ICIC Express Letters
Volume **17**, Number **11**, November **2023**

ICIC International ©2023 ISSN 1881-803X
pp. **1295–1302**

CODE-MIXED SENTIMENT ANALYSIS INDONESIAN-ENGLISH USING TRANSFORMER MODEL

Cuk Tho[1,*], Yaya Heryadi[1], Iman Herwidiana Kartowisastro[1,2] and Widodo Budiharto[3]

[1]Computer Science Department, BINUS Graduate Program – Doctor of Computer Science
[2]Computer Engineering Department
Faculty of Engineering
[3]Computer Science Department
School of Computer Science
Bina Nusantara University
Jl. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia
{ yayaheryadi; imanhk; wbudiharto }@binus.edu
[*]Corresponding author: cuktho@binus.edu

Received November 2022; accepted February 2023

Abstract. *The content of social media in the past decade has become very instrumental as input data in analyzing public opinion toward various issues. The obtained opinion can be positive or negative toward management decisions of organizations or companies, promotion of advertisers, or campaign of political contestants. The advent of machine learning technology in the past decade has changed the way people evaluate sentiment analysis toward public opinions by applying machine learning or a deep learning model. This paper proposed the Transformer model in analyzing the sentiment of a sentence originating from Twitter and containing code-mixed text in Indonesian and English. Transformer models are now widely used for Natural Language Processing (NLP) and have better accuracy than previous models such as Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) as an alternative in Sentiment Analysis. The experimental results show that the use of a transformer in this study has achieved 86% average accuracy using English-to-Indonesian text alignment and 88% average accuracy using Indonesian-to-English text alignment. This experiment result is better than the experiment results of the previous study as the baseline, which is 81% and 80% average accuracy achieved by both CNN and LSTM models for texts translated into Indonesian and 84% and 83% for texts translated into English.*
Keywords: CNN, Code-mixed, LSTM, Sentiment analysis, Transformer

1. **Introduction.** The increasing popularity of social media in the past decade has changed the way people communicate, written or spoken; not only do people tend to use informal language, but they also tend to involve words from more than one language (code-mixed language), which can be a local, national, or international language. Mixing languages has become common practice in communication, including in the social media communication [1]. As we know that communication on social media is done by posting messages, it can influence someone's perception of certain issues based on their content which can be positive or negative sentiment [2]. Understanding sentiment polarity (sentiment for short), which is delivered verbally or in writing, can be very useful for influencers, companies, and politicians.

The advent of deep learning technology in the past decade has made it possible for a human to efficiently recognize the sentiment in various lengths of text or document. However, most prominent sentiment analysis methods are designed to analyze monolingual language text. Hence, finding the right deep-learning model for analyzing sentiment in

DOI: 10.24507/icicel.17.11.1295

certain languages or mixed languages can be very challenging. On the other hand, the high availability of textual documents from social media has encouraged many researchers to do research in this field, and until now, it is still an interesting topic to study, especially topic related to sentiment analysis that involves mixed languages, which is commonly known as code-mixed.

This research proposed a different approach to sentiment analysis by adopting the word alignment and Transformer model on code-mixed texts, especially on code-mixed Indonesian and English, where at this time, no one has conducted research for code-mixed sentiment analysis in Indonesian-English using the Transformer model. Moreover, the experiments showed that the Transformer model could be used as one of the alternatives in sentiment analysis with an accuracy of 88% compared to the Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) models. The remainder of this paper is arranged as follows. Section 2 describes the literature review summary, which discusses the theory and related research. Section 3 presents the proposed method for this research and continues with Section 4, which delivers the result and discussion. Section 5 concludes this paper.

2. **Literature Review.** A survey done by [1] showed that Naïve Bayes (NB) and Support Vector Machine (SVM) are the two most common machine learning algorithms that are used as the baseline. Research disclosed that sentiment analysis containing code-mixed was mostly carried out for Hindi-English, Tamil-English, and Bengali-English [1,2]. However, since the Transformer model was introduced by [3] as an alternative to machine translation, many researchers are starting to adopt the Transformer as an alternative in the sentiment analysis research field. Research conducted using the Transformer model, and attention mechanism for Hindi-English code-mixing has been carried out by Zaharia et al. [4] and Pradhan and Sharma [5]. All of them stated that the use of a Transformer or attention mechanism, which is the basis of the Transformer model, has a higher accuracy than the baseline models such as CNN, LSTM and Bi-LSTM [6]. Sentiment analysis of Code-Mixed Roman Urdu-English using the Transformer model has also been carried out by Younas et al. [7]. In their research, they stated that using the Transformer model also resulted in better accuracy than the baseline model. Based on the results of the literature review, it can be seen that most of the code-mixed sentiment analysis research was conducted for Hindi, Tamil, Bengali and Urdu by using CNN, LSTM or Transformer model.

