2021 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA) | 978-1-6654-2680-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/DATABIA53375.2021.9650096

Sentiment Analysis for Movie Review in Bahasa Indonesia Using BERT

1st Dwi Fimoza
Department of Computer Science
Universitas Sumatera Utara
Medan, Indonesia
dwi.fimoza@gmail.com

2nd Amalia Amalia*
Department of Computer Science
Universitas Sumatera Utara
Medan, Indonesia
amalia@usu.ac.id

3rd T. Henny Febriana Harumy
Department of Computer Science
Universitas Sumatera Utara
Medan, Indonesia
hennyharumy@usu.ac.id

Abstract— **This study aims to analyze the sentiment in Indonesia Language towards the Gundala movie reviews on YouTube. However, sentiment analysis on YouTube comments are varying from positive, negative, and neutral comments which requires some automation in terms of classifying comments based on the polarity of sentiment. Sentiment analysis using traditional machine learning algorithms such as Naïve Bayes, SVM, etc cannot understand the context of comments in depth about the semantic of words because it only learns the given patters such as the frequency of occurrence of words. We need a transfer learning approach such as BERT (Bidirectional Encoder Representations from Transformers) which produces a bidirectional language model. The dataset used to do sentiment analysis goes through a pre-processing step which consists of case folding, data cleaning, tokenization, stop words removal, stemming, and normalization, using libraries from NLTK and Sastrawi. In this study, the hyperparameters used were 10 epochs, learning rate of 2e-5, and a batch size 16 based on experiments of hyperparameters used in another studies. In sentiment analysis, we will be using a multilingual-cased-model $\text{BERT}_{\text{BASE}}$ model and it was carried out with three experiments. During this experiment, the accuracy gained in first experiment is 66%, while the second experiment was 68%, and the third experiment was 66%. So, the average accuracy obtained is 66,7%.**

Keywords— Sentiment Analysis, Indonesian Films, Gundala, YouTube, Bidirectional Encoder Representations from Transformers, Deep Learning, Transformers

I. INTRODUCTION

These days Indonesian film industry is growing very rapidly. This development is shown by data from the Center for Film Development of the Ministry of Education and Culture or Pusbang, which states that the number of Indonesian films shows a growing trend in the last five years. The Creative Economy Agency noted that the number of Indonesian film viewers in 2018 was more than three times in 2015, which was more than 60 million viewers. Pusbang also stated that the number of Indonesian films reached an approximately total audience of over 1 million in 2018 for 14 films. However, with all these achievements, Indonesian films are still less competitive than foreign films in Indonesia. Based on a survey conducted in 2019 by IDN Times, a multi-platform news and entertainment media company, 50.3% of viewers criticized the predictable storyline. Technical in the film, such as cinematography, scoring, or actors, get 27.4%. Monotonous were also taken into consideration by 15% and acting by 7.3%. This problem can be observed further. One way to make observations is to look at the sentiments given to Indonesian films. This problem can be observed further by looking at the sentiments given to Indonesian films. With the highest active internet users in Indonesia, YouTube is a video-sharing social media that is one of the options to see public sentiment towards a film. In general, movie trailers are uploaded on this platform by the official account of the film production house, tempting the other YouTube users to leave a comment. Gundala is an Indonesian superhero film based on a comic character by the same name created by Hasmi, directed and written by Joko Anwar in 2019. As a science-fiction film, it gains a massive popularity, making its trailer trending when it is uploaded on YouTube. Gundala also gains more than one million viewers within seven days since it is released on cinemas. Up until now, Gundala got nine nominations and won some of it. Therefore, Gundala becomes an interesting topic to research especially sentiment analysis of people's reactions towards it.

The sentiments in the comments section can be analyzed and classified. Sentiment analysis studies the perspective, behavior, and feelings or emotions of a person towards individuals, problems, activities, products, or objects [1]. Methods in sentiment analysis are divided into two types, specifically learning-based and lexical-based. Learning-based uses training data and testing data, while lexical-based uses a dictionary (opinion lexicon). To use the learning-based method, traditional algorithmic methods such as Naïve Bayes, Support Vector Machine, and others can be used so that the machine learns the given data. However, the algorithm cannot understand the semantics of the existing words because it only learns the given patterns. Achieve better results requires a more in-depth approach.

