ICIC Express Letters
Volume **16**, Number **10**, October **2022**

ICIC International ©2022 ISSN 1881-803X
pp. 1047–1055

SENTIMENT CLASSIFICATION FOR INDONESIAN SENTENCES USING MULTILINGUAL TRANSFORMERS MODEL

Riccosan*, Rhio Sutoyo and Andry Chowanda

Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta 11480, Indonesia
{ rsutoyo; achowanda }@binus.edu
*Corresponding author: riccosan@binus.edu

Received October 2021; accepted January 2022

Abstract. *Sentiment classification is an essential thing for knowing the value of public opinions and it is one of Natural Language Processing (NLP) tasks that can be carried out using transformers, one of the deep-learning architectures in the machine learning methods. The main goal of this research is to employ the transformers multilingual model for the Indonesian sentiment classification task and to understand its performance. The dataset used in this research called SmSA which is accessed from IndoNLU. The models utilized were Google's Multilingual BERT (M-BERT) and Facebook's XLM RoBERTa (XLM-R). In addition, exploration is carried out in the sentiment classification task using an imbalanced SmSA dataset to see how it affects the model. We do small fine-tuning with the model's parameters and hyper-parameters to improve its performance. Through the experiment, results in the form of F1-Score were obtained, with the score of the uncased M-BERT as 0.85 (batch = 32) and 0.84 (batch = 48); the M-BERT cased as 0.82 (batch = 32) and 0.84 (batch = 48); and the XLM-R model as 0.90 (batch = 32 and 48).*

Keywords: Natural language processing, Sentiment classification, Transformers model

1. **Introduction.** Human populations can be a source of data in the form of opinions such as blog posts, comments, reviews, or tweets, and it can be used for sentiment analysis to obtain considerations for determining steps in the future. For example, to assist the marketing division in a campaign or product launch, determining what things are more popular, or identification sentiment of a feature [1]. Sentiment classification is included in sentiment analysis which is one type of Natural Language Processing (NLP) to track value about a particular topic [1]. Sentiment analysis has been widely used by companies to extract opinions about their products or services [2]. There are three methods for sentiment analysis task: Machine Learning (ML), Hybrid Learning (HL), and Lexicon-based [3].

Our research is using deep learning in ML method. ML method has benefit such as ease of spotting patterns, ongoing improvement, and the capacity to handle multidimensional and multivariate data [4]. The development for Natural Language Understanding (NLU) in tasks of NLP, has been going on until now, especially in the field of deep learning with transformers architecture that have become popular because of advantages for industry and effective in cross-sectional areas, such as language and various models for the task.

The development of applications for NLU often begins with a common international language, for example, English, while only about 5% of the total world population is estimated to use English as their first language [5]. Therefore, there is a challenge to present a product that can be accessed by various speakers of different languages in the

DOI: 10.24507/icicel.16.10.1047

world and it is answered by several large companies by presenting a multilingual model, for example, Google provides the multilingual BERT (M-BERT) from BERT model [6]. The good thing of BERT model is simple yet, but powerful, because you can fine-tune it only by adding an output layer and configurate for the tasks. Next is, XLM-RoBERTa (XLM-R) model [7] from the RoBERTa model [8] that was created by Facebook in order to improve the performance of BERT model with more data to develop the pre-trained model and some extra settings. In this paper, we investigated the performance of both models for related task and fine-tuned them to increase their performance. This study will serve as the foundation for our future research. Other reason why we are using both models because they have been trained in more than 100 languages. Through the development of the two models, there is a definite promise for the implementation and launch of applications for NLU [5].

Various sentiment analysis researches employing ML approaches with Indonesian language datasets have been conducted. We see this as a chance to investigate NLP with transformers architecture, particularly in terms of sentiment classification for datasets in the form of Indonesian sentences. As a result, the M-BERT [6] and XLM-R [7] of multilingual transformers are used. The IndoNLU SmSA dataset [9] was used to train our experimental model for the task. The dataset originally comes from Purwarianti and Crisdayanti's research in an experiment to develop Bi-LSTM (Bi-directional Long Short-Term Memory) performance for sentiment analysis [10]. This study's findings include a comparison of the two models' accuracy displayed in the form of a loss and accuracy report derived from model training results, a confusion matrix with model testing on the test dataset, and a classification report to display the total classification after the model has been tested.

