Emotion Classification of Indonesian Tweets using BERT Embedding

Muhammad Habib Algifari [1]*, Eko Dwi Nugroho [2]*
* Teknik Informatika, Institut Teknologi Sumatera
muhammad.algifari@if.itera.ac.id [1], eko.nugroho@if.itera.ac.id [2]

Article Info

Article history:

Received 2023-09-12
Revised 2023-09-30
Accepted 2023-10-03

Keyword:

BERT,
CNN,
LSTM,
Twitter.

ABSTRACT

Twitter is one of the social media that has the largest users in the world. Indonesia is one of the countries that has the 5th largest number of Twitter users in the world which causes a high possibility of conflict between Indonesian Twitter users due to emotional tension in tweets. In this paper, we will compare the BERT embedding method with CNN and LSTM. The results of this experiment are BERT-CNN has the best performance results which has an accuracy of 61% compared to BERT-LSTM. In the experiment several stages of data preprocessing, data cleaning, data spiting and data training were carried out and the results were evaluated using confusion metrics.

I. INTRODUCTION

Twitter, a widely recognized social media platform, has transformed the methods of communication, information sharing, and global connectivity. Since its inception in 2006, Twitter has become a worldwide sensation, with millions of active users engaging in live discussions covering a wide array of subjects including news, entertainment, sports, politics, and more. This platform has become an influential tool for expressing thoughts, reaching a large audience, and actively participating in valuable conversations for individuals, businesses, organizations, and public figures. According to We Are Social and Hootsuite, the number of Twitter users worldwide has reached 556 million as of January 2023. Specifically for Indonesia, there are approximately 24 million active Twitter users [1].

Twitter users commonly utilize the platform to share updates on their daily lives, express their emotions, and even discuss political perspectives. As a result, disagreements arising from varying viewpoints and opinions are not uncommon on Twitter. Users are granted the freedom to express their feelings openly. Nevertheless, there are instances where tweets are frequently misinterpreted, primarily due to the complexity of accurately understanding emotions solely through written text.

The identification and detection of emotions expressed in Twitter user's tweets can be achieved through the application of machine learning techniques. Machine learning models are commonly to classify the content into specific categories. In this research, the selected machine learning models for the task are convolutional neural networks (CNN) and long short-term memory (LSTM).

CNNs are deep learning models developed to analyse visual data like images and videos. They excel in computer vision tasks such as image classification, object detection, and image recognition. The key feature of CNNs is their capacity to learn and extract hierarchical features from input data. They achieve this through the utilization of convolutional layers. These layers employ small filters or kernels that slide across the input data, performing convolutions to identify local patterns and features. By stacking multiple convolutional layers, CNNs can learn intricate and abstract features at various spatial scales [2].

LSTM, which stands for long short-term memory, is a type of recurrent neural network characterized by its ability to maintain a state memory and its layered cell structure. LSTM is a versatile and effective tool that finds applications in various domains, such as statistics, linguistics, medicine, transportation, computer science, and more [3].

Numerous studies have been conducted thus far on identifying and classifying various emotions based on Twitter user tweets. While most of the research has focused on processing English tweets, there also research focused on Indonesian tweets. Hence, this study aims to build upon the existing state-of-the-art research by combining different methods and approaches. The proposed approach for the machine learning model involves utilizing a combination of a convolutional neural network (CNN) and long short-term memory (LSTM). These two models will be integrated with the BERT encoder method, which will serve as the embedding technique.

II. Literature Review

This research involved conducting experiments on various deep neural network (DNN) architectures to determine their performance in classifying emotions within Indonesian-language tweets. The architectures tested in this study included Long Short-Term Memory (LSTM), bidirectional LSTM (BiLSTM), stacked BiLSTM, and Gated Recurrent Unit (GRU). Each DNN architecture involved embedding the tweets and processing them using either the FastText or Word2Vec embedding methods. The findings of this study indicated that utilizing DNN models resulted in higher accuracy compared to previous studies that used SVM or random forests. Specifically, the BiLSTM architecture demonstrated the best performance, achieving an accuracy rate of 70.83%, which was 8% higher than SVM and 15% higher than random forests according to earlier papers [4].