Deep learning is a branch of machine learning that is part of artificial intelligence. Deep learning has many hidden layers, making the machine not only learns the data but also predict and represent the data better. Transfer learning approaches such as BERT (Bidirectional Encoder Representations from Transformers) can be a solution to gain more understanding. Transfer learning is a techique where a deep learning model is trained on big dataset and can be used to another similar task because it is already had knowledge especially how the language works from the training process. BERT is a trained language representation model developed by researchers at Google AI Language in 2018 [2]. BERT was developed based on deep learning techniques and various

methods such as semi-supervised learning, ELMo, ULMFiT, OpenAI Transformers, and Transformers. As the name implies, the use of Transformers can make the machine learn the contextual relationships between words in the text [3], understand, and convert the understanding obtained by self-attention mechanisms. BERT is pre-trained using BookCorpus (800M words) and English Wikipedia (2,500M words). There is also various released BERT pre-trained model available, for example bert-base, bert-base-multilingual, bert-base-chinese and many more. In related studies, research conducted by [4] produced a reasonably good results of 73.7% by using bert-based-multilingual-cased. However, the sentiment was classified into positive and negative categories only. The researcher also stated that the results were quite good compared to other algorithms such as Naïve Bayes. Another related studies also done by [5][6][7] shows good results in BERT. By using a transfer learning, according to this background, we proposed to use the deep learning method with the BERT language model to analyze sentiment towards the Gundala film obtained from the comments section in Indonesian on YouTube into three types of sentiment categories: positive, neutral, and negative. This paper is organized as follows: Section 2 introduces the related works in sentiment analysis, section 3 describes the research methodology used to do sentiment analysis to Gundala, section 4 discussed the results from the research, and section 5 concluded.

II. RELATED WORKS

Various studies have done sentiment analysis on films using BERT with different datasets and models. Research by [4] used BERT to perform a sentiment analysis on the cornelledu dataset in English. This study resulted in an accuracy value of 73%. Research using BERT was also conducted by [5]. The dataset used in this study is in English and was obtained from IMDB. This study gave satisfactory results, namely 89%. Research related to films was also carried out by [6]. This study analyzed sentiment towards films using a dataset in English obtained from Rotten Tomatoes. The accuracy obtained by this study was 94% using SST-2 model and 83.9% using SST-5 model for BERTBASE. Research conducted by [7] used BERT to analyze the sentiments of tweets to determine their relevance to hashtags about floods that occurred in Jakarta. In this study, the dataset used was in Indonesian. The accuracy obtained by the training dataset is 90% meanwhile the testing dataset is 79%. The author states that the training results showed considerably good accuracy even though it has a lot of noise that affects the accuracy level.

III. METHODOLOGY

This section discusses the methodology used in the process of sentiment analysis on Indonesian films using BERT. The methodology used in this study consists of several steps as shown in Fig. 1.

A. Scraping Dataset

The first step is to collect the dataset used in this study. The dataset was obtained from YouTube by taking comments on the video entitled "Official Trailer GUNDALA (2019)" which was uploaded on July 20, 2019.

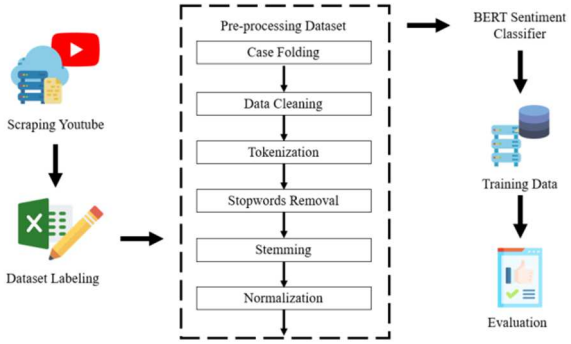

Fig. 1. Methodology used in this research.

B. Dataset Labeling

In sentiment analysis using the supervised learning method, a dataset that already has a label or is annotated by an annotator is needed. The labeling is done to determine the comments into three categories: positive, neutral, and negative. Positive sentiment was given a value of 2, the neutral sentiment was given a value of 1, and the negative sentiment was given a value of 0. The labeling was carried out by five annotators.

C. Dataset Preprocessing

Dataset preprocessing in this study consists of case folding, data cleaning, tokenization, stopwords removal, stemming, and normalization.

a) Case folding: Case folding is used by making all uppercase letters in the dataset into lowercase. By making all words lowercase, it will be beneficial to make generalizations [8].

b) Data cleaning: Every sentence in the dataset is cleaned of everything that can affect the results of the analysis, such as words with two or more repeated characters, links, usernames (@usernames), hashtags (#), numbers, symbols, extra spaces, punctuation marks, and numbers. To perform data cleaning, this study uses regular expressions.

c) Tokenization: Tokenization is a process carried out to break sentences into pieces of words, punctuation marks, and other meaningful expressions by the provisions of the language used.

d) Stopwords removal: Stopwords removal is a process carried out to remove words that have no additional meaning.

e) Stemming: Stemming is a step to change words that have affixes into their root word by removing affixes such as prefixes, suffixes, and confixes.

f) Normalization: The normalization stage is the stage where the dataset that has slang words is converted into standard words or according to spelling. This normalization process used a dictionary by [9].

