Jurnal Ilmu Komputer dan Informatika (JIKI) Vol. 5, No. 2, December 2025, Page. 127-142
P-ISSN: 2807-6664 https://jiki.jurnal-id.com
E-ISSN: 2807-6591 DOI: https://doi.org/10.54082/jiki.307

Comparative Analysis of Bidirectional Encoder Representations from Transformers Models for Twitter Sentiment Classification using Text Mining on Streamlit

Ahmad Fajar Tatang*[1], Mohammad Hasbi Assidiqi[2]

[1,2]Computer Science Department, Faculty of Computing and Information Technology, King Abdulaziz University, Saudi Arabia

Email: [1]atatang0001@stu.kau.edu.sa

Received : Jun 9, 2025; Revised : Nov 20, 2025; Accepted : Nov 22, 2025; Published : Dec 11, 2025

Abstract

Social media platforms like Twitter have become highly influential in shaping public opinion, making sentiment analysis on tweet data crucial. However, traditional techniques struggle with the nuances and complexities of informal social media text. This research addresses these challenges by conducting a comparative analysis between the non-optimized BERT (Bidirectional Encoder Representations from Transformers) model and the BERT model optimized with Fine-Tuning techniques for sentiment analysis on Indonesian Twitter data using text mining methods. Employing the CRISP-DM methodology, the study involves data collection through Twitter crawling using the keyword biznet, data preprocessing steps such as case folding, cleaning, tokenization, normalization, and data augmentation, with the dataset split into training, validation, and testing subsets for modeling and evaluation using the IndoBERT-base-p1 model specifically trained for the Indonesian language. The results demonstrate that the Fine-Tuned BERT model significantly outperforms the non-optimized BERT, achieving 91% accuracy, 0.91 precision, 0.90 recall, and 0.91 F1-score on the test set. Fine-Tuning enables BERT to adapt to the unique characteristics of Twitter sentiment data, allowing better recognition of language and context patterns associated with sentiment expressions. The optimized model is implemented as a web application for practical utilization. This research affirms the superiority of Fine-Tuned BERT for accurate sentiment analysis on Indonesian Twitter data, providing valuable insights for businesses, governments, and researchers leveraging social media data.

Keywords: *BERT, Fine-Tuning, Sentiment Analysis, Social Media, Text Mining, Twitter*

1. INTRODUCTION

Social media platforms, such as Twitter, have become a powerful force in shaping public opinion and influencing decision-making processes. The vast amount of textual data generated on these platforms presents both opportunities and challenges for sentiment analysis. Accurately determining the sentiment expressed in tweets can provide valuable insights for businesses, governments and researchers [1], [2]. However, the inherent complexity and nuances of natural language, coupled with the informal and often ambiguous nature of social media text, pose significant hurdles for traditional sentiment analysis techniques.

To overcome the challenges in sentiment analysis, researchers have explored various methods and algorithms, including deep learning models and ensemble techniques. One promising approach is to use sophisticated language models such as Bidirectional Encoder Representations from Transformers (BERT) with Fine-Tuning methods [3]. Although BERT itself is a powerful model, it has some drawbacks, such as limitations in capturing the nuances and special characteristics of certain domains or tasks such as sentiment analysis, as well as differences in data distribution between BERT training data and sentiment analysis data [4]. By using relevant data, BERT can learn to understand the unique characteristics of language and context associated with sentiment, such as expressions of emotion, sarcasm, or opinion [5]. Refinement helps BERT recognize important aspects in language and context that have an effect on overall sentiment, thus improving its ability to better understand and analyze sentiment [6]. Through a gradual learning process using data relevant to sentiment analysis, BERT with fine-tuning can learn a wider and more diverse range of language and context patterns, thereby improving its accuracy and overall performance in sentiment analysis tasks on social media data compared to using BERT without fine-tuning [7].

In several previous studies, various classification methods have been implemented in the real world. Some of the frequently used classification algorithms are BERT. BERT algorithm with Fine-Tuning was used by Muhammad Bilal [8], to classify online customer reviews into useful or not useful categories. The use of BERT with Fine-Tuning in the study aims to get a more general approach in predicting the usefulness of online customer reviews compared to traditional machine learning methods.

The main objective of this research is to conduct a comparative analysis of the non-optimized BERT model and the optimized BERT model with Fine-Tuning for sentiment analysis on Twitter data using text mining techniques. By utilizing the power of text mining, this research aims to extract relevant features and patterns from textual data, thus enabling a comprehensive evaluation of both approaches. The ultimate goal is to determine which method yields better performance based on the Precision, Recall, and Accuracy results that will be compared, so that conclusions can be drawn regarding the best algorithm for classifying the sentiment of tweets based on user-supplied keywords or phrases.

2. METHOD

The research stages using CRISP-DM include Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Application. Business Understanding is essential for problem analysis. Data Understanding is necessary for data collection. Data Preparation is essential for manual data elimination, preprocessing, and labeling. Modeling is essential for data modeling, word weighting, and classification using the BERT model. Evaluation is essential to test the validity of the data using a confusion matrix. Deployment is essential to ensure the web application can run properly using the Python language and Streamlit framework. By doing deployment, the web application can be published so that it can be accessed and used by other users.

