ICIC Express Letters
Volume **17**, Number **2**, February **2023**

ICIC International ©2023 ISSN 1881-803X
pp. **201–208**

SENTIMENT ANALYSIS OF THE NATIONAL COVID-19 VACCINATION PROGRAM ON TWITTER USING THE BIDIRECTIONAL ENCODER REPRESENTATION FROM TRANSFORMER

Dyan Azka Ingkafi, Gamas Adi Aryana, Arya Krisna Putra and Retno Kusumaningrum*

Department of Informatics
Universitas Diponegoro
Jl. Prof. Soedarto, SH, Tembalang, Semarang 50275, Indonesia
{ ingkafi; addygams; aryakrisnaputra }@students.undip.ac.id
*Corresponding author: retno@live.undip.ac.id

Received March 2022; accepted June 2022

Abstract. *In Indonesia, the implementation of the national COVID-19 (Coronavirus disease of 2019) vaccination programmes has received criticism from various strata of society, especially through social media platforms such as Twitter. Therefore, Twitter can be used as a data source to analyze Indonesian public sentiment regarding the vaccination programme. Various classical machine learning methods exist for sentiment analysis, but these methods require complex feature engineering and do not focus on the importance of word order in a sentence. In this study, a deep learning model, bidirectional encoder representation from transformer (BERT), is used to overcome these problems by conducting experiments to determine the best dataset after pre-processing, the best hyper-parameter, and the best pre-trained model for BERT. The data used in this study were Indonesian Twitter data with a total of 3000 tweets. Our results demonstrate that BERT is suitable for performing sentiment analysis. In our experiments, BERT obtained better results than classical machine learning methods, with a precision of 86.2%, recall of 86%, f1-score of 86%, and accuracy of 86%. The results of the sentiment analysis performed in this study can be considered by the government in formulating policies related to the implementation of vaccination programmes.*

Keywords: Sentiment analysis, COVID-19 vaccination, BERT, Indonesian Twitter

1. **Introduction.** Vaccination is one of the most important efforts made by countries in order to reduce the spread of Coronavirus disease (COVID-19), in addition to implementing various health protocols such as wearing masks, and physical distancing. However, the procurement of vaccines is also not an easy thing, especially for the country which has a very large population like Indonesia. This resulted in the emergence of various opinions in society. For example, in Indonesia, according to public assessment, the policies taken by the government seem rushed; moreover, people feel that the government have not been transparent about the empirical evidence of the continued efficacy of the vaccine [1]. Nevertheless, it is also true that many sections of the Indonesian society have welcomed the national COVID-19 vaccination programme.

Research data from Statista, one of the leading data and statistics portals in the world, shows that Indonesia has 15.1 million active Twitter users, occupying the sixth position among countries with the largest number of Twitter users in the world [2]. Therefore, public discussions about the COVID-19 national vaccination programmes are often found on Twitter. This can be seen from the frequent hashtags related to COVID-19 becoming a trending topic in Indonesia. Although the public opinion on Twitter is diverse and

DOI: 10.24507/icicel.17.02.201

abundant, most opinions have similar terms, structures, and meanings that express the same domain of knowledge. Therefore, these data can prove highly useful if processed further using a sentiment analysis approach.

Sentiment analysis can be used to explore whether the content of a document has a positive, negative, or neutral sentiment polarity, which can then be processed further to obtain useful information as per the need. Various studies on sentiment analysis on Indonesian in various domains have been conducted using classical machine learning methods, such as Naïve Bayes [3,4], logistic regression [5,6], and support vector machine [7,8]. The main problem in applying the classical machine learning method is that it is difficult to determine which feature extractor should be included in the given model. If the feature is missing or incomplete, the results of the model might not be perfect. If too many features are extracted and not all of them contribute to the output of the model, the model may not provide optimal performance. This can be overcome by applying a deep learning method, namely, a convolutional neural network (CNN) [9]. However, the application of the CNN method also has the disadvantage that it cannot work on long sequential data. This is because CNNs have no memory, and thus, they cannot store information about the meaning of a word [9]. This weakness can be overcome using long short-term memory (LSTM). LSTM is a type of recurrent neural network (RNN) architecture that is designed to maintain the previously obtained values for a certain period. LSTM can also overcome the problem of classical machine learning methods, which do not focus on the order of words in a sentence. However, LSTM still has some problems of its own; the data are processed sequentially and generated sequentially as well, which makes data training a time-intensive task. Furthermore, even bidirectional LSTM is not really "bidirectional" because technically bidirectional LSTM learns from left to right and from right to left, and the results are combined at the end, which results in the loss of the context of the word that has been learned. In addition, when using LSTM to process long text (more than 50 words), the gradient disappears [10].