D. BERT (Bidirectional Encoder Representations from Transformers)

BERT model architecture is a multi-layer bidirectional Transformer, consists of a stack of encoders. The mechanism in the Transformer, namely self-attention, functions to change the understanding of other related words into words that the mechanism will process. BERT is different from the

directional model, which looks at the order of the text from left-to-right, right-to-left, or a combination of left-to-right and right-to-left. BERT will capture both left and right context. Hence, a bi-directionally trained language model can have a deeper understanding of context than a one-way language model. BERT can be trained to understand a language and can also be fine-tuned to learn specific tasks. The training at BERT consists of two stages: pre-training and fine-tuning. The first stage, namely pre-training, is where BERT is made to understand and learn the language and its context; meanwhile, fine-tuning is where BERT has adjusted the hyperparameters to perform a similar task. In pre-training, two unsupervised tasks are performed simultaneously, namely Masked Language Model and Next Sentence Prediction. Masked Language Modelling (Masked LM) gives mask with [MASK] token to random word words in sentences with low probability, and BERT will try to predict the masked words based on the context given by other words on left and right. In BERT, the model can also accept a pair of sentences and is trained to predict if the second sentence in the pair is the following sentence in the original document. During training, 50% of the input is a pair of sentences, where the second sentence is the next sentence from the original document. While the other 50% are sentences taken randomly from the corpus as the second sentence.

BERT uses WordPiece embeddings with 30,000 vocabulary tokens. The first token of each sequence is always a special classification token which is [CLS]. In its implementation, two size models exist in BERT, namely $BERT_{BASE}$ and $BERT_{LARGE}$ shown in Fig. 2. $BERT_{BASE}$ has 12 layers of encoder, 12 self-attention heads, hidden size 768, and 110M parameters. In comparison, $BERT_{LARGE}$ has 24 layers, 16 self-attention heads, 1024 hidden sizes, and 340M parameters. In this study, the model used is bert-multilingual-base-cased which has 119,547 data from Wikipedia. Models specifically made using Indonesian only are still limited, therefore using this model has become an option, as it is also used in related studies such as [7] and [10]. This model supports 104 languages, including Indonesian. There are four types of vocabularies in bert-multilingual-base-cased: all words, sub-words that appear in front of the word or separately, sub-words that are not in front of the word, and individual characters.

For a dataset containing sentences to be processed by BERT, the sentence must represent input to the BERT done by the BERT tokenizer. The tokenizer results will be input to each encoder. Fig. 3 shows the process of BERT Tokenizer BERT accepts a fixed, same length for each of its inputs with maximum sentence sequence length, which is 512. It is because the encoder on the Transformer only produces output with dimensions of 512 only. If the length of the sentence is more than the maximum length, it will truncate the sentence. Fig. 4 shows the input and output representation in BERT. Therefore, the input representation can be seen in Fig. 5.

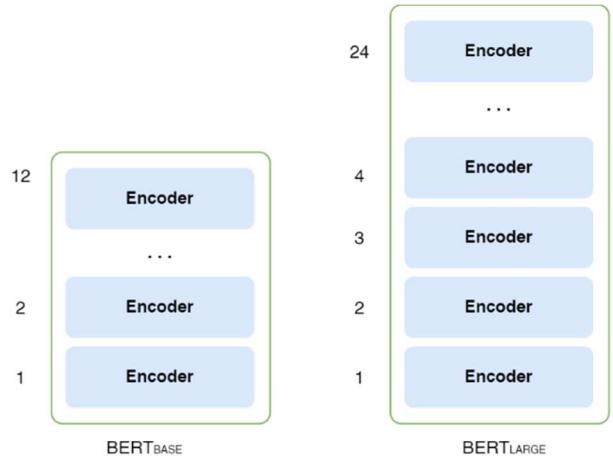

Fig. 2. $BERT_{BASE}$ and $BERT_{LARGE}$.