A machine learning model was developed to classify Indonesian-language tweets into five emotional categories: happy, angry, scared, sad, and love. The experiment involved combining a convolutional neural network (CNN) with different embedding methods, including ELMo (Embedding from Language Models), BERT (Bidirectional Encoder Representation from Transformers), and Word2Vec. The results of the experiment showed that the CNN model using BERT, named BERT-CNN, achieved an f1 value of 72.83%. The ELMo-CNN model achieved an f1 value of 55.69%, while the Word2Vec-CNN model achieved an f1 value of 65.57% [5].

Various embedding methods, including Glove, Word2Vec, and FastText, were assessed to determine the most effective method when combined with a convolutional neural network (CNN) for classifying emotions in Indonesian tweets. The study utilized a publicly available dataset that categorized emotions into five groups: love, joy, anger, sadness, and fear. The findings revealed that the combination of CNN with Word2Vec achieved an F1 score of 72.06% [6].

The authors created a publicly available dataset comprising tweets written in Indonesian, which was utilized for emotion classification purposes. The study also involved conducting feature engineering to identify the most effective features for emotion classification. The experimented features included lexicon-based features, bag of words, word embedding, orthography, and part of speech tags. The findings demonstrated that implementing a combination of all these features resulted in an F1 score of 69.73% [7].

This study aimed to conduct experiments on detecting and analyzing emotional sentiment in tweets posted by Twitter users. The researcher collected tweets related to specific topics and compiled them into a dataset. This dataset was then used for emotion detection in tweets. The paper introduces several novel aspects, including (i) incorporating tweet replies in the dataset and analysis, (ii) introducing agreement scores, sentiment scores, and emotional response scores in calculating influence scores, and (iii) generating recommendations for public and personalized lists of users who share the same topic and express similar emotions and sentiments towards the topic [8] .

In a previous study [6], researchers examined the performance of various embedding methods including ELMo, BERT, and Word2Vec in conjunction with the CNN model. The findings indicated that the combination of the BERT embedding method with the CNN model yielded highly promising results. On the other hand, in another study [5], the authors compared the performance of LSTM and BiLSTM architectures with multiple embedding methods. The experiments revealed that the BiLSTM architecture demonstrated superior performance. Therefore, the aim of the present research is to compare the performance of the CNN and LSTM models when combined with the BERT embedding method.

III. Methodology

This research will compare the performance of the BERT embedding method with CNN and LSTM and will compare it with the performance on the baseline paper [6]. The experimental stages to be carried out in this study can be seen in Figure 1.

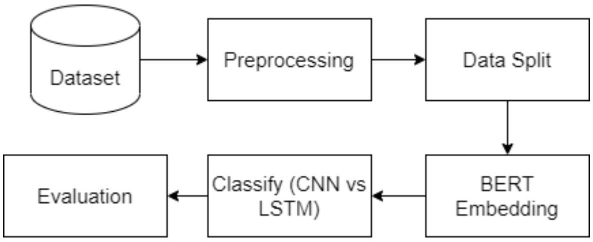

Figure 1. Proposed Method

A. Dataset

The dataset employed in this study is the same as several prior state-of-the-art studies, and it was compiled specifically for this research [8]. The data was collected by streaming tweets using the Twitter Streaming API over a span of four days, from June 1, 2018, to June 14, 2018. The streamed tweets were written in Indonesian and amounted to approximately 4403 tweets. These tweets were manually labeled and categorized into five emotion classes: anger, happy, love, fear, and sadness.

During the experimental process, the dataset was partitioned into three subsets: 80% for training data, 10% for validation data, and 10% for testing data. The distribution of the training, validation, and testing data is outlined in Table 1 below.