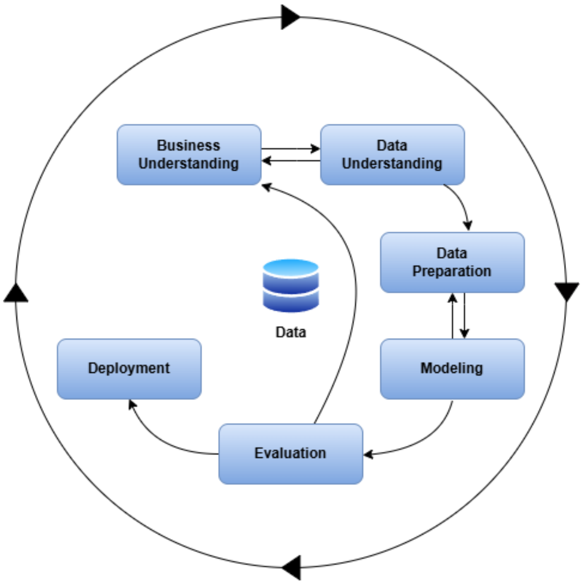


Figure 1. CRISP-DM Diagram

Figure 1 shows the stages of CRISP-DM [9]. This research aims to classify positive, negative, and neutral sentiments from tweets using Unoptimized BERT and Optimized BERT through the Fine-Tuning method.

2.1. Business Understanding

This first stage was to identify and understand the existing challenges in Twitter sentiment analysis using text mining techniques; the main challenge relates to the efficiency of data collection, where the previous process done manually was not only time consuming but also prone to errors. As a solution this research proposes the implementation of Node Package Manager (npm) tweet-harvest package run through Python, enabling automated and real-time data collection; which significantly speeds up the process while increasing the volume of data that can be processed.

The second challenge is improving the accuracy of sentiment analysis; where traditional analysis is often unable to capture complex language nuances and specific sentence contexts. To address this, this research integrates the use of the IndoBERT model, a language model that has been specially trained for the Indonesian language [10]. This model is designed to be more sensitive to language context and nuances, so it is expected to improve accuracy in sentiment classification. This research aims to compare the performance of the standard BERT model with the BERT model optimized through Fine-Tuning techniques; evaluate the effectiveness of this technique in improving model performance on Twitter sentiment analysis.

2.2. Data Understanding

The second stage is data collection, which is very important in this research. The data used comes from Twitter social media. The data collection process is done by crawling tweet data based on keywords inputted by the user.

The users can enter keywords that are relevant to the topic or subject they want to analyze, then tweet data containing those keywords will be collected for further processing. In this research, the dataset used is obtained from crawling tweet data using the keyword biznet in Indonesian. The selection of the keyword biznet is based on the consideration that the word refers to an Indonesian telecommunication company, so it is expected that people's sentiment towards the keyword tends to be neutral, not too positive or negative. In addition, with Biznet as an Indonesian company, the tweet data obtained is expected to be in Indonesian, so it is in accordance with the qualifications of the IndoBERT model used in this research to perform sentiment analysis on Indonesian tweet data.

Crawling is a text mining technique to automatically collect text data from online resources such as social media, websites, and blogs based on certain keywords [11]. This process involves a script or computer program that searches for resources, identifies, and downloads relevant text data.

In this research, crawling is used to collect tweet data from Twitter by utilizing npm tweet-harvest. This technique allows for efficient and automated collection of large amounts of text data as opposed to manual collection which is time and labor consuming.

2.3. Data Preparation

In the third stage, the collected Twitter data will go through a series of preprocessing processes to prepare it before being classified using the BERT model that has not been optimized and the BERT model that has been optimized with Fine-Tuning techniques. This preprocessing stage is carried out because it can improve the accuracy of the BERT model; the preprocessing process in this study includes case folding, data cleaning, tokenization, normalization, data augmentation, and split dataset using techniques such as Fine-Tuning Tokenizer [12].

2.3.1. Case Folding

This stage is used to convert all characters into lowercase letters, because the data obtained is not always structured and consistent in the use of capital letters, so case folding is done to equalize the use of capital letters [13]. Case folding is done using the lower() function that is available in the Python library. This can help in the sentiment classification process by eliminating capitalization differences.

2.3.2. Data Cleaning

The Data Cleaning stage is very important in the sentiment classification process from social media data. This process serves to clean noise or irrelevant information such as mentions, hashtags, and URLs in text data before further processing, so as to increase the accuracy of the classification model in determining sentiment [14]. In this study, the Data Cleaning process is carried out by removing mentions (@username), hashtags (#), and irrelevant URLs in order to avoid bias or noise in the form of words, symbols, or characters that are not needed. When modeling is done from data that is not optimal in the presence of excessive noise, the prediction accuracy of the model can decrease and become inefficient in analyzing sentiment.

2.3.3. Tokenization

Tokenization is the process of separating text into smaller semantic units (such as words or phrases) before classification. It is important to isolate "tokens" or semantic components that may affect sentiment classification [15].

2.3.4. Normalization

Normalization is an important process in sentiment analysis to make the text more consistent and uniform [16]. Normalization steps include: replacing nonstandard or slang words into standard Indonesian forms (e.g., "*gak*" becomes "*enggak*"), correcting typos (e.g., "*sech*" becomes "*sih*"), removing unnecessary special characters (such as inappropriate punctuation marks, emojis, and other symbols), and changing all letters to lowercase to ensure consistency (such as "INDONESIA" and "Indonesia" are considered the same). With normalization, the text data becomes cleaner and more structured, which improves the accuracy of the model in sentiment classification.

2.3.5. Data Augmentation

Data augmentation is performed using the back translation technique, which translates text from Indonesian to English, then translates it back to Indonesian [17]. The process includes converting the text to tensor, translating the tensor to English, translating the translation to Indonesian, ensuring the meaning does not change significantly, and merging the augmented data with the original dataset. This technique increases data diversity by preserving context and meaning, helping the model learn new patterns to improve sentiment classification performance.