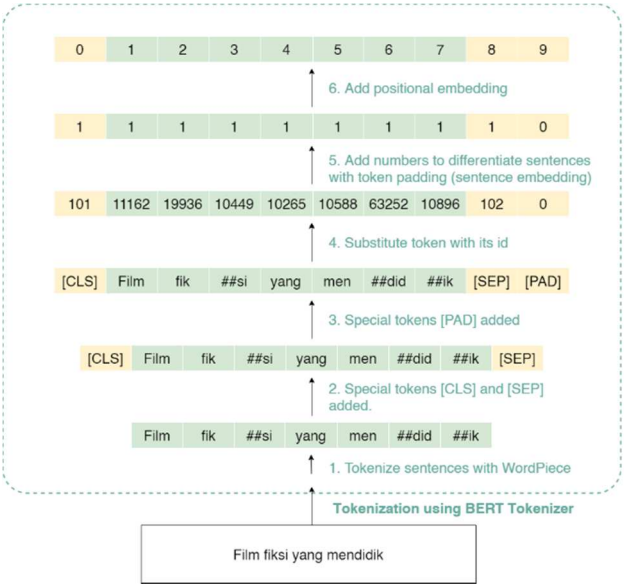

Fig. 3. Tokenization process.

Fig. 4. Input and output representation in BERT.

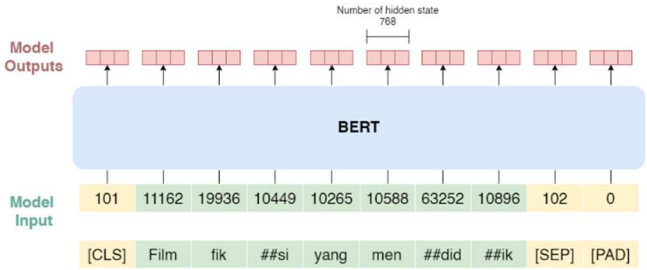

Fig. 5. Input and output representation in BERT.

After going through all the encoders in $BERT_{BASE}$, the output is entered into a fully connected neural network as a classifier with a softmax function to obtain the probability from comments to sentiment categories as shown in Fig. 6. The softmax function will change the logits or output in the form of a rough probability prediction of the sentence to be classified into probability by taking the exponent of each logits value so that the total probability is exactly 1. So the probability value will be between 0 and a positive number. The softmax function is given as below:

$$softmax(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad (1)$$

with $i = 1, \ldots, K$ where:
$z = (z_1, \ldots, z_k)\ \epsilon\ \mathbb{R}^K$ = logits
e^{z_i} = exponential of each element of the input vector.
$\sum_{j=1}^{K} e^{z_j}$ = normalization process to ensure that all the output values of softmax will add up to exactly 1 and each value is in the range (0,1)

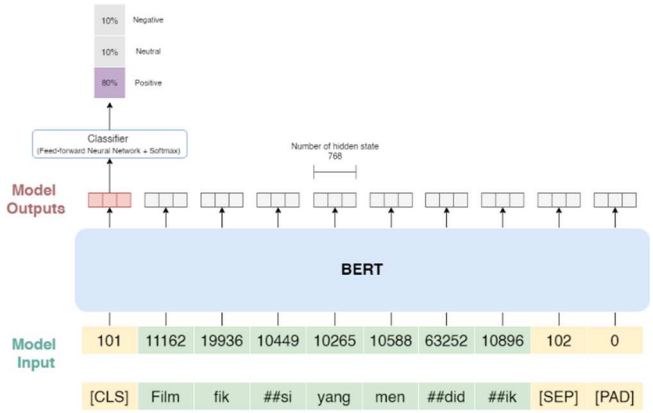

Fig. 6. Classification process using BERT.

After BERT is pre-trained with the existing dataset, training is carried out. At this stage, BERT is fine-tuned by adjusting its hyperparameters. The hyperparameters used in BERT training include:

- Batch size is the number of samples entered into the network before the weight is adjusted. The larger the batch size, the longer it will take to complete one batch [11].
- Epoch is the number of times the network viewed the entire dataset. One epoch occurs when all instances have passed through the network, both forward and backward passes.
- Learning rate determines how much weight on the neural network will be changed.

The sentiment analysis process with BERT is carried out using hyperparameters. There is a recommendation of possible values of hyperparameters. However, the optimal hyperparameters are task-specific. Hence, we will use hyperparameters based on related studies. To determine which epoch to use, moreover, a test with epochs was carried out based on research by [5][12][13][14][10]. We chose batch size 16 because the larger the batch size, the longer the time required to complete one batch[11]. In addition, the use of a learning rate of 2e-5 was chosen because it can make BERT overcome the problem of catastrophic forgetting [12], which is a problem in which the understanding gained from pre-training is erased while learning new information or data.

In this research, we also need a baseline to compare the performance achieved by BERT. One of the traditional machine learning algorithms, Multinomial Naïve Bayes, tends to work well for text classification and is often used as a baseline for classification [15]. Multinomial Naïve Bayes is one of Naïve Bayes algorithm. It is based on the Bayes Theorem, which calculates the probability of data to each label or category we have decided. In this case, the labels would be negative, neutral, and positive. The label with highest probability will be the output. The algorithm will be used for comparison and evaluation with the results obtained by BERT. Thus, the process will not be explained further. The dataset used is the same as the dataset used in the BERT implementation.