TABLE 1.
DATASET

Dataset	Train	Valid	Test
Anger	880	111	110
Fear	813	102	101
Sadness	519	65	64
Love	797	100	99
Happy	509	64	63

B. *Evaluation Metrics*

The comparison will be made between the experimental outcomes and the findings of the most recent state-of-the-art paper [6], which utilized the same dataset. The comparison will focus on evaluating the performance of the CNN-BERT model, both in the study [6] and in the current research paper. Additionally, the performance of the LSTM-BERT architecture, a recent innovation employed for classifying the identical dataset, will also be included in the performance analysis.

Metrics used for performance comparisons are precision, recall, accuracy, and F1-score. Precision is calculated by dividing the number of accurately retrieved instances by the total number of instances that were retrieved.

$$Precission : P = \frac{TP}{TP+FP} \qquad [9]$$

Recall is determined by dividing the number of correctly retrieved instances by the total number of instances that should have been retrieved.

$$Recall : R = \frac{TP}{TP+FN} \qquad [9]$$

Accuracy is a metric that quantifies the proportion of correctly retrieved instances, including both positive and negative instances, among all the instances that were retrieved.

$$Accuracy : A = \frac{TP+TN}{TP + TN+FP+FN} \qquad [9]$$

The F-score is a metric that combines precision and recall by taking their weighted average, where the weighting is determined by the function β.

$$F - score : F_1 = 2 * \frac{P*R}{P+R} \qquad [9\text{-}10]$$

IV. RESULT AND ANALYSIS

A. *Data Pre-processing*

The collected tweet data may contain non-standard words and irrelevant components, such as usernames and links, which do not contribute to emotion classification. To enhance the accuracy of classification and simplify the data, several stages of data cleaning are necessary. In this experiment, the data cleaning process involved case folding, word normalization, removal of special characters, and stemming. These steps were taken to streamline the data and improve the accuracy of emotion classification.

B. *Training Parameter*

In conducting CNN and LSTM training the number of epochs used is 32, and the loss calculation method is sparse categorical cross entropy, and the accuracy metrics used are sparse categorical cross entropy. While the optimization used is Adam.

C. *Result BERT-CNN*

Experiments with BERT-CNN were carried out by forming several layers. The first layer is the input layer which is used to initialize the data train used. next is the preprocessing layer, the embedding layer which uses the BERT method. Then there are 2 convolutional layers with output sizes (127.32) and (126.64). The next layer is the global max pooling layer and ends with the later classifier.