2.3.6. Split Dataset

The dataset obtained from the crawling process, using the keyword biznet*,* was divided into three subsets for further processing and analysis. The first subset, 70% of the data, is used as a training data subset to train the model to learn patterns from the data, both for the BERT sentiment classification model before optimization and BERT with Fine-Tuning, where the more training data, the more patterns the model can learn, thus improving classification accuracy [18]. The second subset, 20% of the data, is the validation data subset used to monitor the performance of the model during training and prevent overfitting, by periodically evaluating the model during training, so that if performance decreases, adjustments can be made [19]. Meanwhile, the third subset, 10% of the data, is a test data subset used to evaluate the final performance of the model after training is complete, separate from the training-validation data, and used to calculate evaluation metrics such as accuracy, precision, recall, and F1-score.

2.4. Modeling

The fourth stage is Modeling. This fourth stage is the core of data classification. The classification used in this study is BERT without optimization and BERT optimization with Fine-Tuning.

2.4.1. BERT without Optimization

BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained language model that has achieved remarkable results in various NLP (Natural Language Processing) tasks [20]. To apply BERT to specific tasks such as sentiment analysis, a common approach used is fine-tuning techniques.

BERT (Bidirectional Encoder Representations from Transformers) without optimization refers to the application of pre-trained language models used directly in analysis tasks without any special adjustments[21]. This model, which has achieved significant results in various natural language processing (NLP) tasks, is used as the baseline in this study, allowing users to understand the basic performance of BERT under standard conditions, without a fine-tuning process that further adapts the model to specific contexts or data.

Nonetheless, the BERT model without optimization itself, lacks broader world knowledge or higher reasoning that can help in more complex tasks [20]. Consequently, a fine-tuning process is needed to prevent overfitting of the data, and to better understand the context of the initial data so as to have higher accuracy in sentiment analysis tasks [21], [22].

2.4.2. BERT with Fine-Tuning

BERT with Fine-Tuning is a technique, language model training that involves two stages. First, the BERT model is initialized with weights that have been pre-trained on a large dataset. This stage aims to utilize the general knowledge acquired by BERT during the initial training. Next, the training process continues by adjusting those weights using the specific data of the task to be accomplished, such as sentiment analysis. These adjustments allow BERT to more effectively map the nuances of language and relevant context from the given data [22].

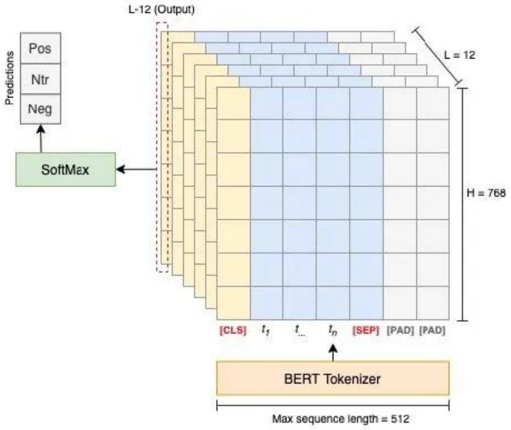


Figure 2. BERT Tokenizer

Fine-Tuning BERT involves the concept of Transfer Learning, where the general knowledge that BERT gains when trained on large data can be transferred and adapted for specific tasks with smaller amounts of data [23], [24]. Although BERT models themselves lack the wider world knowledge or higher reasoning that can help in more complex tasks, the Fine-Tuning process helps prevent overfitting on data, and allows BERT to better understand the context of the initial data and thus have higher accuracy [25], [26].

In this research, the IndoBERT-base-p1 model based on BERT-base architecture is used. Before the Fine-Tuning process, the dataset must be prepared by tokenizing the sentences using BertTokenizer, so that the input received by BERT matches the expected format. BertTokenizer tokenizes the text, input into a sequence of tokens with a maximum length of 512 tokens. During the fine-tuning process, the tokens are represented as embedding vectors through several embedding layers, namely token embedding, segment embedding, and position embedding. These embedding layers are used to incorporate contextual and positional information of the tokens.

The embedding vectors of the tokens are then fed into the BERT encoder stack which consists of 12 identical layers. Each encoder layer uses a self-attention mechanism to extract contextual relationships between tokens in the sentence. The output of the BERT encoder stack is a vector representation for each token. The output of interest is the [CLS] token which represents the entire sentence. This vector of [CLS] tokens is then used as input to the classifier, which generates a logit (rough probability prediction) of the sentence to be classified into sentiment classes (positive, negative, or neutral) [27]. Softmax is then used to convert the logit into probabilities with values between 0 and 1. To optimize BERT's performance in specific tasks, hyperparameters can be adjusted such as batch size, number of epochs, and learning rate. In this study, the hyperparameters used are encode max sequence length 512, batch size 32, workers 8, learning rate (AdamW) 5e-5, and epoch 5.

With Fine-Tuning techniques, BERT can utilize the general knowledge that has been previously acquired and customize it for specific tasks such as sentiment analysis, thus achieving optimal performance on the task.

2.5. Evaluation

This fifth stage aims to evaluate the model that has been implemented. Evaluation is done by conducting tests. In this research, the evaluation is carried out by dividing the dataset into three subsets, namely the training subset, validation subset, and test subset, then, the performance of BERT before optimization with Fine-Tuning, and BERT after optimization with Fine-Tuning on the dataset will be evaluated using confusion matrix, such as accuracy, precision, recall, and F1-score [28].