This paper is divided into five chapters with Chapter one about the introduction why this research was carried out; Chapter two, which contains the foundation in the form of previous research; Chapter three, which contains the methodology and flow of the research; Chapter four, which contains the explanations of the experiments conducted and their results; and finally, Chapter five, which contains conclusions and future research plans.

2. **Recent Work.**

2.1. **Sentiment analysis methods.**

There are 3 methods for sentiment analysis mentioned by Ligthart and friends, namely Machine Learning (ML), Lexicon Based, and Hybrid Learning (HL) [3]. The first is ML divided into 3 components: supervised, unsupervised, and semi-supervised. Support-Vector Machine (SVM) [11], Naïve Bayes, Maximum Entropy (ME), and Artificial Neural Network (ANN) [12] are examples of supervised learning. LDA (Latent Dirichlet Allocation) and PLSA (Probabilistic Latent Semantic Analysis) are examples of unsupervised learning [13]. Semi-supervised learning combines supervised and unsupervised learning, such as combined SVM, Naïve Bayes, and Logistic Regression (LR) [14] to build the model.

ML comes out as the superior one because it can automate tasks using computer technology that is still evolving, can handle large collections of datasets [3], and there are various models available for different tasks (for example, SVM, Naïve Bayes, ANN, and transformers). As a result, our experiment is using ML approaches with multilingual transformers models. The second method is Lexicon Based, focusing on keywords from opinions to carry out the process of sentiment analysis or classification. This method is separated into two approaches: dictionary-based (focused on keywords from the dataset, then searching for synonyms and antonyms) and corpus-based (starting with a list of words in the opinion, then carry out a search in a larger corpus for generating opinion terms that are contextually relevant). The identification of Indonesian-language opinion tweets uses the Lexicon-learning approach [15] and sentiment analysis using Lexicon Based on social websites [16] which are two examples with this method. The third way is Hybrid Learning (HL), a method that combines ML with Lexicon Based. Sentiment analysis of twitter data for application suggestions [17] is one of its examples. Gandhe and his colleagues used a combination of Naïve Bayes and Lexicon Based approach in the research.

2.2. **Sentiment classification with deep learning architecture.** In sentiment analysis research and opinion mining [1] paper generally explained that ML methods for the sentiment analysis process is mostly classified as supervised and included as a text classification technique. Deep learning in ML method for the NLP task has multiple advantages, including the ability of automation, ability to process vast volumes of data, and various models. Many studies for sentiment analysis for Indonesian datasets utilizing with deep learning in ML method have been carried out. Some of them are sentiment analysis for informal Indonesian tweets using Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) [18], sentiment analysis for tweets of Indonesian cellular operators using the SVM-Linear technique, Naïve Bayes, and Decision Tree [19], sentiment analysis on reviews of Indonesian mobile applications using the fine-tuned Google BERT transformers architecture [20], and many other studies.

The transformers architecture is well-liked and outperforms others. Wang et al. mentioned the benefits of it [21], which is stated that transformers architecture is superior to the Recurrent Neural Network (RNN) model since the process may be done in parallel and increase the efficiency. Furthermore, this architecture includes numerous models for various tasks as well as a pre-trained form [22]. This enables the transformer architecture to be developed with the researcher's dataset and configured for specialized tasks in the research. As a result, for our research, we employ transformers architecture with multilingual models, notably M-BERT [6] and XLM-R [6], with a dataset specifically for Indonesian. Several types of dataset are required for sentiment analysis in this method and ours use three types: train, valid, and test. Train data is used to train the model for the task. During the training, valid data is used to verify the model's accuracy. The last is test data for evaluating the performance of the model in terms of sentiment analysis when applied to an NLP task.

Fine-tuning the Google BERT model and its multilingual for Indonesian mobile application review [20] and improving Indonesian text classification using multilingual language model [23] are the examples of the recent work with multilingual transformers model and especially with Indonesian datasets. Opportunities have been seen and experiment was carried out with the models to obtain a comparison of performance in the task of Indonesian sentiment classification.

3. **Methodology.** This section describes the steps we propose for our research in employing the multilingual transformers models to perform sentiment classification on Indonesian datasets. The model is tasked with determining if a sentence is positive, negative, or neutral. There are 3 base models used, namely M-BERT Base Uncased, M-BERT Base Cased, and XLM-RoBERTa Base. These are models that have been pre-trained. The following is a comparison of the base model for this experiment in Table 1 based on the total parameters, trainable parameters, non-trainable parameters, and classifier.