As indicated before, it can be concluded that research on sentiment analysis using Indonesian-English code-mixed using the Transformer model still needs to be well-known. In addition to this research, we put additional word alignment [8] to enhance the model's accuracy and also to address the issue of sentence structure in Indonesian that is reversed from English.

3. **Methodology.** Sentiment classification using the Transformer model was obtained from the Indonesian-English code-mixed training dataset, which had previously been preprocessed, applied word embedding and alignment. The Transformer model obtained from the training results will classify mixed words in Indonesian and English, whether they fall into positive, negative, or neutral sentiment categories.

3.1. **Transformer model.** The Transformer model used in this study uses a Transformer model that refers to Vaswani et al. [3]. In general, the Transformer architecture in Figure 1(a) is a sequence-to-sequence model that consists of an encoder and a decoder. The encoder maps the input string of words into a series of vector numbers representing each word. The output of the Transformer model lies in the decoder, which also generates a vector sequence for each word representation. The encoder and decoder layers consist of sub-layers that are called multi-head attention and feed-forward fully connected layers.

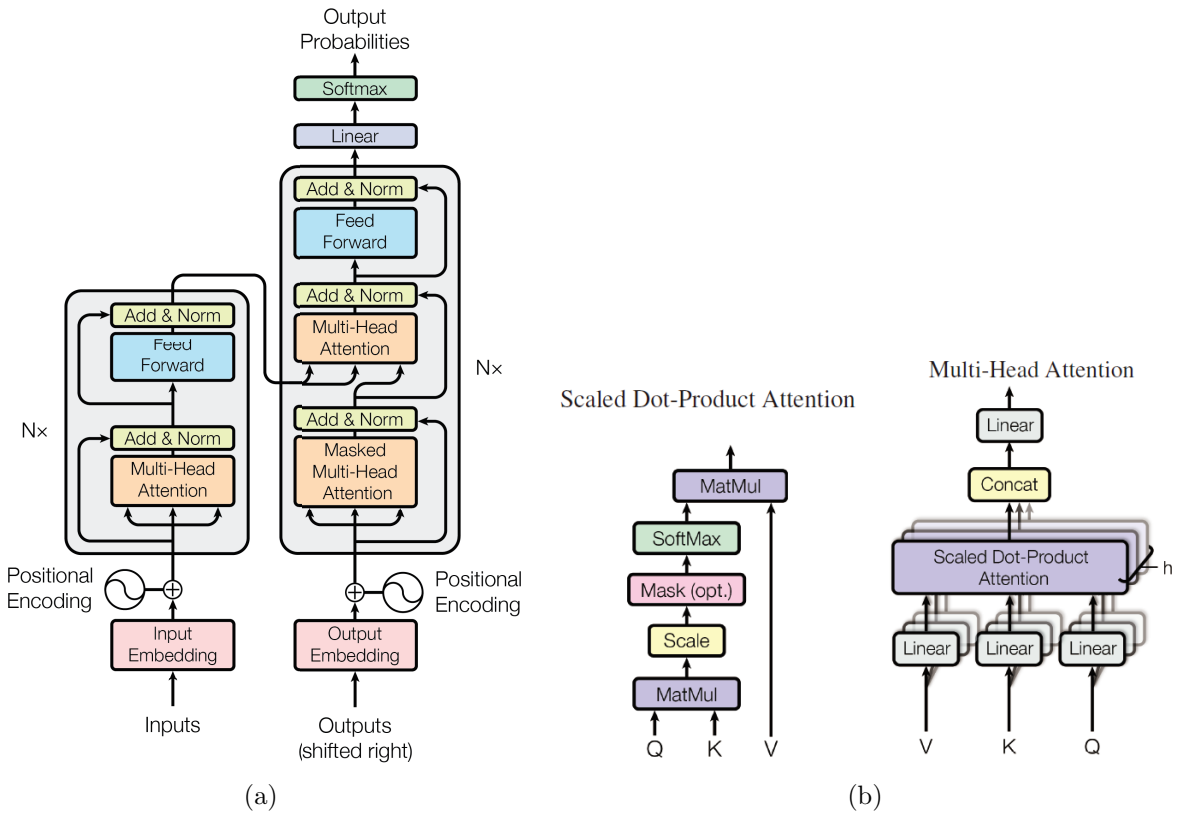


FIGURE 1. (a) Transformer model architecture; (b) attention mechanism [3]

The output of the Encoder, including the Key and Value, will be concatenated with the Query from the decoder from the previous layer. For each sub-layer, a layer-normalization is inserted to speed up the training process and produce better generalizations [8].