This problem can be solved using bidirectional encoder representation from transformers (BERT) because it can solve the problem of forgotten information from serial models such as LSTM. BERT also speeds up the data training process as training is carried out simultaneously in two directions. This means BERT can better understand the context of a sentence. Therefore, in this study, the BERT model is applied for sentiment analysis. This study compares the performance results of the BERT model with classical machine learning methods such as logistic regression, support vector machine, Naïve Bayes classifier, random forest classifier, and k-nearest neighbours. To make the research more focused and achieve the specified goals, a limitation of scope is needed. In this study, the Twitter data used were 3000 tweets in Indonesian, uploaded on 7-19 July 2021 with the keyword "vaccination". Sentiment data are classified into three classes: neutral, positive, and negative.

Many previous studies using a dataset from the cornelledu website used the BERT model for sentiment analysis, achieving an accuracy of 73.7% [11]. However, in this study, no fine-tuning of hyper-parameters was performed. Another study [12] analyzed the sentiment of IMDb movie reviews dataset using BERT. The study used a batch size of 32, a learning rate of 2e-5, and four epochs. The results obtained were as follows: 89% accuracy, 91% precision, and 88% recall. Another study [13] produced a sentiment analysis using a dataset obtained from Rotten Tomatoes, an English-language film review site. The accuracy obtained in this study was 94% on the SST-2 model and 83.9% on the SST-5 model for $\text{BERT}_{\text{BASE}}$.

There are very few cases of sentiment analysis on datasets in Indonesian. A study [14] using a tweet dataset in Indonesian that discusses the emergency response during floods in Jakarta achieved a training dataset accuracy of 90% and a test dataset accuracy of 79%. In this study, no pre-processing of the dataset was carried out; hence, the dataset

contained noise such as slang words, typos, or non-standard abbreviations, and as a result, the model could not classify tweets properly. Another study [15] performed a sentiment analysis on the application review dataset. In this study, hyperparameter fine-tuning and comparison of two pre-trained models were performed, namely, IndoBERT and BERT Multilingual. The IndoBERT pre-trained model obtained better results with an accuracy of 84%. However, no pre-processing experiments were performed on that study. Therefore, we conducted several experiments in this study, such as determining the pre-processing of the dataset used, fine-tuning the hyperparameters, and comparing the pre-trained models. The rest of this paper is organized as follows. Section 2 describes the method used in this paper and the detailed explanation of system architecture. The results and analysis of this paper are presented in Section 3, including the performance comparison with other models. It is followed by Section 4, in which conclusions are drawn.

2. **Methods.** The proposed method consists of several steps, as shown in Figure 1.