The model is created using the Python 3 in the Google Collaboratory [24]. The model consists of starting set-up, retrieving files from Google Drive (G-Drive), pre-processing datasets into pandas data frames along with adding column names and converting labels to numeric, creating input sequences for converting data into input examples and tensor datasets, section configurating, fine-tuning, and training models, section testing models for sentiment prediction of random Indonesian sentences, and the last section is evaluation models. The research process which is divided into 3 phases is shown in Figure 1.

TABLE 1. Transformers multilingual model for sentiment analysis/classification

Model name	M-BERT Base Uncased	M-BERT Base Cased	XLM-R Base
Total parameters	167,358,723	177,855,747	278,045,955
Trainable parameters	167,358,723	177,855,747	278,045,955
Non-trainable parameters	0	0	0
Classifier	2,307	2,307	592,899

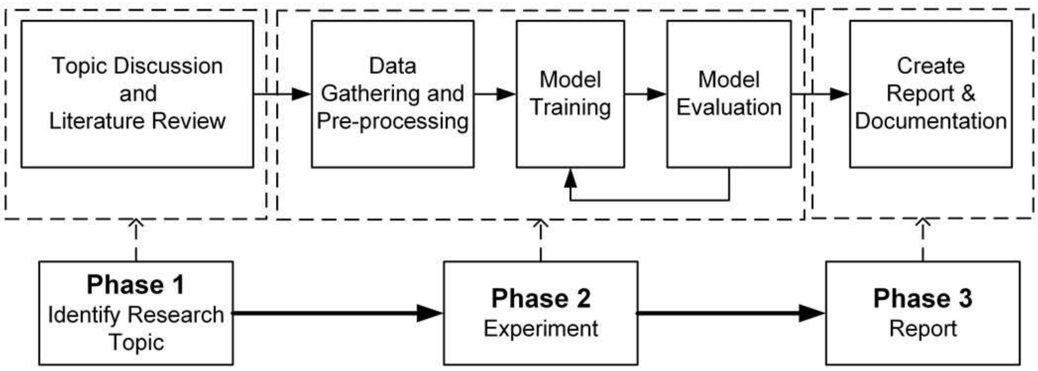

FIGURE 1. Research and experiment flow

3.1. **Phase 1 – Identify research topic.** To begin, a team meeting was conducted to discuss the topic and the result is NLP for sentiment classification using the multilingual transformers. Based on the topic, we go on to the literature review stage, where we look for prior studies and examples of multilingual transformers models to use as reference material in this study in the Phase 2: Experiment.

3.2. **Phase 2 – Experiment.** This phase contains model creation, training, and the collection of post-experimental data. We begin by searching the dataset to be used. The SmSA dataset [9] (a dataset for Indonesian sentence sentiment analysis) is used in the experiment, and accessed from IndoNLU. For study efficiency, the datasets are stored in G-Drive. The dataset is divided into 3 parts: training, validation, and testing. After the calling procedure, the dataset is accommodated into variables according to their respective functions.

The dataset is pre-processed by converting into a pandas data frame, adding column names, and transform the label to numeric, with "0" as neutral, "1" as positive, and "2" as negative. In this experiment, we propose additional experiment about utilizing the random sampling method with the goal of balancing the dataset. From a quantitative standpoint, this experiment was effective in balancing the dataset, but it had a detrimental influence on model training and assessment. The trained model is overfitting after the experiment, and the prediction is a failure. As a result, the resampling experiment holded up, and the research was continued using the basic dataset value.

The second is creating the Input Sequence function. With the first, convert input sequences into input examples; and the second, convert it to tensor datasets with features like input ids, attention mask, and token type ids. The model tokenizer is used to tokenize the dataset of input instances in the second part of the function. Next, the batches for each dataset are configured. The training dataset is the only ones that are mixed and repeated. After the datasets ready, we carry out the model training step.

In the model training, we propose minor fine-tuning to increase the performance by setting the learning rate, epsilon, and clipnorm. The last is model evaluation phase with 3 methods, namely the Loss and Accuration Report, the Confusion Matrix to see the accuracy of the model in predicting sentiment towards the test dataset, and the Classification Report to see the report detail related to the Indonesian Sentiment Classification task. If the evaluation yields poor result, the process returns to the Model Training [Figure 1].

3.3. **Phase 3 – Report.** Report takes the form of a research file that includes the research's background, theoretical reviews about previous research as the basic foundation, research stages, elaboration of experimental results and evaluation, conclusions, and future plans.