Multi-head attention is the mechanism that distinguishes the Transformer model from other deep learning models. Multi-head attention that is mentioned above is shown in Figure 1(b). The attention mechanism used by the Transformer is to map Query from the order of incoming words by weighting Key-Values in the resulting output. The attention of Query, Key, and Value (Q, K, V) is calculated using the scaled dot product equation as shown in Equation (1). Q refers to Query, K refers to Key, V refers to Values and $\sqrt{d_k}$ is the key dimension. All Q, K and V are vectors.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

Multi-head attention in Figure 1(b) is a composition of several self-attention to enhance the performance of the single attention mechanism and can be calculated using Equation (2). Through this attentional equation, each incoming word order has its weight (W) against the others.

$$MultiHead(Q, K, V) = Concat\left(head_1, head_2, \ldots, head_h\right)W^O \tag{2}$$

where $head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right)$.

3.2. **Word alignment.** Word alignment is needed to find the equivalence of a source word to the target word. This word alignment is very useful, particularly in word translation or code-mixed text. Song et al. [9] conducted a research on word alignment from Transformer model. The target word is calculated using Equation (3) [9].

$$l_t^{word} = -\log\left(p\left(y_t \mid S_t^j, X\right)\right) \tag{3}$$

where l is likelihood translation and S_t^j hidden state of the decoder final layer, X is the source sentence and y_t is the target words.

3.3. **Feed-forward multi-layer perceptron.** The sentiment classification process is carried out by a feed-forward multi-layer perceptron to determine sentiment predictions based on the training results using a code-mixed dataset labeled sentiment in each sentence. Equation (4) represents the multi-layer perceptron output in the form of a sentiment classification results [3].

$$FFN(X) = \max(0, XW_1 + b_1)W_2 + b_2 \tag{4}$$

where W is the matrices parameter and b is bias.

3.4. **Proposed method.** The proposed method for the code-mixed sentiment analysis of Indonesian English using the Transformer model can be seen in Figure 2. Extracting information from code-mixed text requires word embedding, and word alignment must be carried out. Word embedding represents words into vector spaces [10]. In this vector space, words will be matched against words in other languages with similar meanings. In addition, word alignment brings those two vectors closer in the vector space [9]. This paper will also use word embedding and alignment by implementing the Transformer model, which is considered capable of producing a better word alignment [9].

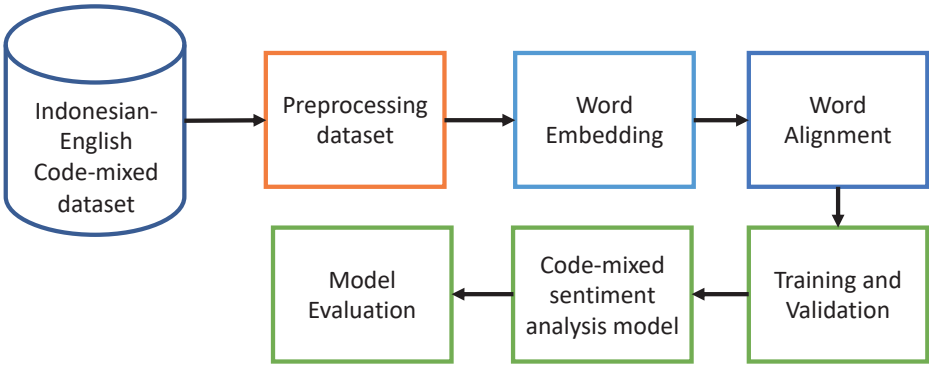


FIGURE 2. The proposed method

The dataset used in this experiment is a code-mixed Indonesian-English dataset taken from Twitter. Each sentence is labeled as the sentiment of each sentence, namely positive, negative, or neutral. Remembering text data coming from Twitter is very unstructured, and then the dataset is then preprocessed in the form of tokenization so that each word can be processed for embedding and alignment. Preprocessing also removes characters that are not useful for the sentiment analysis process, such as the @#%* character and the like. In addition, preprocessing also converts all uppercase letters to all lowercase letters to save the embedding process because they have the same meaning. Excess space is also removed in this preprocessing [11]. After preprocessing is done, then the embedding and alignment process is carried out to convert words into vectors. After the words are converted into vectors, then the alignment process is carried out to find equivalent words that have the same meaning in the same language. This search is done by finding the position of the same word vector in one vector space.