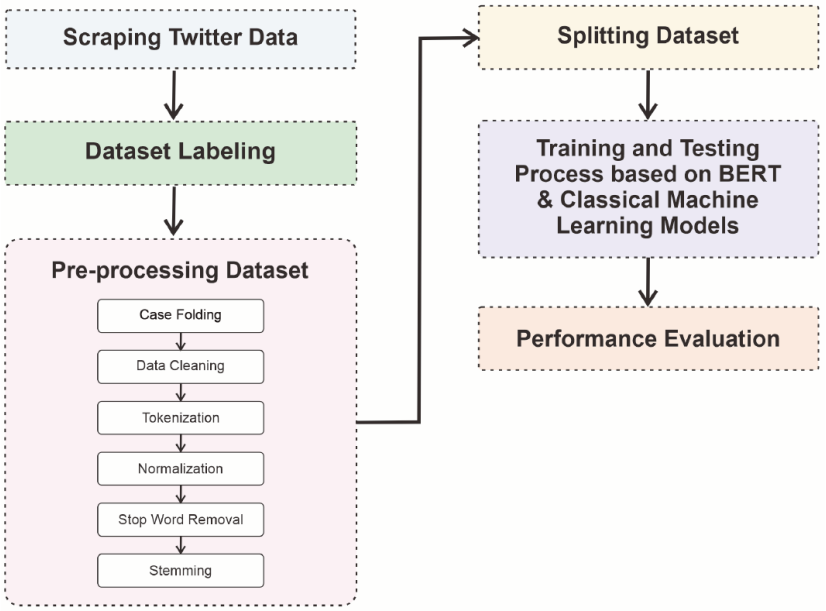

FIGURE 1. System architecture

The architecture of the system based on Figure 1 shows that the research started by scraping Twitter data using the Twitter API. The scraping results were then collected and used as a dataset. The dataset was then labelled according to its sentiment, which is negative, neutral, or positive. The next stage is the pre-processing stage. Dataset pre-processing is the process of preparing data that was initially unstructured into more structured data by passing data through several stages, such as data cleaning, case folding, tokenization, normalization, stemming, and stop word removal. The preprocessed training datasets are subsequently trained based on BERT model as well as several classical machine learning models for comparison study. In this stage, the testing process was conducted to predict testing dataset into three sentiment categories, i.e., negative, neutral, and positive. The final stage is the evaluation of the results for advanced analysis.

2.1. **Scraping Twitter data.** The data used are in the form of tweets from Twitter users in Indonesia (Indonesian) with a language filter. Data were taken from July 7-19, 2021 with the keyword "vaccination" with 3600 tweets. Of the 3600 tweets, 3000 were selected, which were divided into neutral, positive, and negative sentiments of 1000 tweets each. Scraping was performed using Tweepy, a package or library from the Python programming language, which allows interaction with the API provided by Twitter.

2.2. **Dataset labelling.** Dataset labelling was carried out to classify tweets into negative, neutral, or positive categories by assigning a value as a marker. Tweets with positive sentiments were assigned a score of 2, whereas tweets with neutral and negative sentiments were given a value of 1 and 0, respectively. The inclusion criteria for positive sentiment category were tweets containing an invitation to vaccination, an invitation to obey health protocols, satisfaction with the implementation of vaccination, and appreciation of the vaccination programmes. Inclusion criteria for neutral sentiment category were tweets that do not show any emotion towards vaccination, such as information on vaccination administration and registration. Finally, for inclusion in the negative sentiment category, the criteria were tweets that criticize the government and the vaccination services, contain thoughts about refusal to get vaccinated, and include hate speech. An example of labelling the tweet dataset is presented in Table 1.

Table 1. Example of dataset labelling

Tweet	Sentiment
"Kita harus bersama-sama menangani pandemi ini. Pak @jokowi dan jajaran sudah sekuat tenaga bekerja. Ayo sukseskan vaksinasi untuk menyelamatkan diri kita dan orang yg kita sayangi. https://t.co/utfYo2OyAY" – "We have to deal with this pandemic together. Mr. @jokowi and his staff are working as hard as they can. Let us make vaccinations successful to save ourselves and our loved ones. https://t.co/utfYo2OyAY"	Positive
Masyarakat saat ini bisa vaksinasi sendiri di Klinik Kimia Farma, berikut ini lokasinya. #kimiafarma https://t.co/wTkQGuyZfi – People can now vaccinate themselves at Kimia Farma Clinic; here are the locations. #kimiafarma https://t.co/wTkQGuyZfi	Neutral
Super ngawur. Meluaskan distribusi bukan berarti harus dibuat berbayar. Vaksinasi di pandemi begini harusnya tetap GRATIS. https://t.co/GFoGxwANG7 – Super inconsequential. Expanding distribution does not mean it has to be paid for. Vaccination in this pandemic should still be FREE. https://t.co/GFoGxwANG7	Negative

2.3. **Pre-processing dataset.** In this study, to transform unstructured datasets into structured ones to facilitate data processing, we performed a pre-processing step in several stages: case folding, data cleaning, tokenization, normalisation, stop word removal, and stemming.

1) Case folding: Case folding is a pre-processing stage that converts all uppercase letters into lowercase letters.
2) Data cleaning: At this stage, the dataset is cleaned of everything that can affect the results of the analysis, such as links, usernames (@usernames), hashtags (#), numbers, symbols, extra spaces, and punctuation.
3) Tokenization: Tokenization is a process carried out to break sentences into chunks of words, punctuation marks, and other meaningful expressions according to the terms of the language used.
4) Word normalization: In the normalization stage, non-standard words are converted into standard words according to the correct spelling.
5) Stop word removal: Stop word removal is a process carried out to remove words that have no meaning.
6) Stemming: Stemming is a process carried out to change words that have affixes into root forms by removing affixes such as prefixes, suffixes, and confixes.

2.4. **Splitting dataset.** Before implementing the model, the data are divided into testing data, training data, and validation data. Data split ratio used in this research is 70% as the train set and 30% as the test set. Furthermore, 10% of the training data is used for validation data.