4. Experimental Settings.

4.1. **Data gathering and pre-processing.** The dataset used is SmSA which is accessed from IndoNLU [9]. This dataset was created by compiling comments and answers from various Indonesian Internet sites. The sentiment label has been divided into three types: neutral, positive, and negative. The dataset is separated into two columns: text and sentiment labels. The dataset has been split into three parts: training, validation, and test. Training and validation data are used in the model training phase, whereas test data is used in the model assessment phase. The description of the amount of data for each segment in terms of labels is shown in Table 2.

Table 2. SmSA total data based on sentiment label

Sentiment label	**(SmSA) dataset type**				
	Training	**Training (Random oversampling)**	**Training (Random undersampling)**	Validation	Test
Neutral	1,148	**6,000**	**1,148**	131	101
Positive	6,416	**6,416**	**2,100**	735	198
Negative	3,436	**6,000**	**2,100**	394	201
Total data	11,000	**18,416**	**5,348**	1,260	500

Based on the labels, there is a large distribution gap. From that, a resampling procedure was carried out with random oversampling and random undersampling. M-BERT Base Uncased was the model used with the results of the experiment as shown in Table 2. After resampling, we are configurating the batch for dataset (batch = 32) and training the model for 10 epochs. Unsatisfactory results were achieved after training the process because the model became overfitting. This is depicted in Figure 2 which shows the degree of loss and accuracy as report throughout training and model validation.

The validation loss rate increased, from 0.63 with No Resampling to 6.08 with the oversampling and to 5.19 with the undersampling approach, an increase of 9.65 times and 8.23 times, respectively. Meanwhile, the degree of validation accuracy fell from 0.91 to 0.31 while utilizing the original No Resampling method, and fell by 2.9 times when using the oversampling and undersampling processes. As result, the dataset resampling process was stopped, and the experiment was restored using the basic dataset value.

4.2. **Model configuration.** In this experiment, the total label parameter in the basic model is set to 3 pieces to modify the amount of sentiment labels for the dataset, which is divided into three categories: neutral, positive, and negative. The base model's hyperparameter setup takes the form of batch settings for reading datasets. Batch 32 and batch 48 were used in this experiment to split the processing batch. In the training phase of the whole model, we set the total repetition to 10 epochs. The amount of loss and accuracy of

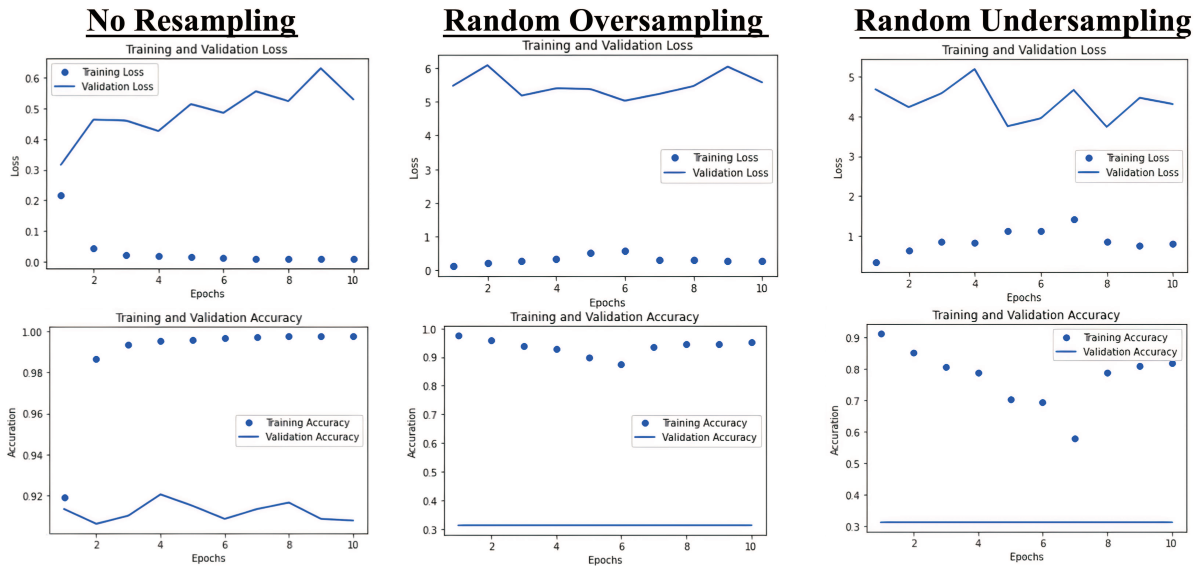

FIGURE 2. Loss and accuracy report (training) – Before & after resampling dataset

the model for the execution of natural language tasks in terms of sentiment classification using the Indonesian dataset are found to be affected by the batch setting for reading the dataset.