Table 1. Confusion matrix

True	Predicted sentiment		
sentiment	Positive	Neutral	Negative
Positive	TP	FNt	FN
Neutral	FP	TNt	FN
Negative	FP	FNt	TN

Precision is the calculation of the estimated proportion of positive cases formulated in (1):

$$Precision \ = \frac{TP}{TP+FP} \times 100\% \tag{1}$$

Recall is the calculation of the estimated proportion of correctly identified positive cases and is as shown in (2):

$$Recall \ = \frac{TP}{TP+FNt+FN} \times 100\% \tag{2}$$

F1-score is a calculation of precission and recall that balances the trade-off between these metrics, providing a summary of the model's ability to correctly identify positive cases while minimizing false identifications. It is formulated in equation (3):

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \times 100\% \tag{3}$$

Accuracy is a calculation of the proportion of the total number of correct predictions and is as shown in (4):

$$Accuracy = \frac{TP + TNt + TN}{TP + FP + TNt + FNt + TN + FN} \times 100\% \tag{4}$$

Where:	TP	: True Positive
	TNt	: True Neutral
	TN	: True Negative
	FP	: False Positive
	FNt	: False Neutral
	FN	: False Negative

2.6. Deployment

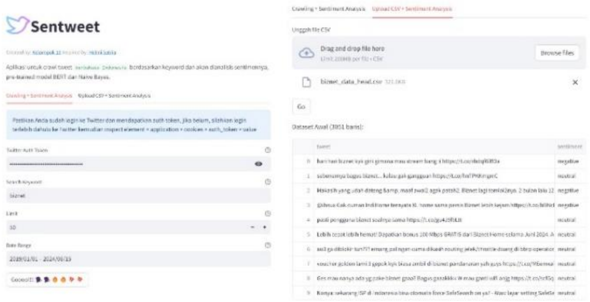
The sixth stage is an important implementation in this research. In this stage, the crawling, preprocessing, and the best model from the training and evaluation process will be implemented on the Streamlit web application. This application allows users to enter certain keywords or phrases and get sentiment predictions from the processed Twitter data, whether positive, negative, or neutral. With the Streamlit web application, the results of this research can be widely accessed and utilized by users or other parties who need sentiment analysis from Twitter data.

3. RESULT

This section is a discussion of the research that has been done. Starting from the data understanding, data preparation, modeling, evaluation, deployment.

3.1. Crawling Result

The crawling process is done to collect tweet data from Twitter using the keyword biznet. Crawling technique with npm tweet-harvest package allows automatic and real-time data collection. From the crawling results in Figure 3, a dataset of 3.051 tweet data was obtained which will be used for the training and evaluation process of the sentiment analysis mode.



Sentweet

Aplikasi untuk crawl tweet [illegible] berdasarkan keyword dan akan dianalisis sentimennya, pre-trained model BERT dan Naive Bayes.

Pastikan Anda sudah login ke Twitter dan mendapatkan auth token, jika belum, silahkan login terlebih dahulu ke Twitter kemudian inspect element > application > cookies > auth_token > value

Twitter Auth Token: ••••••••••••••••••••

Search Keyword: biznet

Limit: 10

Date Range: 2019/01/01 - 2024/06/15

Unggah file CSV

biznet_data_head.csv

Dataset Awal (3051 baris):

	tweet	sentiment
0	[illegible] https://t.co/[illegible]	negative
1	sebenarnya bagus biznet... kalau gak gangguan https://t.co/[illegible]	neutral
2	[illegible] 2 bulan lalu 12	negative
3	[illegible] biznet lebih kayam https://t.co/[illegible]	negative
4	pasti pengguna biznet [illegible] https://t.co/[illegible]	neutral
5	Lebih cepat lebih hemat! Dapatkan bonus 100 Mbps GRATIS dari Biznet Home selama Juni 2024. A	neutral
6	[illegible]	neutral
7	[illegible]	neutral
8	[illegible] https://t.co/[illegible]	neutral
9	[illegible] ISP di Indonesia bisa domain force SafeSearch on ya? - [illegible]	neutral

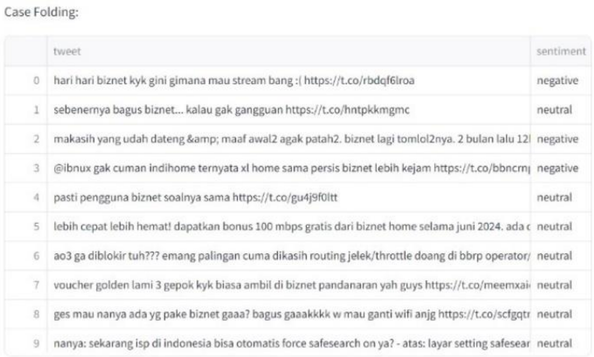
Crawling + Sentiment Analysis
Upload CSV + Sentiment Analysis
Drag and drop file here
Limit 200MB per file • CSV
Browse files
Go

Figure 3. Crawling Result

From the results of crawling 3.051 data using the keyword biznet, the distribution of sentiment in the dataset is obtained as follows: 660 positive sentiments (21.6%), 949 neutral sentiments (31.1%), and 1.442 negative sentiments (47.3%).

3.2. Case Folding Result

After obtaining the dataset from the crawling results, the first preprocessing stage performed is case folding. By performing case folding, the entire text is converted into lowercase letters as can be seen in Figure 4, thus eliminating the difference in the use of uppercase and lowercase letters. This helps improve data consistency and reduce redundancy in the tokenization and sentiment modeling process.