2.5. **Training and testing process.** BERT [16] is a trained language representation model developed by researchers at Google AI Language in 2018. It is based on deep learning techniques and various methods such as semi-supervised learning, ELMo, ULMFiT, OpenAI transformers, and transformers. BERT uses transformers architecture, which is mechanisms that study contextual relationships between words in a text [17]. Transformers can understand and convert the understanding obtained by a mechanism called the self-attention mechanism. A self-attention mechanism is a way of transformer to check the attention from all words in the same sentence at once, which makes it a simple matrix calculation and able to parallel compute on computing units. In this study, the transformer library provided by HuggingFace was used. Sentiment analysis was performed using $\text{BERT}_{\text{BASE}}$, which has an encoder with 12 layers, 12 self-attention heads, a hidden size of 768, and 110M parameters. As comparison, we also train and test based on several classical machine learning algorithms, such as logistics regression, support vector machine (SVM), Naïve Bayes, random forest, and k-nearest neighbors (kNN).

2.6. **Performance evaluation.** In this study, we conducted several experiments to obtain the best results. Several settings were made, such as determining the best dataset after pre-processing stages, the best hyper-parameters, and the best pre-trained model for BERT.

2.6.1. *Pre-processing dataset.* In this study, we used two datasets that undergo different pre-processing stages. Dataset A was put through all the existing pre-processing stages, whereas dataset B was processed without stop word removal and stemming stages.

2.6.2. *Hyper-parameter.* We performed fine-tuning using the following hyper-parameters: (i) Batch size (16), (ii) epoch (4, 10, 16), and (iii) learning rate (2e-5). The selection of hyper-parameters was determined for several reasons. A batch size of 16 was chosen because the larger the batch size, the longer it takes to complete one batch [18]. In addition, the use of learning rate 2e-5 was chosen because it can overcome the problem of catastrophic forgetting in BERT [20]; catastrophic forgetting is a problem in which the understanding gained from pre-training is erased while learning new information or data. To determine which epoch to use, we tested 4 epochs [12,19], 10 epochs [20], and 16 epochs [21].

2.6.3. *Pre-trained models.* Two pre-trained models were selected: IndoBERT and BERT Multilingual. The IndoBERT pre-trained model was chosen because it is one of the most comprehensive models of the natural language processing (NLP) dataset for Indonesian [22]. The BERT Multilingual pre-trained model was chosen because it supports 104 languages, including Indonesian [14,21].

2.7. **Evaluation criteria.** This study uses accuracy, precision, recall, and f1-score to measure the performance of the model.

1) Accuracy: This metric indicates how often the model correctly classifies items. Accuracy can be obtained through the formula shown in Equation (1).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

2) Precision: This metric measures the accuracy of the model when predicting positively and how often the prediction is correct. The formula for precision is shown in Equation (2).

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

3) Recall: This indicates how often the model predicts a positive, when the actual class is positive. The formula for recall is given in Equation (3).

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

4) F1-score: This is the harmonic average of precision and recall. The formula for f1-score is shown in Equation (4).

$$F1\text{-}score = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \tag{4}$$

TP or True Positive is a condition when the model correctly predicted the positive class, whereas TN or True Negative is a condition when the model correctly predicted the negative class. Furthermore, FP or False Positive is a condition when the model incorrectly predicted the positive class, whereas FN is the same condition when the model incorrectly predicted but as negative class.

3. **Results and Analysis.** All experiments were conducted on Google Colab using Python version 3.711. For fine-tuning experiments, we used Torch version 1.9.0+cu102 and transformers version 4.9. The test results from the experimental settings are listed in Table 2.

Table 2. Experimental results

Epoch	Pre-trained model	Dataset	Wall time	Precision	Recall	F1-score	Accuracy
4	IndoBERT	B	4min 3s	0.856	0.856	0.856	0.855
4	IndoBERT	A	4min 2s	0.826	0.83	0.83	0.828
4	BERT Multilingual	B	4min 23s	0.82	0.828	0.82	0.822
4	BERT Multilingual	A	4min 21s	0.803	0.803	0.803	0.804
10	IndoBERT	B	10min11s	0.852	0.856	0.851	0.854
10	IndoBERT	A	9min 51s	0.83	0.83	0.828	0.828
10	BERT Multilingual	B	10min 54s	0.83	0.833	0.83	0.832
10	BERT Multilingual	A	10min 43s	0.826	0.83	0.823	0.827
16	IndoBERT	B	16min 26s	**0.862**	**0.86**	**0.86**	**0.86**
16	IndoBERT	A	15min 57s	0.846	0.838	0.838	0.837
16	BERT Multilingual	B	17min 30s	0.84	0.84	0.84	0.841
16	BERT Multilingual	A	16min 51s	0.8	0.798	0.8	0.8