Furthermore, a training and validation process is carried out to produce a code-mixed sentiment analysis model in Indonesian English. After the training and validation process was completed, a code-mixed sentiment analysis model of Indonesian English was obtained. The next step is to evaluate the metrics in the form of a confusion matrix to determine the accuracy of the model [12].

The algorithm used in this Transformer model can be seen in Algorithm 1.

Algorithm 1. Training model.

load *Indonesian-English code-mixed dataset*
for each *sentence in Indonesian-English code-mixed dataset*
  tokenize (sentence)
  Word embedding
  Vector alignment Indonesian and English
for each *word in sentence find similar vector*
    normalize vector sentence
initialize test_size to 0.2
initialize train_size to 0.8
split *dataset into training and testing data according to test_size and train_size*
set *parameters for transformers model*
set *batch size to 64, learning rate to 0.001*, set *epoch to 10, activation function Adam, Loss function Sparse Categorical Cross Entropy*
for *epoch = 1: number of epochs*
  for *batch = 1: number of batches*
    Generate another batch
    Train the model
    Backpropagate the loss
    Update model parameter
calculate *sentiment classification prediction*

4. **Experiment Results.** The Indonesian-English code-mixed sentiment analysis dataset in this experiment used a secondary dataset taken from [13]. The dataset consists of 40,000 lines of sentences and is manually labeled as 0 for negative sentiment, 1 for neutral sentiment, and 2 for positive sentiment. Of the 40,000 rows of the dataset, 80% were used for training, and 20% were used for testing or validation. The training process uses Google Collaboratory with GPU mode, which runs for 10 epochs. The results of the training and validation can be seen in Figure 3.

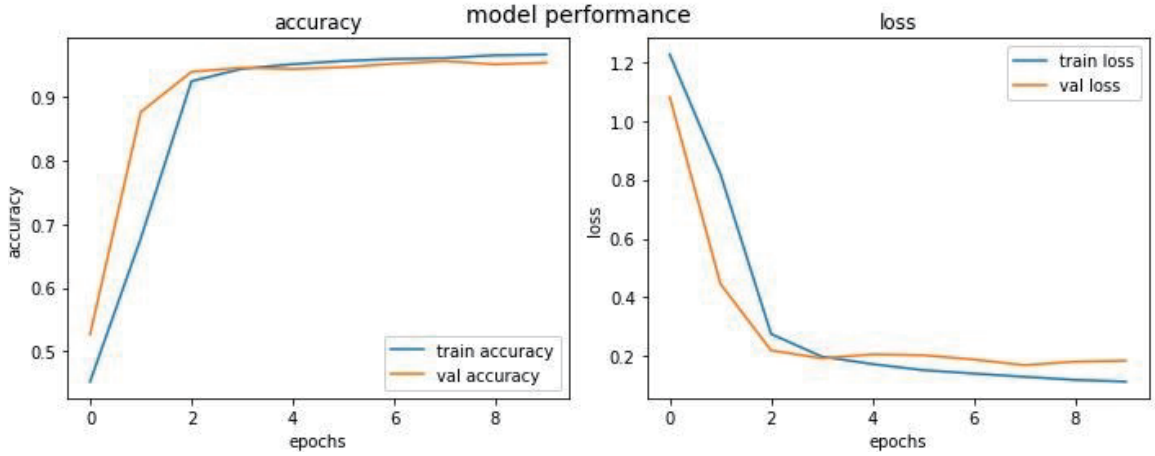


Figure 3. Training and validation results after 10 epochs

The validation results show an accuracy close to 0.95 and a loss close to 0.25. The training results also show no overfitting or underfitting, which indicates that the training process has produced a convergent model and can be generalized to new data [14,15]. The results of the text position of words in Indonesian and their equivalents in English can be seen in Figure 4.