4.3. **Experimental result.** Our experiment to perform the related task was quite promising. The comparison of the models with the experimental result is presented in Table 3. The XLM-R Base model, got F1-Score of 0.90 for batches 32 and 48, and had the best results. During validation, the maximum loss level for the model dropped from 0.507 (batch = 32) to 0.429 (batch = 48). Figure 3 shows the degree of accuracy and loss for XLM-RoBERTa Base, the fundamental model in the experiment, during training and validation.

TABLE 3. Model evaluation report

Base model	M-BERT Base Uncased		M-BERT Base Cased		**XLM-RoBERTa Base**	
Batch	32	48	32	48	**32**	**48**
F1-Score	0.85	0.84	0.82	0.84	**0.90**	**0.90**
Misclassification	0.148	0.158	0.184	0.162	**0.104**	**0.096**
(Highest) Model validation loss	0.631	0.714	0.604	0.618	**0.507**	**0.429**

The confusion matrix obtained during the assessment step of the XLM-R Base model using a test dataset is shown in Figure 4. A simple experiment with batch settings using the basic M-BERT Base Cased and XLM-R Base models, shows that the larger the batch, the lower the rate of misclassification (the failure rate of the model in classifying data according to the proper label). However, due to the increased misclassification of the M-BERT Base Uncased model and the fact that the quality of the model is also connected to the quality of the dataset used for training in carrying out sentiment classification tasks, this knowledge still requires additional assessment. The classification report in Figure 5, which shows the final assessment report for two batches using the basic XLM-R model. An exploration was carried out to balance the dataset as described in Section 4.1. Random oversampling and random undersampling are used as the approach. Because they resample data from the same data that is repeated, the result is unsatisfactory because the model becomes overfitting and the experiment employs back the dataset with base value.

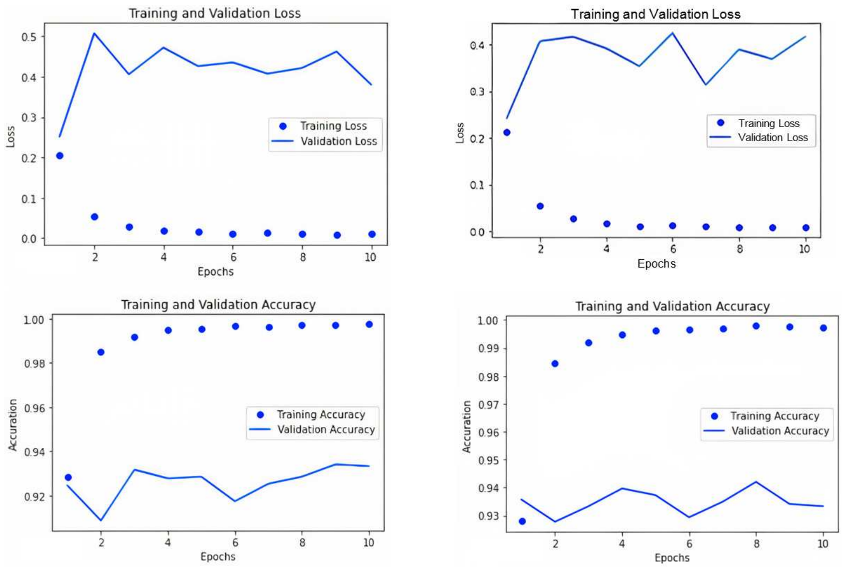

FIGURE 3. XLM-R, train and valid loss and accuracy, batch 32 (left) – batch 48 (right)