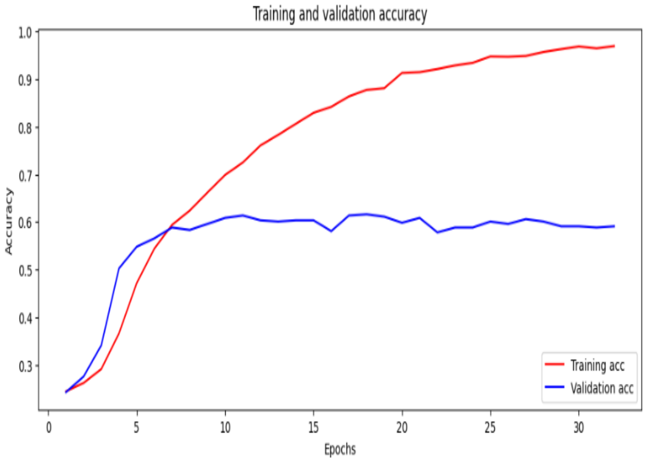

Figure 2. Accuracy BERT-CNN

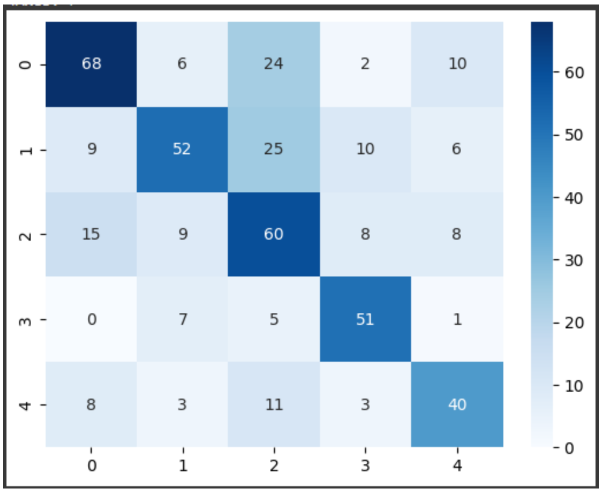

Figure 3. Confusion Metrics BERT-CNN

The result of training the model with parameters that match the previous training parameters is the resulting model accuracy of 61.45% with a training duration of 2 minutes for each existing epoch. The results of the training carried out were still below the baseline, which was used as a reference in this experiment, which at the baseline reached 72% for

BERT-CNN. In this research and the experiments carried out in this paper, there are differences in dividing train, test, and split data, as well as in normalizing datasets.

D. Result BERT-LSTM

Experiments with BERT-LSTM were carried out by forming several layers. The first layer is the input layer which is used to initialize the data train used. next is the preprocessing layer, the embedding layer which uses the BERT method. Then there are LSTM layer and following with dense layer. By using the previous training parameters, the results of the BERT-LSTM training are the accuracy obtained by 55%.

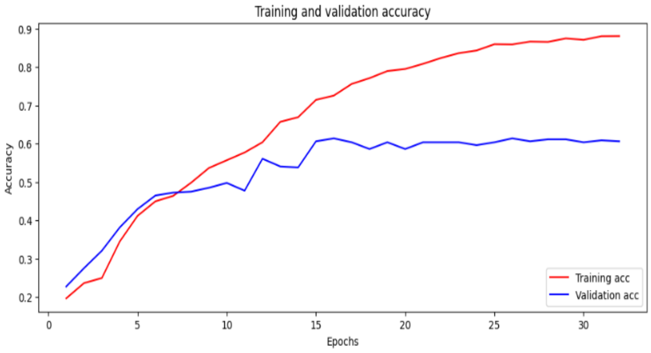

Figure 4. Accuracy BERT-LSTM

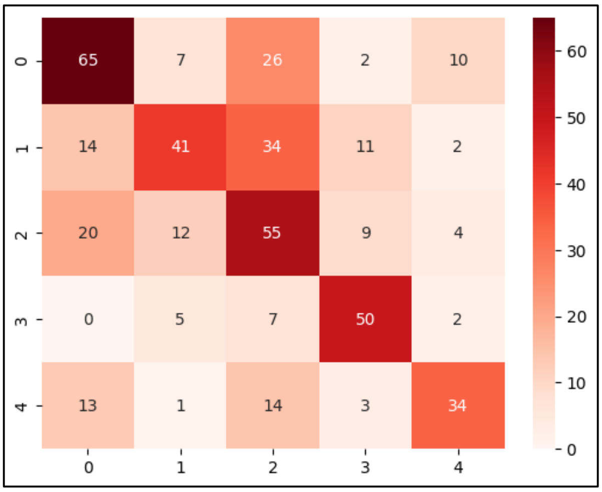

Figure 5. Confusion Metrics BERT-LSTM

Therefore, the overall results of the experiments conducted in this study can be seen in Table 2 below.

TABLE 2.
EXPERIMENT SUMMARY

Model	Accuracy	Precision	Recall	F1-Score
BERT-CNN	61.45%	62.53%	61.45%	61.48%
BERT-LSTM	55%	55%	57.32%	55.47%
Baseline	70.91%	73.02%	70.26%	71.29%

There are several reasons for the lower performance of BERT-LSTM compared to BERT-CNN, one of which is the determination of training parameters. In general, the parameters needed by CNN are relatively fewer than LSTM to produce a good model. However, in the experiment conducted in this study, the parameters used were limited to the application of loss calculation, accuracy metrics, and optimization. Therefore, with the few parameters used in this experiment, the performance of BERT-CNN is much better than BERT-LSTM. Suggestions for the future so that the performance of BERT-LSTM can achieve maximum results, the parameters used can be added again, such as by adding the appropriate regularization method to BERT. In addition to the limited parameters, the normalization method in the baseline is also different. The baseline uses an abbreviation dictionary, whereas in this experiment, normalization was performed by default based on the dictionary provided along with the dataset.