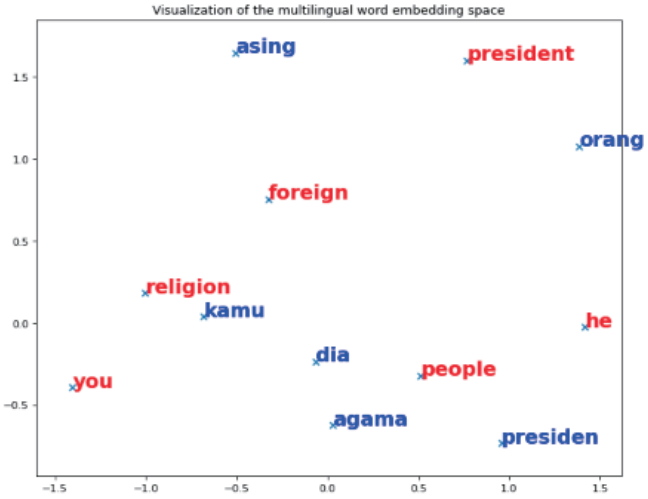


FIGURE 4. Word alignment result

From Figure 4, it can be seen that the blue Indonesian words have their red counterparts in English. Both are in the same vector space and are adjacent. Sentences containing code-mixed Indonesian and English will be translated into Indonesian and English before sentiment analysis is carried out. The purpose of this step is to determine the level of accuracy of sentiment analysis for each sentence in Indonesian and English. The results of the translation and prediction of these sentiments can be seen in Figure 5.

```
Original text:
ganti presiden now

Translate to English:
replace president now
1/1 [==============================] - 0s 31ms/step

Sentiment prediction: negative
```

(a)

```
Original text:
ganti presiden now

Translate to Indonesian:
ganti presiden sekarang
1/1 [==============================] - 0s 39ms/step

Sentiment prediction: negative
```

(b)

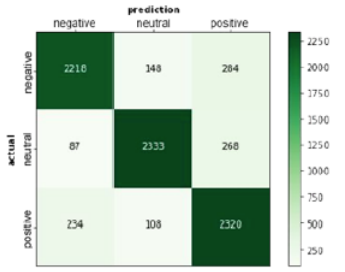
FIGURE 5. Sentiment analysis results (a) English and (b) Indonesian

From Figure 5, it can be seen that the Indonesian-English code word mix, namely "ganti presiden now" is translated into English, all of which become "replace president now" with negative sentiment results (a). Meanwhile, if translated into Indonesian, it becomes "ganti presiden sekarang" with negative sentiments (b). The comparison results using the CNN model and the LSTM-based model developed by Konate and Du [6] can be seen in Table 1.

Confusion matrix testing using test data of 20% of the dataset shows very promising results. Figure 6 shows the result evaluation using a confusion matrix, in which (a) the vector in English is mapped into Indonesian vector space; meanwhile, (b) the vector in Indonesian is mapped into English. The confusion matrix shows an accuracy of 86% for sentiment analysis in (a) and 88% for sentiment analysis in (b). The same result

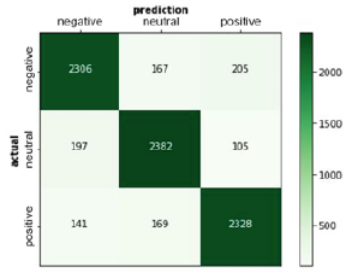
TABLE 1. Comparison results

Model	Indonesian	English
Our model	86%	88%
One-layer CNN	81%	84%
Two-layer Bi-LSTM	80%	83%



	precision	recall	f1-score	support
0	0.87	0.84	0.85	2650
1	0.90	0.87	0.88	2688
2	0.81	0.87	0.84	2662
accuracy			0.86	8000
macro avg	0.86	0.86	0.86	8000
weighted avg	0.86	0.86	0.86	8000

(a)



	precision	recall	f1-score	support
0	0.87	0.86	0.87	2678
1	0.88	0.89	0.88	2684
2	0.88	0.88	0.88	2638
accuracy			0.88	8000
macro avg	0.88	0.88	0.88	8000
weighted avg	0.88	0.88	0.88	8000

(b)

FIGURE 6. Sentiment analysis results (a) Indonesia and (b) English

was also applied to average recall, precision, and f1-score, which showed that the model performance was a good fit. Adding the word alignment on the Transformer showed that our model performs better than the CNN and LSTM as stated in Table 1. Furthermore, our model reaches its maximum performance in neutral sentiments and performs better if the code-mixed is translated into English.

5. **Conclusions.** This work investigated the implementation of the Transformer model as an alternative deep learning model on sentiment analysis that focused on code-mixed text in Indonesian and English. The experimental results show that the use of a transformer in this study has achieved 86% average accuracy using English-to-Indonesian text alignment and 88% average accuracy using Indonesian-to-English text alignment. This experiment result is better than the experiment results of the previous study as the baseline, which is 81% average accuracy achieved by One-layer CNN for texts translated into Indonesian and 84% for texts translated into English. This experiment also showed better performance than Two-layer Bi-LSTM with 80% for translated to Indonesian and 83% translated to English. The experiment showed promising results to be applied in practical applications to develop a tool to make social media more useful as a communication media alternative. In addition, the Transformer model is feasible to be improved further for analyzing sentiment from code-mixed text.