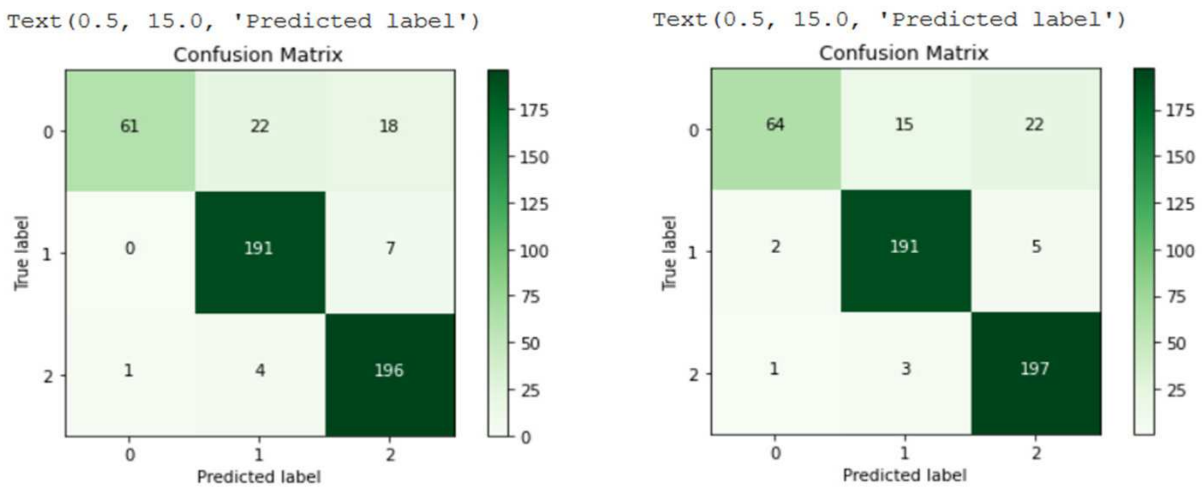

FIGURE 4. XLM-R confusion matrix for batch = 32 (left) & batch = 48 (right)

	precision	recall	f1-score	support
Neutral	0.98	0.60	0.75	101
Positive	0.88	0.96	0.92	198
Negative	0.89	0.98	0.93	201
accuracy			0.90	500
macro avg	0.92	0.85	0.87	500
weighted avg	0.90	0.90	0.89	500

Batch = 32

	precision	recall	f1-score	support
Neutral	0.96	0.63	0.76	101
Positive	0.91	0.96	0.94	198
Negative	0.88	0.98	0.93	201
accuracy			0.90	500
macro avg	0.92	0.86	0.88	500
weighted avg	0.91	0.90	0.90	500

Batch = 48

FIGURE 5. Base model = XLM-R, model evaluation classification report

5. **Conclusion and Future Work.** In our experiment, the quality of the models for the specified task is affected by the balance of data distribution in the dataset. A conclusion is reached as an understanding that the model correctness level will decrease, rise, or remain unchanged as a result of adjusting the batch reading of the dataset. This study still needs to be improved, despite the fact that the base model produced high degree of accuracy. Through the experiment of sentiment classification for Indonesian dataset with

multilingual transformers, we were able to gain an understanding of how multilingual transformers work for the related task, the effect of the balance distribution of data in the dataset on the basic model used, and the impact of random sampling in an attempt to balance the data distribution. In the future, our research on NLP with multilingual transformers will continue in the form of hyper-parameters configuration, continue the exploration of resampling process, and implement the models embedding for the related models.

REFERENCES

[1] G. Vinodhini and R. Chandrasekaran, Sentiment analysis and opinion mining: A survey, *International Journal of Adv. Research in Comp. Sci. and Software Eng.*, vol.2, no.6, pp.282-292, 2012.

[2] A. N. Azhar and M. L. Khodra, Fine-tuning pretrained multilingual BERT model for Indonesian aspect-based sentiment analysis, *2020 7th International Conference on Advance Informatics: Concepts, Theory and Applications (ICAICTA)*, DOI: 10.1109/ICAICTA49861.2020.9428882, 2020.

[3] A. Ligthart, C. Catal and B. Tekinerdogan, Systematic reviews in sentiment analysis: A tertiary study, *Artificial Intelligence Review*, DOI: 10.1007/s10462-021-09973-3, 2021.

[4] K. C. A. Khanzode and R. D. Sarode, Advantages and disadvantages of artificial intelligence and machine learning: A literature review, *International Journal of Library and Information Science*, vol.9, no.1, pp.30-36, 2020.

[5] C. Wang and M. Banko, Practical transformer-based multilingual text classification, *Proc. of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pp.121-129, DOI: 10.18653/v1/2021.naacl-industry.16, 2021.

[6] 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*, vol.1, pp.4171-4186, DOI: 10.18653/v1/N19-1423, 2019.

