International Journal of Advanced Computer Science and Applications(IJACSA)*, vol. 11, no. 9, 2020, doi: 10.14569/IJACSA.2020.0110921.
[8] N. P. Arthamevia, Adiwijaya, and M. D. Purbolaksono, "Aspect-Based Sentiment Analysis in Beauty Product Reviews Using TF-IDF and SVM Algorithm," in *2021 9th International Conference on Information and Communication Technology (ICoICT)*, 2021, pp. 197–201, doi: 10.1109/ICoICT52021.2021.9527489.
[9] A. N. Azhar, M. L. Khodra, and A. P. Sutiono, "Multi-label Aspect Categorization with Convolutional Neural Networks and Extreme Gradient Boosting," in *International Conference on Electrical Engineering and Informatics (ICEEI)*, 2019, pp. 35–40, doi: 10.1109/ICEEI47359.2019.8988898.
[10] A. Ilmania, Abdurrahman, S. Cahyawijaya, and A. Purwarianti, "Aspect Detection and Sentiment Classification Using Deep Neural Network for Indonesian Aspect-Based Sentiment Analysis," in *2018 International Conference on Asian Language Processing (IALP)*, 2018, pp. 62–67, doi: 10.1109/IALP.2018.8629181.
[11] A. Ishaq, S. Asghar, and S. A. Gillani, "Aspect-Based Sentiment Analysis Using a Hybridized Approach Based on CNN and GA," *IEEE Access*, vol. 8, pp. 135499–135512, 2020, doi: 10.1109/ACCESS.2020.3011802.
[12] R. Catelli, S. Pelosi, and M. Esposito, "Lexicon-Based vs. Bert-Based Sentiment Analysis: A Comparative Study in Italian," *Electronics*, vol. 11, no. 3, p. 374, 2022, doi: 10.3390/electronics11030374.
[13] P. Amalia and E. Winarko, "Aspect-Based Sentiment Analysis on Indonesian Restaurant Review Using a Combination of Convolutional Neural Network and Contextualized Word Embedding," *Indonesian Journal of Computing and Cybernetics Systems (IJCCS)*, vol. 15, no. 3, pp. 285–294, 2021, doi: 10.22146/ijccs.67306.
[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, vol. 1, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.
[15] A. Vaswani *et al.*, "Attention is All you Need," in *31st Conference on Neural Information Processing Systems (NIPS)*, pp. 6000–6010, 2017.
[16] E. Yulianti, N. Pangestu, and M. A. Jiwanggi, "Enhanced TextRank Using Weighted Word Embedding For Text Summarization," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 5, pp. 5472–5482, 2023, doi: 10.11591/ijece.v13i5.pp5472-5482.
[17] N. K. Nissa and E. Yulianti, "Multi-label text classification of Indonesian customer reviews using bidirectional encoder representations from transformers language model," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 5, pp. 5641–5652, 2023, doi: 10.11591/ijece.v13i5.pp5641-5652.
[18] D. Wahyudi and Y. Sibaroni, "Deep Learning for Multi-Aspect Sentiment Analysis of TikTok App using the RNN-LSTM Method," *Building of Informatics, Technology and Science (BITS)*, vol. 4, no. 1, pp. 169–177, 2022, doi: 10.47065/bits.v4i1.1665.