IV. Results and Discussion

In this research, we aimed to do sentiment analysis towards Gundala, an Indonesian film, using BERT. After scraping, we found 9,484 comments from the comments column with a date range from the oldest (July 20th, 2019 to August 10th, 2020). Furthermore, the comments go through a sentence-splitting process to separate the comments into a sentence. This process produced a dataset with 12,852 sentences stored in an excel worksheet. The dataset was then annotated by five annotators with three categories: positive, neutral, and negative. Comments that use languages other than Indonesian are deleted in the labeling process. Thus, the dataset used for sentiment analysis is 10,652 comments with 816 negative comments, 4759 neutral comments, and 5077 positive comments. The stopwords removal stage uses the Indonesian language stopwords library provided by NLTK and a dictionary made by Tala (link). However, there are some words in the Tala dictionary that are not used because they will affect sentiment analysis, such as "sangat", "terlalu", "kurang", and "sekali" which are booster words for a sentiment. In addition, the words "enggak" and "tidak" are also removed because they can change the polarity of an opinion. In the normalization stage, we add new words based on slang words obtained when annotating the dataset. Some examples of the dataset can be seen in Table 1.

TABLE I. Dataset

Comment	Sentiment
banyak yang beli film tayang negara	1
cium aroma aroma kenalannya film gundala	1
wajib menonton mah	2
keren	2
keren film indonesia pasuk era mantap	2

After the normalization process, we input the dataset into BERT and then going through classifier consists of feed neural network and softmax function. Suppose, a comment written is “Film fiksi yang mendidik” which means “This film is an educational fiction film” has logits vector = (7 5 0). We convert the logits into a probability distribution of a comment to be classified into negative, neutral, and positive categories. Therefore, the steps to get the probability using softmax function are:

i. Calculate every exponential of each element in the vector.

$$e^{z_1} = e^7 = 1096.6$$
$$e^{z_2} = e^5 = 148.4$$
$$e^{z_3} = e^0 = 1.0$$

ii. Normalize the values by adding up all the exponentials.

$$\sum_{j=1}^{K} e^{z_j} = 1096.6 + 148.4 + 1 = 1246$$

iii. Divide the exponential of each element by normalization to get the softmax output of each element.

$$\sigma(\overrightarrow{z})_1 = \frac{1096.6}{1246} = 0.8801$$
$$\sigma(\overrightarrow{z})_2 = \frac{148.4}{1246} = 0.1191$$
$$\sigma(\overrightarrow{z})_3 = \frac{1}{1246} = 0.0008$$

To determine whether the predicted result of each probability is 1, we sum all the probabilitie:$\sigma(\overrightarrow{z})_1 + \sigma(\overrightarrow{z})_2 + \sigma(\overrightarrow{z})_3$ = 0.8801+0.1191+0.0008 = 1. The results of the probability prediction will show the probability of the comments to the sentiment category. It means that the comment has a probability of 0.8801 for the positive sentiment category, while for the negative sentiment, it is 0.1191, and the neutral sentiment is 0.0008. Therefore, this comment is considered as positive comment. This process also applies to all comments in the dataset.

We did three experiments based on research by [5][12][13][14][10], which is 4, 10, and 16 epochs. Each of the experiments used batch size 16 and learning rate 2e-5. Fig. 7, 8, and 9 respectively show the results from each epoch. Based on this comparison, 10 epochs is chosen since the accuracy is relatively higher compare to 4 epoch and 16 epoch. Therefore, in this study we use hyperparameters: batch size 16, 10 epoch, and learning rate 2e-5.

	precision	recall	f1-score	support
negative	0.43	0.29	0.34	42
neutral	0.78	0.77	0.78	253
positive	0.82	0.87	0.85	238
accuracy			0.78	533
macro avg	0.68	0.64	0.66	533
weighted avg	0.77	0.78	0.77	533

Fig. 7. Accuracy from 4 epoch.

	precision	recall	f1-score	support
negative	0.65	0.40	0.50	42
neutral	0.78	0.77	0.77	253
positive	0.78	0.84	0.81	238
accuracy			0.77	533
macro avg	0.74	0.67	0.69	533
weighted avg	0.77	0.77	0.77	533

Fig. 8. Accuracy from 10 epoch.

	precision	recall	f1-score	support
negative	0.43	0.31	0.36	42
neutral	0.76	0.73	0.74	253
positive	0.76	0.83	0.79	238
accuracy			0.74	533
macro avg	0.65	0.62	0.63	533
weighted avg	0.73	0.74	0.73	533

Fig. 9. Accuracy from 16 epoch.