[7] A. Conneau, K. Khandelwal, N. Goyal et al., Unsupervised cross-lingual representation learning at scale, *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics*, pp.8440-8451, DOI: 10.18653/v1/2020.acl-main.747, 2020.

[8] Y. Liu, M. Ott, N. Goyal et al., RoBERTa: A robustly optimized BERT pretraining approach, *arXiv Preprint*, arXiv: 1907.11692, 2019.

[9] B. Wilie, K. Vincentio, G. I. Winata et al., IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding, *Proc. of the 1st Conf. of the Asia-Pacific Chapter of the Assoc. for Comp. Linguistics and the 10th International Joint Conf. on Natural Language Processing*, Suzhou, China, pp.843-857, 2020.

[10] A. Purwarianti and I. A. Crisdayanti, Improving Bi-LSTM performance for Indonesian sentiment analysis using paragraph vector, *2019 6th International Conference on Advance Informatics: Concepts, Theory and Applications (ICAICTA)*, pp.1-5, DOI: 10.1109/ICAICTA.2019.8904199, 2019.

[11] H. X. Shi and X. J. Li, A sentiment analysis model for hotel reviews based on supervised learning, *Proc. of the 2011 International Conference on Machine Learning and Cybernetics*, Guilin, China, vol.3, pp.950-954, DOI: 10.1109/ICMLC.2011.6016866, 2011.

[12] M. Anjaria and R. M. Guddeti, A novel sentiment analysis of social networks using supervised learning, *Social Network Analysis and Mining*, vol.4, DOI: 10.1007/s13278-014-0181-9, 2014.

[13] Z. Madhoushi, A. R. Hamdan and S. Zainudin, Sentiment analysis techniques in recent works, *2015 SAI Conference*, London, UK, pp.288-291, DOI: 10.1109/SAI.2015.7237157, 2015.

[14] N. F. Silva, L. F. Coletta and E. R. Hruschka, A survey and comparative study of tweet sentiment analysis via semi-supervised learning, *ACM Computing Surveys*, vol.49, no.1, pp.1-26, DOI: 10.1145/2932708, 2016.

[15] Y. Azhar, Metode lexicon-learning based untuk identifikasi tweet opini berbahasa Indonesia, *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, vol.6, no.3, pp.237-242, DOI: 10.23887/janapati.v6i3.11739, 2017 (in Indonesia).

[16] F. M. Kundi, A. Khan, S. Ahmad and M. Z. Asghar, Lexicon-based sentiment analysis in the social web, *Journal of Basic and Applied Scientific Research*, vol.4, no.6, pp.238-248, 2014.

[17] K. Gandhe, A. S. Varde and X. Du, Sentiment analysis of Twitter data with hybrid learning for recommender applications, *2018 9th IEEE Annual UEMCON*, New York, pp.57-63, DOI: 10.1109/UEMCON.2018.8796661, 2018.

[18] T. A. Le, D. Moeljadi, Y. Miura and T. Ohkuma, Sentiment analysis for low resource languages: A study on informal Indonesian tweets, *Proc. of the 12th Workshop on Asian Language Resources*, Osaka, Japan, pp.123-131, 2016.

[19] H. Wijaya, A. Erwin, A. Soetomo and M. Galinium, Twitter sentiment analysis and insight for Indonesian mobile operators, *ISICO*, Bali, Indonesia, pp.367-372, 2013.

[20] K. S. Nugroho, A. Y. Sukmadewa, H. W. DW, F. A. Bachtiar and N. Yudistira, BERT fine-tuning for sentiment analysis on Indonesian mobile apps reviews, *arXiv Preprint*, arXiv: 2107.06802, 2021.

[21] C. Wang, M. Li and A. J. Smola, Language models with transformers, *arXiv Preprint*, arXiv: 1904.09408, 2019.

[22] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi and A. M. Rush, HuggingFace's transformers: State-of-the-art natural language processing, *Proc. of the 2020 Conf. on Empirical Methods in NLP: System Demonstrations*, pp.38-45, DOI: 10.18653/v1/2020.emnlp-demos.6, 2020.

[23] I. F. Putra and A. Purwarianti, Improving Indonesian text classification using multilingual language model, *2020 7th International Conference on Advance Informatics: Concepts, Theory and Applications (ICAICTA)*, DOI: 10.1109/ICAICTA49861.2020.9429038, 2020.

[24] Google Research, *Google Colab*, https://research.google.com/colaboratory/, 2021.