[19] U. Sirisha and B. S. Chandana, "Aspect based Sentiment & Emotion Analysis with ROBERTa, LSTM," *International Journal of Advanced Computer Science and Applications(IJACSA)*, vol. 13, no. 11, 2022, doi: 10.14569/IJACSA.2022.0131189.
[20] A. Mewada and R. K. Dewang, "SA-ASBA: A Hybrid Model for Aspect-Based Sentiment Analysis Using Synthetic Attention in Pre-Trained Language BERT Model with Extreme Gradient Boosting," *The Journal of Supercomputing*, vol. 79, no. 5, pp. 5516–5551, Oct. 2022, doi: 10.1007/s11227-022-04881-x.
[21] U. Nuha and C.-H. Lin, "Aspect-Based Sentiment Analysis with Semi-Supervised Approach on Taiwan Social Distancing App User Reviews," in *2023 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, 2023, pp. 444–447, doi: 10.1109/ICAIIC57133.2023.10067048.
[22] C. Bahri and L. Suadaa, "Aspect-Based Sentiment Analysis in Bromo Tengger Semeru National Park Indonesia Based on Google Maps User Reviews," *Indonesian Journal of Computing and Cybernetics Systems (IJCCS)*, vol. 17, no. 1, pp. 79–90, 2023, doi: 10.22146/ijccs.77354.
[23] F. Said and L. P. Manik, "Aspect-Based Sentiment Analysis on Indonesian Presidential Election Using Deep Learning," *Paradigma - Journal of Computer Science and Informatics*, vol. 24, no. 2, pp. 160–167, 2022, doi: 10.31294/paradigma.v24i2.1415.
[24] A. Tiwari, K. Tewari, S. Dawar, A. Singh, and N. Rathee, "Comparative Analysis on Aspect-based Sentiment using BERT," in *2023 7th International Conference on Computing Methodologies and Communication (ICCMC)*, 2023, pp. 723–727, doi: 10.1109/ICCMC56507.2023.10084294.
[25] A. N. Azhar and M. L. Khodra, "Fine-tuning Pretrained Multilingual BERT Model for Indonesian Aspect-based Sentiment Analysis," in *2020 7th International Conference on Advance Informatics: Concepts, Theory and Applications (ICAICTA)*, 2020, pp. 1–6, doi: 10.1109/ICAICTA49861.2020.9428882.
[26] Z. Wu and D. C. Ong, "Context-Guided BERT for Targeted Aspect-Based Sentiment Analysis," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 16, pp. 14094–14102, 2021, doi: 10.1609/aaai.v35i16.17659.
[27] X. Li *et al.*, "Enhancing BERT Representation With Context-Aware Embedding for Aspect-Based Sentiment Analysis," *IEEE Access*, vol. 8, pp. 46868–46876, 2020, doi: 10.1109/ACCESS.2020.2978511.
[28] C. Sun, L. Huang, and X. Qiu, "Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol.1, 2019, pp. 380–385, doi: 10.18653/v1/N19-1035.
[29] A. Pathak, S. Kumar, P. P. Roy, and B.-G. Kim, "Aspect-Based Sentiment Analysis in Hindi Language by Ensembling Pre-Trained mBERT Models," *Electronics*, vol. 10, no. 21, 2021, doi: 10.3390/electronics10212641.
[30] M. M. Abdelgwad, T. H. A. Soliman, and A. I. Taloba, "Arabic aspect sentiment polarity classification using BERT," *Journal of Big Data*, vol. 9, no. 1, p. 115, 2022, doi: 10.1186/s40537-022-00656-6.
[31] N. Nuryani, A. Purwarianti, and D. H. Widyantoro, "Identification of Conflict Opinion in Aspect-Based Sentiment Analysis Using BERT-Based Method," in *Proceedings of the 2022 International Conference on Computer, Control, Informatics and Its Applications*, 2023, pp. 276–280, doi: 10.1145/3575882.3575935.
[32] H. Jafarian, A. H. Taghavi, A. Javaheri, and R. Rawassizadeh, "Exploiting BERT to Improve Aspect-Based Sentiment Analysis Performance on Persian Language," in *2021 7th International Conference on Web Research (ICWR)*, 2021, pp. 5–8, doi: 10.1109/ICWR51868.2021.9443131.
[33] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pp. 843–857, 2020.
[34] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 757–770, doi: 10.18653/v1/2020.coling-main.66.

BIOGRAPHIES OF AUTHORS

Evi Yulianti is a lecturer and researcher at Faculty of Computer Science, Universitas Indonesia. She received the B.Comp.Sc. degree from the Universitas Indonesia in 2010, the dual M.Comp.Sc. degree from Universitas Indonesia and Royal Melbourne Institute of Technology University in 2013, and the Ph.D. degree from Royal Melbourne Institute of Technology University in 2018. Her research interests include information retrieval and natural language processing. She can be contacted at email: evi.y@cs.ui.ac.id.

Nuzulul Khairu Nissa received B.Math. degree from Diponegoro University in 2019. She is currently pursuing a Master of Computer Science in University of Indonesia. Her research interests are related to machine learning and natural language processing. She can be contacted at email: nuzulul.khairu@ui.ac.id.