Acknowledgment. The authors thankfully acknowledge the helpful comments and suggestions of the reviewers.

REFERENCES

[1] G. I. Ahmad, J. Singla, A. Ali, A. A. Reshi and A. A. Salameh, Machine learning techniques for sentiment analysis of code-mixed and switched Indian social media text corpus: A comprehensive review, *Int. J. Adv. Comput. Sci. Appl.*, vol.13, no.2, pp.455-467, doi: 10.14569/IJACSA.2022.0130254, 2022.

[2] M. Wankhade, A. C. S. Rao and C. Kulkarni, A survey on sentiment analysis methods, applications, and challenges, *Artificial Intelligence Review*, vol.55, pp.5731-5780, 2022.

[3] A. Vaswani et al., Attention is all you need, *Adv. Neural Inf. Process. Syst.*, vol.30, pp.5999-6009, 2017.

[4] G.-E. Zaharia, G.-A. Vlad, D.-C. Cercel, T. Rebedea and C.-G. Chiru, UPB at SemEval-2020 task 9: Identifying sentiment in code-mixed social media texts using transformers and multi-task learning, *Proc. of the 14th Int. Work. Semant. Eval.*, pp.1322-1330, doi: 10.18653/v1/2020.semeval-1.179, 2020.

[5] R. Pradhan and D. K. Sharma, An ensemble deep learning classifier for sentiment analysis on code-mix Hindi-English data, *Soft Comput.*, doi: 10.1007/s00500-022-07091-y, 2022.

[6] A. Konate and R. Du, Sentiment analysis of code-mixed Bambara-French social media text using deep learning techniques, *Wuhan Univ. J. Nat. Sci.*, vol.23, no.3, pp.237-243, doi: 10.1007/s11859-018-1316-z, 2018.

[7] A. Younas, R. Nasim, S. Ali, G. Wang and F. Qi, Sentiment analysis of code-mixed Roman Urdu-English social media text using deep learning approaches, *Proc. of the 23rd Int. Conf. Comput. Sci. Eng. (CSE 2020)*, pp.66-71, doi: 10.1109/CSE50738.2020.00017, 2020.

[8] M. Zeineldeen, A. Zeyer, R. Schlüter and H. Ney, Layer-normalized LSTM for hybrid-HMM and end-to-end ASR, *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020)*, pp.7679-7683, 2020.

[9] K. Song et al., Towards better word alignment in transformer, *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol.28, no.8, pp.1801-1812, doi: 10.1109/TASLP.2020.2998278, 2020.

[10] Q. Jiao and S. Zhang, A brief survey of word embedding and its recent development, *IEEE 5th Adv. Inf. Technol. Electron. Autom. Control Conf. (IAEAC 2021)*, vol.2021, pp.1697-1701, doi: 10.1109/IAEAC50856.2021.9390956, 2021.

[11] S. Pradha, M. N. Halgamuge and N. T. Q. Vinh, Effective text data preprocessing technique for sentiment analysis in social media data, *Proc. of 2019 11th Int. Conf. Knowl. Syst. Eng. (KSE 2019)*, pp.1-8, doi: 10.1109/KSE.2019.8919368, 2019.

[12] R. D. Tan et al., LMS content evaluation system with sentiment analysis using lexicon-based approach, *2022 10th Int. Conf. Inf. Educ. Technol. (ICIET 2022)*, pp.93-98, doi: 10.1109/ICIET55102.2022.9778976, 2022.

[13] R. S. Perdana, *Indonesian Sentiment Tweet Dataset Unlabeled*, https://github.com/ridife/dataset-idsa/blob/master/Indonesian Sentiment Tweet Dataset Unlabeled.zip, accessed on August 10, 2022.

[14] X. Guyon and J. Yao, On the underfitting and overfitting sets of models chosen by order selection criteria, *J. Multivar. Anal.*, vol.70, no.2, pp.221-249, doi: 10.1006/jmva.1999.1828, 1999.

[15] X. Ying, An overview of overfitting and its solutions, *J. Phys. Conf. Ser.*, vol.1168, no.2, doi: 10.1088/1742-6596/1168/2/022022, 2019.