V. CONCLUSION

Based on the results of the experiments conducted in this study, it can be concluded that the performance of CNN combined with the BERT embedding method has a higher advantage compared to LSTM. This can be seen in the training results where the accuracy of BERT-CNN is 61% and LSTM is 55%. However, in this study, it was not able to exceed the performance of the baseline used as a research comparison.

Related analysis why the performance carried out in this experiment has not been able to exceed the baseline is because there is a difference in the split data ratio. In addition, the data normalization method used is also different. The baseline paper uses an abbreviation dictionary, while in this experiment, it uses the default dictionary from the dataset.

REFERENCES

[1] C. M. Annur, “Pengguna Twitter di Indonesia Capai 24 Juta hingga Awal 2023, Peringkat Berapa di Dunia?,” Databoks, 2023, [Online]. Available: https://databoks.katadata.co.id/datapublish/2023/02/27/pengguna-twitter-di-indonesia-capai-24-juta-hingga-awal-2023-peringkat-berapa-di-dunia

[2] S. Albawi, T. A. Mohammed, and S. Al-Zawi, “Understanding of a convolutional neural network,” Proc. 2017 Int. Conf. Eng. Technol. ICET 2017, vol. 2018-January, pp. 1–6, 2018, doi: 10.1109/ICEngTechnol.2017.8308186.

[3] K. Smagulova and A. P. James, “A survey on LSTM memristive neural network architectures and applications,” Eur. Phys. J. Spec. Top., vol. 228, no. 10, pp. 2313–2324, 2019, doi: 10.1140/epjst/e2019-900046-x.

[4] A. Glenn, P. LaCasse, and B. Cox, “Emotion classification of Indonesian Tweets using Bidirectional LSTM,” Neural Comput. Appl., vol. 35, no. 13, pp. 9567–9578, 2023, doi: 10.1007/s00521-022-08186-1.

[5] M. F. Heldiansyah and E. Winarko, “Emotion Detection on Indonesian Tweets Using CNN and Contextualized Word Embedding,” Proc. 2022 Int. Conf. Data Softw. Eng. ICoDSE 2022, pp. 53–58, 2022, doi: 10.1109/ICoDSE56892.2022.9972229.

[6] F. M. Rusli, R. Rismala, and H. Nurrahmi, “Emotion Classification on Indonesian Twitter Using Convolutional Neural Network (CNN),” 2021 9th Int. Conf. Inf. Commun. Technol. ICoICT 2021, pp. 213–218, 2021, doi: 10.1109/ICoICT52021.2021.9527447.

[7] M. S. Saputri, R. Mahendra, and M. Adriani, “Emotion Classification on Indonesian Twitter Dataset,” Proc. 2018 Int. Conf. Asian Lang. Process. IALP 2018, pp. 90–95, 2019, doi: 10.1109/IALP.2018.8629262.

[8] K. Sailunaz and R. Alhajj, "Emotion and sentiment analysis from Twitter text," J. Comput. Sci., vol. 36, p. 101003, 2019, doi: 10.1016/j.jocs.2019.05.009.

[9] H. Dalianis, "Evaluation Metrics and Evaluation," Clin. Text Min., no. 1967, pp. 45–53, 2018, doi: 10.1007/978-3-319-78503-5_6.

[10] A. Bruns and S. Stieglitz, "Metrics for understanding communication on Twitter," in Twitter and society [Digital Formations, Volume 89], A. Bruns, M. Mahrt, K. Weller, J. Burgess, and C. Puschmann, Eds., United States of America: Peter Lang Publishing, 2014, pp. 69–82. Accessed: Nov. 30, 2023. [Online]. Available: https://eprints.qut.edu.au/66326/.