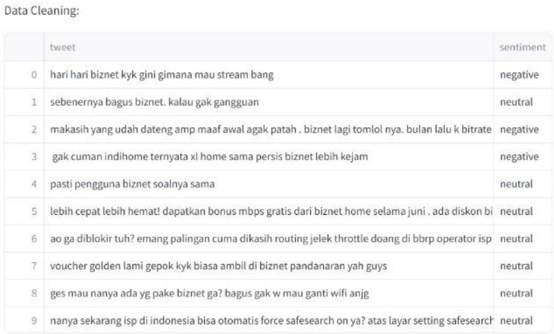
Case Folding:

	tweet	sentiment
0	hari hari biznet kyk gini gimana mau stream bang :(https://t.co/rbdqf6lroa	negative
1	sebenernya bagus biznet... kalau gak gangguan https://t.co/hntpkkmgmc	neutral
2	makasih yang udah dateng &amp; maaf awal2 agak patah2. biznet lagi tomlol2nya. 2 bulan lalu 12	negative
3	@ibnux gak cuman indihome ternyata xl home sama persis biznet lebih kejam https://t.co/bbncrn	negative
4	pasti pengguna biznet soalnya sama https://t.co/gu4j9f0ltt	neutral
5	lebih cepat lebih hemat! dapatkan bonus 100 mbps gratis dari biznet home selama juni 2024. ada	neutral
6	ao3 ga diblokir tuh??? emang palingan cuma dikasih routing jelek/throttle doang di bbrp operator/	neutral
7	voucher golden lami 3 gepok kyk biasa ambil di biznet pandanaran yah guys https://t.co/meemxai	neutral
8	ges mau nanya ada yg pake biznet gaaa? bagus gaaakkkk w mau ganti wifi anjg https://t.co/scfgqtr	neutral
9	nanya: sekarang isp di indonesia bisa otomatis force safesearch on ya? - atas: layar setting safesear	neutral

Figure 4. Case Folding Result

3.3. Data Cleaning Result

The data cleaning stage is the second preprocessing stage performed after the case folding stage. The data cleaning process that can be seen in Figure 5 aims to clean the noise so that the accuracy of the model in sentiment analysis can be improved. In addition, the removal of mentions, hashtags, and URLs can also help maximize the next tokenization process and text normalization.



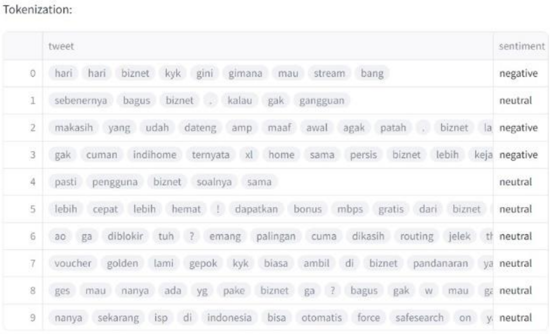
Data Cleaning:

	tweet	sentiment
0	hari hari biznet kyk gini gimana mau stream bang	negative
1	sebenernya bagus biznet. kalau gak gangguan	neutral
2	makasih yang udah dateng amp maaf awal agak patah . biznet lagi tomlol nya. bulan lalu k bitrate	negative
3	gak cuman indihome ternyata xl home sama persis biznet lebih kejam	negative
4	pasti pengguna biznet soalnya sama	neutral
5	lebih cepat lebih hemat! dapatkan bonus mbps gratis dari biznet home selama juni . ada diskon bi	neutral
6	ao ga diblokir tuh? emang palingan cuma dikasih routing jelek throttle doang di bbrp operator isp	neutral
7	voucher golden lami gepok kyk biasa ambil di biznet pandanaran yah guys	neutral
8	ges mau nanya ada yg pake biznet ga? bagus gak w mau ganti wifi anjg	neutral
9	nanya sekarang isp di indonesia bisa otomatis force safesearch on ya? atas layar setting safesearch	neutral

Figure 5. Data Cleaning Result

3.4. Tokenization Result

The next preprocessing stage is tokenization. In Figure 6 of the tokenization results, it can be seen that the text has been separated into tokens. This separation of text into tokens allows the model to more easily recognize language patterns.



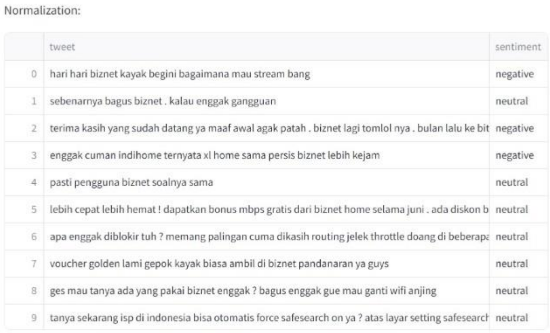
Tokenization:

	tweet	sentiment
0	hari hari biznet kyk gini gimana mau stream bang	negative
1	sebenernya bagus biznet . kalau gak gangguan	neutral
2	makasih yang udah dateng amp maaf awal agak patah . biznet la	negative
3	gak cuman indihome ternyata xl home sama persis biznet lebih keja	negative
4	pasti pengguna biznet soalnya sama	neutral
5	lebih cepat lebih hemat ! dapatkan bonus mbps gratis dari biznet	neutral
6	ao ga diblokir tuh ? emang palingan cuma dikasih routing jelek th	neutral
7	voucher golden lami gepok kyk biasa ambil di biznet pandanaran ya	neutral
8	ges mau nanya ada yg pake biznet ga ? bagus gak w mau ga	neutral
9	nanya sekarang isp di indonesia bisa otomatis force safesearch on y	neutral

Figure 6. Tokenization Result

3.5. Normalization Result

In Figure 7, we can see the result of the text normalization process. By performing normalization, the text becomes cleaner, structured, and uniform. This helps machine learning models to more easily recognize patterns and extract relevant features from the text, thus improving accuracy in tasks such as sentiment analysis.