Based on Table 2, it can be seen that the model obtains the best results when using dataset B. In addition, the model obtains the best results when using an epoch of 16 and the pre-trained IndoBERT model with a precision value of 86.2%, recall of 86%, f1-score of 86%, and accuracy of 86%. IndoBERT obtained the best results because the BERT Multilingual was only trained on the Wikipedia corpus. IndoBERT was trained on larger data in Indonesian and contains formal and slang languages, such as those common in Twitter [16]. IndoBERT was trained with over 220M words from various sources, such as 1) Indonesian Wikipedia (74M words); 2) news articles from Kompas, Tempo, and Liputan6 (55M words in total); and 3) an Indonesian Web Corpus (90M words) [22]. This makes IndoBERT specially trained for Indonesian, unlike BERT Multilingual, which also supports other languages. The best results were obtained when using dataset B because

the process of stop word removal and stemming can change the context of a sentence. After testing, we conducted experiments using the classical machine learning method for comparison. The five methods we used were logistic regression, SVM, Naïve Bayes classifier, random forest classifier, and kNN. The experiment was carried out using dataset B, the results of which are listed in Table 3. Subsequently, based on Table 3, it can be seen that the best results were obtained by the SVM model with a precision value of 83.1%, recall of 81.1%, f1-score of 81.2%, and accuracy of 81.1%.

Table 3. Experimental results using classical machine learning methods

Machine learning	Precision	Recall	F1-score	Accuracy
Logistic Regression	0.825	0.81	0.81	0.81
SVM	**0.831**	**0.811**	**0.812**	**0.811**
Naïve Bayes Classifier	0.785	0.775	0.776	0.775
Random Forest Classifier	0.815	0.803	0.803	0.803
kNN	0.782	0.755	0.757	0.756

A comparison between the BERT and SVM models is presented in Table 4. Based on the table, BERT obtained better results compared with SVM on the four indicators: precision, recall, f1-score, and accuracy. The experiments also obtained slightly better results than those of [16].

Table 4. Comparison of BERT model results with SVM

Model	Precision	Recall	F1-score	Accuracy
BERT	**0.862**	**0.86**	**0.86**	**0.86**
SVM	0.831	0.811	0.812	0.811

4. **Conclusions.** In this study, we performed sentiment analysis on Twitter database containing views on Indonesia's national COVID-19 vaccination programme. The experiment conducted uses various settings such as epoch settings (4, 10, and 16), pre-trained model for BERT (IndoBERT and BERT Multilingual), and two datasets that undergo different pre-processing stages. The BERT model achieved the best results with 16 epoch settings and the IndoBERT pre-trained model; the best dataset was the one for which stop word removal and stemming stages were not performed as these stages can change the context of a sentence. IndoBERT achieved better results because this pre-trained model was specially trained for Indonesian with over 220M words from various sources. The results of this study demonstrate that the use of the BERT model in sentiment analysis yields better results than the classical machine learning methods. In a comparison between BERT and SVM, BERT achieved an accuracy of 86%, whereas SVM achieved an accuracy of 81.1%. However, our experiments had several limitations. First, various tests of batch size settings were not conducted, and second, the datasets we used were not sufficiently large. We intend to address these issues in our future research.

REFERENCES

[1] I. Akbar, COVID-19 vaccination and state policy: A political economy perspective, *Journal of Acad. Praja*, vol.4, no.1, pp.244-254, 2021.

[2] Twitter: Most users by country, *Statista*, https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/, Jul. 27, 2021.

[3] Y. Pratama, A. R. Tampubolon, L. D. Sianturi, R. D. Manalu and D. F. Pangaribuan, Implementation of sentiment analysis on Twitter using Naïve Bayes algorithm to know the people responses to debate of DKI Jakarta Governor Election, *Journal of Physics Conference Series*, vol.1175, no.1, DOI: 10.1088/1742-6596/1175/1/012102, 2019.

[4] Pristiyono, M. Ritonga, M. A. Al Ihsan, A. Anjar and F. H. Rambe, Sentiment analysis of COVID-19 vaccine in Indonesia using Naïve Bayes algorithm, *IOP Conference Series in Material Science Engineering*, vol.1088, no.1, DOI: 10.1088/1757-899x/1088/1/012045, 2021.