The hyperparameters used in this study can be seen in Table 2.

TABLE II. HYPERPARAMETERS

Parameters	Value
Batch size	16
Epoch	10
Learning rate	2e-5

At each stage of training, validation, and testing, the dataset used is randomized. It will affect the process as every training, validation, and testing did not have the exact same dataset. After going through the iterative process per epoch on the model, the results that get the best accuracy values are stored and become a reference for conducting sentiment analysis. Fig. 10, Fig. 11, and Fig. 12 show the comparison results between those obtained during training and validation from each experiment.

In conducting sentiment analysis with BERT, the model with the best accuracy value is then tested using dataset testing. After getting the accuracy value, the model will then try to predict it using the testing dataset. Fig. 13, Fig. 14, and Fig. 15, respectively, show the test results in each experiment. Based on the testing obtained, it is known that the overall accuracy results using BERT reached 66%, 68%, and 66% in the first, second, and third experiments, respectively. With that, the average obtained from sentiment analysis with BERT is 66.7%.

In comparison, the results obtained by Multinomial Naïve Bayes can be seen in Fig. 16. The result of BERT is lower than the result obtained by Multinomial Naïve Bayes, which is 72%. The difference in accuracy results obtained by the system is influenced by the dataset that is randomized when it was divided in each experiment into datasets for training, testing, and evaluation. In addition, the precision for negative sentiments tends to be lower than the precision for positive and neutral sentiments.

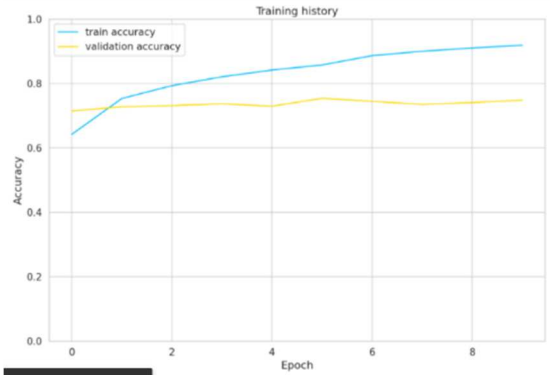

Fig. 10. Training and validation performance results in first try.

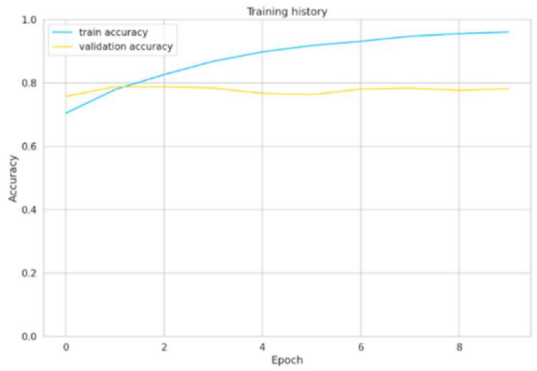

Fig. 11. Training and validation performance results in second try.

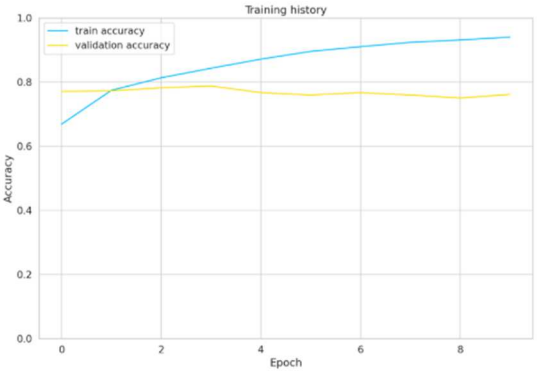

Fig. 12. Training and validation performance results in third try.

	precision	recall	f1-score	support
negative	0.38	0.31	0.34	39
neutral	0.76	0.79	0.77	252
positive	0.84	0.84	0.84	242
accuracy			0.77	533
macro avg	0.66	0.64	0.65	533
weighted avg	0.77	0.77	0.77	533

Fig. 13. Accuracy from the first try.

	precision	recall	f1-score	support
negative	0.48	0.36	0.41	42
neutral	0.74	0.77	0.75	238
positive	0.80	0.81	0.81	253
accuracy			0.76	533
macro avg	0.68	0.65	0.66	533
weighted avg	0.75	0.76	0.75	533

Fig. 14. Accuracy from the second try.

	precision	recall	f1-score	support
negative	0.43	0.53	0.47	34
neutral	0.74	0.74	0.74	238
positive	0.81	0.78	0.79	261
accuracy			0.75	533
macro avg	0.66	0.68	0.67	533
weighted avg	0.75	0.75	0.75	533

Fig. 15. Accuracy from the third try.