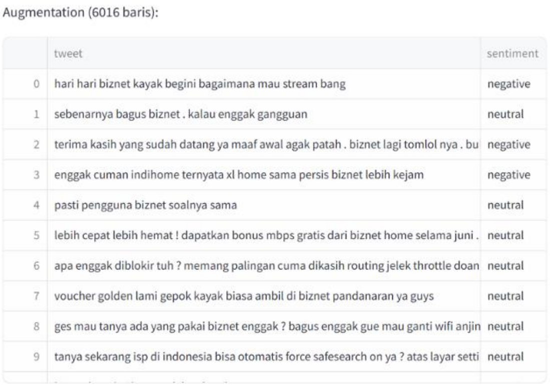
Normalization:

	tweet	sentiment
0	hari hari biznet kayak begini bagaimana mau stream bang	negative
1	sebenarnya bagus biznet . kalau enggak gangguan	neutral
2	terima kasih yang sudah datang ya maaf awal agak patah . biznet lagi tomlol nya . bulan lalu ke bit	negative
3	enggak cuman indihome ternyata xl home sama persis biznet lebih kejam	negative
4	pasti pengguna biznet soalnya sama	neutral
5	lebih cepat lebih hemat ! dapatkan bonus mbps gratis dari biznet home selama juni . ada diskon b	neutral
6	apa enggak diblokir tuh ? memang palingan cuma dikasih routing jelek throttle doang di beberapa	neutral
7	voucher golden lami gepok kayak biasa ambil di biznet pandanaran ya guys	neutral
8	ges mau tanya ada yang pakai biznet enggak ? bagus enggak gue mau ganti wifi anjing	neutral
9	tanya sekarang isp di indonesia bisa otomatis force safesearch on ya ? atas layar setting safesearch	neutral

Figure 7. Normalization Result

3.6. Data Augmentation Result

From the initial dataset of 3.051 data, after augmentation with this technique, the amount of data increased to 6.016 data as can be seen in Figure 8. The addition of 2.965 data or about 49.3% of the total data. The addition of this data aims to increase data diversity and help the model learn new patterns, so it is expected to improve sentiment classification performance.



Augmentation (6016 baris):

	tweet	sentiment
0	hari hari biznet kayak begini bagaimana mau stream bang	negative
1	sebenarnya bagus biznet . kalau enggak gangguan	neutral
2	terima kasih yang sudah datang ya maaf awal agak patah . biznet lagi tomlol nya . bu	negative
3	enggak cuman indihome ternyata xl home sama persis biznet lebih kejam	negative
4	pasti pengguna biznet soalnya sama	neutral
5	lebih cepat lebih hemat ! dapatkan bonus mbps gratis dari biznet home selama juni .	neutral
6	apa enggak diblokir tuh ? memang palingan cuma dikasih routing jelek throttle doan	neutral
7	voucher golden lami gepok kayak biasa ambil di biznet pandanaran ya guys	neutral
8	ges mau tanya ada yang pakai biznet enggak ? bagus enggak gue mau ganti wifi anjin	neutral
9	tanya sekarang isp di indonesia bisa otomatis force safesearch on ya ? atas layar setti	neutral

Figure 8. Data Augmentation Result

3.7. Split Dataset Result

From the dataset of 6.016 data, it was divided into three subsets, namely the train subset of 4.211 data (70%), the validation subset of 1.209 data (20%), and the test subset of 596 data (10%). The details of the split result are shown in table 2.

Table 2. Split Dataset Result

Train	Validation	Test
70%	20%	10%
4.211	1.209	596

In the train subset consisting of 4.211 data, the proportion of positive sentiment is 911 data (21.6%), negative 1.991 data (47.3%), and neutral 1.309 data (31.1%). While in the validation subset of 1.209 data, the proportion of positive sentiments is 262 data (21.7%), negative 571 data (47.2%), and neutral 376 data (31.1%). Finally, the test subset consisting of 596 data has a distribution of 129 data (21.6%) positive sentiment, 282 data (47.3%) negative, and 185 data (31%) neutral. In general, the pattern of sentiment distribution in the three subsets is quite balanced.

3.8. Result Analyze

After preprocessing the data, configuring the model is essential to achieve optimal performance. This research investigates two configurations: BERT without optimization and BERT with Fine-Tuning.

The results of each configuration will be analyzed using metrics such as accuracy, precision, recall and F1-score. By comparing performance metrics across configurations, this research aims to identify the model that achieves the most robust and accurate results in twitter sentiment analysis.

3.8.1. BERT without Optimization

Tests were conducted on the BERT model without optimization using example tweets outside the dataset, as follows:

a. Text: "*wifi biznet lambat sekali*", Label: neutral (41.313%)
b. Text: "*wifi biznet stabil atau tidak?*", Label: neutral (40.774%)
c. Text: "*wifi biznet cepat dan lancar*", Label: neutral (40.166%)

From the three example sentences, it can be seen that the BERT model without optimization tends to predict all sentences as neutral sentiment, even though the first sentence is actually negative and the third sentence is positive.

Table 3. BERT without Optimization Confusion Matrix

True	Predicted sentiment		
sentiment	Positive	Neutral	Negative
Positive	36	534	1
Neutral	25	347	4
Negative	25	236	1

Furthermore, in Table 3. BERT without Optimization Confusion Matrix, we can see the model performance on the validation data subset. From the confusion matrix, results are obtained as in Table 4. Classification Report on BERT without Optimization, with an accuracy of only 0.32 or 32%. The precision, recall, and F1-Score values for each sentiment class are also relatively low.

Table 4. Classification Report on BERT without Optimization

	Precision	Recall	F1-Score	Support
Negative	0.17	0.00	0.01	262
Neutral	0.31	0.92	0.46	376
Positive	0.42	0.06	0.11	571
Accuracy			0.32	1209
Macro Avg	0.30	0.33	0.19	1209
Weighted Avg	0.33	0.32	0.20	1209

These results show that the BERT model without optimization is less able to recognize sentiment patterns in tweet data well. This is because the BERT model used is still general and has not been adapted to the characteristics of Twitter sentiment data. Therefore, it is necessary to optimize with Fine-Tuning techniques to improve the performance of the model in performing sentiment analysis on Twitter data. Fine-Tuning will help the BERT model adjust its weight to the characteristics of Twitter sentiment data, so that the model can better recognize language and context patterns associated with sentiment expressions.