[5] P. Bharti, M. Bakshi and R. A. Uthra, Fake news detection using logistic regression, sentiment analysis and web scraping, *International Journal of Advanced Science and Technology*, vol.29, no.7, pp.1157-1167, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85084811301&partnerID =40 &md5=4f174c99317c1c86e7f4c6be34aafa1a, 2020.

[6] S. P. Sheela, Sentiment analysis and prediction of online reviews with empty ratings, *International Journal of Applied Engineering Research*, vol.13, no.14, pp.11532-11539, 2018.

[7] J. S. Pasaribu, Appilication of K-Means algorithm to predict consumer interest according to the season on place reservation and food online software, *Journal of Physics: Conferenc Series*, vol.1477, no.2, DOI: 10.1088/1742-6596/1477/2/022023, 2020.

[8] S. Fransiska and A. I. Gufroni, Sentiment analysis provider By.U on Google Play Store Reviews with TF-IDF and support vector machine (SVM) method, *Scientific Journal of Informatics*, vol.7, no.2, http://journal.unnes.ac.id/nju/index.php/sji, 2020.

[9] P. F. Muhammad, R. Kusumaningrum and A. Wibowo, Sentiment analysis using Word2vec and long short-term memory (LSTM) for Indonesian hotel reviews, *Procedia Computer Science*, vol.179, no.2020, pp.728-735, DOI: 10.1016/j.procs.2021.01.061, 2021.

[10] M. Jiang, J. Wu, X. Shi and M. Zhang, Transformer based memory network for sentiment analysis of web comments, *IEEE Access*, vol.7, pp.179942-179953, DOI: 10.1109/ACCESS.2019.2957192, 2019.

[11] C. A. Putri, Classification of sentiment analysis on English film reviews with analisis sentimen review film Berbahasa Inggris Dengan Pendekatan bidirectional encoder representations from transformers approval, *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, vol.6, no.2, pp.181-193, DOI: 10.35957/jatisi.v6i2.206, 2020.

[12] S. Abdul, Y. Qiang, S. Basit and W. Ahmad, Using BERT for checking the polarity of movie reviews, *International Journal of Computer Applications*, vol.177, no.21, pp.37-41, DOI: 10.5120/ijca 2019919675, 2019.

[13] M. Munikar, S. Shakya and A. Shrestha, Fine-grained sentiment classification using BERT, *IEEE International Conference on Artificial Intelligence for Transforming Business Society (AITB2019)*, pp.2-5, DOI: 10.1109/AITB48515.2019.8947435, 2019.

[14] W. Maharani, Sentiment analysis during Jakarta flood for emergency responses and situational awareness in disaster management using BERT, *2020 8th International Conference on Information and Communication Technology (ICoICT)*, pp.1-5, DOI: 10.1109/ICoICT49345.2020.9166407, 2020.

[15] K. S. Nugroho, A. Y. Sukmadewa, F. A. Bachtiar and N. Yudistira, BERT fine-tuning for sentiment analysis on Indonesian mobile apps reviews, *arXiv.org*, arXiv: 2107.06802v1, 2020.

[16] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, *2019 Conference of the North American Chapter of the Association for Computational Linguistic: Human Language Technology (NAACL HLT 2019)*, pp.4171-4186, 2019.

[17] A. Vaswani et al., Attention is all you need, *Advances in Neural Information Processing Systems*, pp.5999-6009, 2017.

[18] D. Osinga, *Deep Learning Cookbook*, O'Reilly Media, Inc., 2018.

[19] C. Sun, X. Qiu, Y. Xu and X. Huang, How to fine-tune BERT for text classification?, in *Chinese Computational Linguistics. CCL 2019. Lecture Notes in Computer Science*, M. Sun, X. Huang, H. Ji, Z. Liu and Y. Liu (eds.), Cham, Springer, 2019.

[20] Y. Song, J. Wang, Z. Liang, Z. Liu and T. Jiang, Utilizing BERT intermediate layers for aspect based sentiment analysis and natural language inference, *arXiv.org*, arXiv: 2002.04815, 2020.

[21] M. R. Yanuar and S. Shiramatsu, Aspect extraction for tourist spot review in Indonesian language using BERT, *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pp.298-302, DOI: 10.1109/ICAIIC48513.2020.9065263, 2020.

[22] F. Koto, A. Rahimi, J. H. Lau and T. Baldwin, IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP, *Proc. of the 28th International Conference on Computational Linguistics*, pp.757-770, DOI: 10.18653/v1/2020.coling-main.66, 2021.