The confusion matrix diagram for each experiment can be seen in Fig. 17, Fig. 18, Fig. 19. Based on the confusion matrix diagram for the three experiments, the system has difficulty classifying negative comments. Whereas in positive and neutral comments, the system tends to be able to classify sentiment correctly. This happens because the number of comments is not balanced between negative, neutral, and positive comments. Based on these observations, the author tries to conduct a sentiment analysis with the same dataset, but the ratio between each comment is the same as the results shown in Fig. 20.

	precision	recall	f1-score	support
0	0.63	0.16	0.25	152
1	0.76	0.75	0.76	941
2	0.77	0.87	0.82	1038
accuracy			0.77	2131
macro avg	0.72	0.59	0.61	2131
weighted avg	0.76	0.77	0.75	2131

Fig. 16. Accuracy from Multinomial Naïve Bayes.

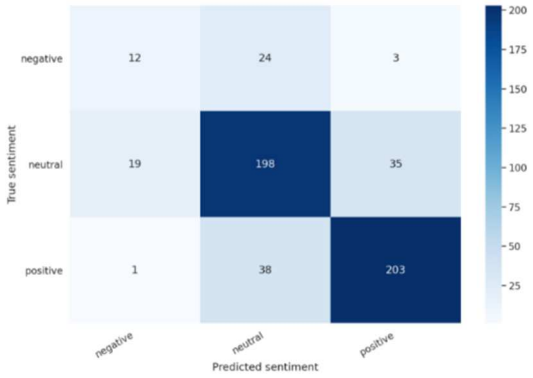

Fig. 17. Confusion matrix diagram in first try.

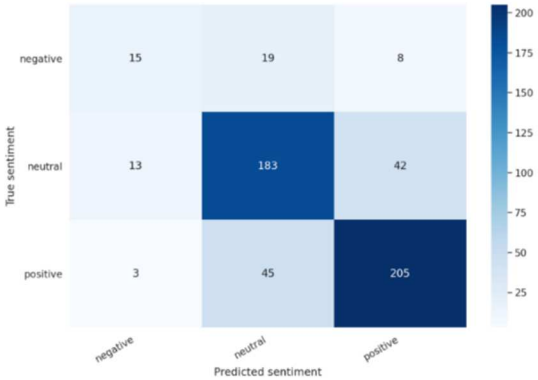

Fig. 18. Confusion matrix diagram in second try.

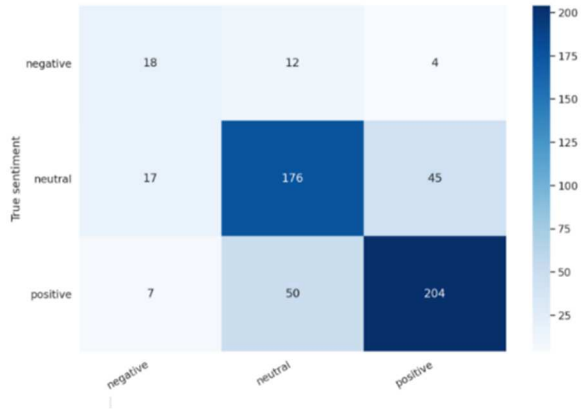

Fig. 19. Confusion matrix diagram in third try.

Based on these results, it can be concluded that the result of sentiment analysis using bert-multilingual-base-cased on unbalanced datasets has a lower accuracy than using bert-multilingual-base-cased on balanced datasets. It also shows that a balanced dataset affects because it improves performance and accuracy as the accuracy obtained is 74% even with the same model. In addition to comparing with the baseline and bert-multilingual-base-cased on balanced dataset, a comparison was also made with IndoBERT [16] using the same unbalanced data, as shown in Fig. 21 to see and evaluate performance. Based on this comparison as shown in Fig. 13, Fig. 14, and Fig. 15, BERT with bert-multilingual-base-cased with the unbalanced dataset has not provided higher and better accuracy results than Multinomial Naïve Bayes or IndoBERT as in the average accuracy is only 66.7%, especially on negative comments. However, the precision and recall results of negative comments obtained by the bert-multilingual-base-cased on a balanced dataset are better than Multinomial Naïve Bayes. It also indicates that using a balanced dataset on BERT will give a better result at predicting negative sentiment or another sentiment.

	precision	recall	f1-score	support
negative	0.71	0.70	0.70	46
neutral	0.70	0.74	0.72	62
positive	0.83	0.78	0.80	49
accuracy			0.74	157
macro avg	0.74	0.74	0.74	157
weighted avg	0.74	0.74	0.74	157

Fig. 20. Accuracy from sentiment analysis using balanced dataset.

	precision	recall	f1-score	support
negative	0.56	0.45	0.50	42
neutral	0.80	0.75	0.77	253
positive	0.78	0.87	0.82	238
accuracy			0.78	533
macro avg	0.71	0.69	0.70	533
weighted avg	0.77	0.78	0.77	533

Fig. 21. Accuracy from sentiment analysis using IndoBERT.