3.8.2. BERT with Fine-Tuning

Fine-Tuning Process in this stage, the BERT model pre-trained on a large dataset is weighted using a more specific Twitter sentiment dataset. This adjustment is done through several additional training epochs as can be seen in table 5 by further learning the language and context patterns associated with sentiment analysis on Twitter data.

Table 5. Epoch Process on BERT with Fine-Tuning

		Accuracy	Precision	Recall	F1-Score
Epoch 1	Train 1	0.76	0.75	0.73	0.74
	Validation 1	0.80	0.80	0.79	0.79
Epoch 2	Train 2	0.91	0.90	0.90	0.90
	Validation 2	0.86	0.87	0.83	0.85
Epoch 3	Train 3	0.96	0.96	0.96	0.96
	Validation 3	0.89	0.89	0.88	0.88
Epoch 4	Train 4	0.98	0.98	0.98	0.98
	Validation 4	0.89	0.88	0.88	0.88
Epoch 5	Train 5	0.98	0.98	0.98	0.98
	Validation 5	0.89	0.89	0.87	0.88

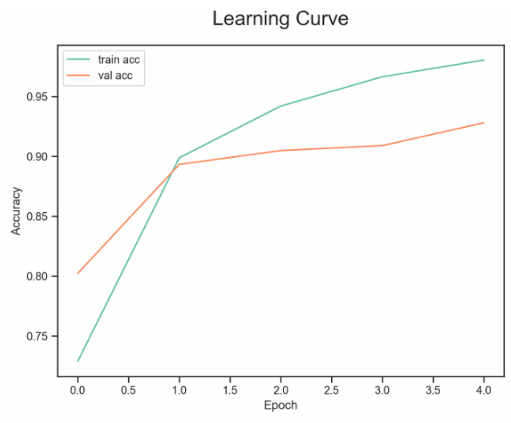


Figure 9. Learning Curve

Figure 8 shows the accuracy graph of the BERT model against training and validation data during the fine-tuning process for 5 epochs. At epoch 1, the accuracy is still low. However, the accuracy continues to increase significantly in subsequent epochs, both for train and validation data. This shows that the Fine-Tuning process successfully adjusted the model weights to better learn sentiment analysis patterns on Twitter data.

Tests were also conducted on the BERT model with Fine-Tuning using example tweets outside the dataset, as follows:

a. Text: "*wifi biznet lambat sekali*", Label: negative (99.872%)
b. Text: "*wifi biznet stabil atau tidak?*", Label: neutral (99.852%)
c. Text: "*wifi biznet cepat dan lancar*", Label: positive (98.274%)

The test results show that the BERT model with Fine-Tuning is able to classify sentiment well according to the context of the sentence. In the sentence "*wifi biznet lambat sekali*", the model can classify it as a negative sentiment (poor service). On "*wifi biznet cepat dan lancar*", the model classifies it as positive (good service). This success is because the Fine-Tuning process helps the model to efficiently learn sentiment-related language and context patterns, resulting in more accurate sentiment classification performance on Twitter data.

Table 6. BERT with Fine-Tuning Confusion Matrix Validation Set

True	Predicted sentiment		
sentiment	Positive	Neutral	Negative
Positive	535	30	9
Neutral	38	329	10
Negative	13	36	209

Table 7. Classification Report on BERT with Fine-Tuning Validation Set

	Precision	Recall	F1-Score	Support
Negative	0.91	0.93	0.92	574
Neutral	0.83	0.87	0.85	377
Positive	0.92	0.81	0.86	258
Accuracy			0.89	1209
Macro Avg	0.89	0.87	0.88	1209
Weighted Avg	0.89	0.89	0.89	1209

Results on Validation Data in Table 6. BERT with Fine-Tuning Confusion Matrix Validation Set, we can see the performance of the model on the validation data subset after going through the Fine-Tuning process. Based on Table 7. Classification Report on BERT with Fine-Tuning Validation Set, the model accuracy reaches 0.89 or 89%. The precision, recall, and F1-Score values for each sentiment class are also very good, ranging from 0.81 to 0.93.

Table 8. BERT with Fine-Tuning Confusion Matrix Test Set

True	Predicted sentiment		
sentiment	Positive	Neutral	Negative
Positive	272	8	3
Neutral	21	157	8
Negative	9	5	113

Table 9. Classification Report on BERT with Fine-Tuning Test Set

	Precision	Recall	F1-Score	Support
Negative	0.90	0.96	0.93	283
Neutral	0.92	0.84	0.88	186
Positive	0.91	0.89	0.90	127
Accuracy			0.91	596
Macro Avg	0.91	0.90	0.90	596
Weighted Avg	0.91	0.91	0.91	596

Results on the subsequent Test Data, in Table 8. BERT with Fine-Tuning Confusion Matrix Test Set and Table 9. Classification Report on BERT with Fine-Tuning Test Set, the model performance on a separate subset of test data is shown. There was an increase in accuracy compared to the validation data, to 0.91 or 91% from the previous 0.89 in the validation data. The precision, recall, and F1-Score values for each class are also still very good.

The above results show that by performing Fine-Tuning, the performance of the BERT model in performing sentiment analysis on Twitter data becomes much better than BERT without optimization. Fine-Tuning helps the model adjust its weights to the characteristics of Twitter sentiment data, so that it can better recognize language and context patterns associated with sentiment expressions. This can be seen from the significant increase in accuracy, precision, recall, and F1-Score after Fine-Tuning.