V. CONCLUSIONS

Based on the test results of the implementation of the BERT bert-multilingual-base-cased in conducting sentiment analysis, we found that the accuracy obtained was 66%, 68%, and 66% with three experiments with the selection of hyperparameters, namely batch size 16, learning rate 2e-5, and epoch 10. Based on the test results between 4, 10, and 16 epochs, the epoch that gives good results is 10 epoch. We decided that 10 epoch is used to analyze sentiment. We also find out that an unbalanced dataset affects the accuracy obtained when implementing BERT. Although the number of balanced datasets is less than that of unbalanced datasets, the accuracy obtained is better at 74%. This research also found that BERT produces a lower accuracy than using Multinomial Naïve Bayes, which is 72%. We also do sentiment analysis using IndoBERT model. The result shows accuracy in 71%. According to this result, sentiment analysis using BERT bert-multilingual-base-cased on the Indonesian dataset has not given better results for Indonesian. In the future work, we are looking forward to doing sentiment analysis in more Indonesian datasets with a significant same amount of positive, negative, and neutral data using BERT model in Indonesian too.

REFERENCES

[1] A. S. H. Basari, B. Hussin, I. G. P. Ananta, and J. Zeniarja, "Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization," *Procedia Eng.*, vol. 53, pp. 453–462, 2013, doi: 10.1016/j.proeng.2013.02.059.

[2] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, no. Mlm, pp. 4171–4186, 2019.

[3] A. Vaswani *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Nips, pp. 5999–6009, 2017.

[4] C. A. Putri, "Analisis Sentimen Review Film Berbahasa Inggris Dengan Pendekatan Bidirectional Encoder Representations from Transformers," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 6, no. 2, pp. 181–193, 2020, doi: 10.35957/jatisi.v6i2.206.

[5] S. Abdul, Y. Qiang, S. Basit, and W. Ahmad, "Using BERT for Checking the Polarity of Movie Reviews," *Int. J. Comput. Appl.*, vol. 177, no. 21, pp. 37–41, 2019, doi: 10.5120/ijca2019919675.

[6] M. Munikar, S. Shakya, and A. Shrestha, "Fine-grained Sentiment Classification using BERT," *Int. Conf. Artif. Intell. Transform. Bus. Soc. AITB 2019*, vol. 1, pp. 1–5, 2019, doi: 10.1109/AITB48515.2019.8947435.

[7] W. Maharani, "Sentiment Analysis during Jakarta Flood for Emergency Responses and Situational Awareness in Disaster Management using BERT," *2020 8th Int. Conf. Inf. Commun. Technol. ICoICT 2020*, 2020, doi: 10.1109/ICoICT49345.2020.9166407.

[8] D. Jurafsky and J. H. Martin, *Speech and language processing*. 2019.

[9] N. Aliyah Salsabila, Y. Ardhito Winatmoko, A. Akbar Septiandri, and A. Jamal, "Colloquial Indonesian Lexicon," *Proc. 2018 Int. Conf. Asian Lang. Process. IALP 2018*, pp. 226–229, 2019, doi: 10.1109/IALP.2018.8629151.

[10] M. R. Yanuar and S. Shiramatsu, "Aspect Extraction for Tourist Spot Review in Indonesian Language using BERT," *2020 Int. Conf. Artif. Intell. Inf. Commun. ICAIIC 2020*, pp. 298–302, 2020, doi: 10.1109/ICAIIC48513.2020.9065263.

[11] D. Osinga, *Deep Learning Cookbook*, no. June. 2018.

[12] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to Fine-Tune BERT for Text Classification?," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol.

11856 LNAI, no. 2, pp. 194–206, 2019, doi: 10.1007/978-3-030-32381-3_16.

[13] H. Xu, B. Liu, L. Shu, and P. S. Yu, "BERT post-training for review reading comprehension and aspect-based sentiment analysis," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, pp. 2324–2335, 2019.

[14] 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*, 2020.

[15] S. Xu, Y. Li, and Z. Wang, "Bayesian Multinomial Na ï ve Bayes Classi fi er to Text Classi fi cation," no. 15, 2017, doi: 10.1007/978-981-10-5041-1.

[16] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," 2020, [Online]. Available: http://arxiv.org/abs/2011.00677.