3.9. Deployment Result

At this stage, the developed system is deployed using the Streamlit Community Cloud platform, a free hosting service provided by Streamlit. Users can access and utilize the system through the following URL: https://soeara-sentweet.streamlit.app. The deployment process involves integrating the core components of the research, including the data crawling module, preprocessing techniques, and the optimized BERT model with Fine-Tuning.

4. DISCUSSIONS

In this study, we have conducted a comparative analysis between the BERT model without optimization and the BERT model with Fine-Tuning optimization for sentiment analysis on Twitter data using text mining techniques. From the results obtained, it can be seen that the BERT model with Fine-Tuning optimization shows much better performance compared to the BERT model without optimization. In the BERT model without optimization, the accuracy achieved is only 0.32 or 32% on validation data. The precision, recall, and F1-Score values for each sentiment class are also relatively low. This shows that the BERT model without optimization is less able to recognize sentiment patterns in tweet data well. This is because the BERT model used is still general and has not been adapted to the characteristics of Twitter sentiment data.

In contrast, after optimization with Fine-Tuning techniques, the performance of the BERT model improved significantly. On the validation data, the accuracy reached 0.89 or 89%, with precision, recall, and F1-Score values for each sentiment class in the range of 0.81 to 0.93. On split test data, the accuracy even increased to 0.91 or 91%, with other evaluation metrics also being excellent. This significant performance improvement indicates that the Fine-Tuning process successfully adapted the BERT model weights to the characteristics of Twitter sentiment data, so that the model can better recognize language and context patterns associated with sentiment expressions.

The results of this study are in line with several previous studies that also show the superiority of BERT models with Fine-Tuning optimization in sentiment analysis tasks. For example, research conducted by Muhammad Bilal used BERT with Fine-Tuning for online customer review classification and obtained a more generalized approach in predicting review usability compared to traditional machine learning methods. However, this study has its own advantages because it focuses on Twitter data in Indonesian and uses the IndoBERT model specifically trained for Indonesian. In addition, this research also applies several preprocessing techniques such as data augmentation with back translation to improve the performance of the model. Overall, the results of this study make an important contribution to the development of sentiment analysis methods on Twitter data, especially for the Indonesian language. With the advantages of the BERT model optimized through Fine-Tuning, sentiment analysis can be performed more accurately and reliably, thus providing valuable insights for various parties that utilize Twitter data.

5. CONCLUSION

This research has conducted a comparative analysis between the BERT model without optimization and the BERT model with Fine-Tuning optimization for sentiment analysis on Twitter data using text mining techniques. The analysis results show that the BERT model with Fine-Tuning optimization is superior in performance compared to the BERT model without optimization. On validation data, the accuracy, precision, recall, and F1-score of the BERT model without optimization are relatively low for each sentiment class. But after optimization with Fine-Tuning, there is a significant improvement in these metrics. In the test data, the improvement in accuracy and other evaluation metrics is higher using the Fine-Tuning optimized BERT model. This indicates that the Fine-Tuning process successfully adapted the BERT model to the characteristics of Twitter sentiment data, resulting in better recognition of language and context patterns related to sentiment expressions. Fine-Tuning facilitates the transfer of BERT model learning from previously acquired general knowledge to the specific task of Twitter sentiment analysis. Through additional training stages, the model is able to understand the nuances and context associated with sentiment expressions, resulting in more accurate sentiment classification. Overall, this research confirms the superiority of the Fine-Tuning optimized BERT model for sentiment analysis on Twitter data, particularly Indonesian language, with the ability to provide valuable insights for Twitter data users.

CONFLICT OF INTEREST

The authors declares that there is no conflict of interest between the authors or with research object in this paper.

REFERENCES

[1] Y. Qi and Z. Shabrina, "Sentiment analysis using Twitter data: a comparative application of lexicon- and machine-learning-based approach," *Soc. Netw. Anal. Min.*, vol. 13, no. 1, pp. 1–14, 2023, doi: 10.1007/s13278-023-01030-x.

[2] A. Ghorbanali and M. K. Sohrabi, "A comprehensive survey on deep learning-based approaches for multimodal sentiment analysis," *Artif. Intell. Rev.*, vol. 56, no. 1, pp. 1479–1512, 2023, doi: 10.1007/s10462-023-10555-8.

[3] L. Zhang, H. Fan, C. Peng, G. Rao, and Q. Cong, "Sentiment analysis methods for hpv vaccines related tweets based on transfer learning," *Healthc.*, vol. 8, no. 3, pp. 1–18, 2020, doi: 10.3390/healthcare8030307.

[4] C. Sun, Z. Yang, L. Wang, Y. Zhang, H. Lin, and J. Wang, "Deep learning with language models improves named entity recognition for PharmaCoNER," *BMC Bioinformatics*, vol. 22, no. 1, pp. 1–16, 2021, doi: 10.1186/s12859-021-04260-y.

[5] F. B. Oliveira, A. Haque, D. Mougouei, S. Evans, J. S. Sichman, and M. P. Singh, "Investigating the Emotional Response to COVID-19 News on Twitter: A Topic Modelling and Emotion Classification Approach," *IEEE Access*, vol. 10, no. 1, pp. 16883–16897, 2022, doi:

10.1109/ACCESS.2022.3150329.

[6] A. Pathak, S. Kumar, P. P. Roy, and B. G. Kim, "Aspect-based sentiment analysis in hindi language by ensembling pre-trained mbert models," *Electron.*, vol. 10, no. 21, pp. 1–15, 2021, doi: 10.3390/electronics10212641.

[7] V. Gupta, J. Patel, S. Shubbar, and K. Ghazinour, "COVBERT: Enhancing Sentiment Analysis Accuracy in COVID-19 X Data through Customized BERT," *CS IT Conf. Proc.*, vol. 14, no. 2, pp. 171–182, 2024, doi: 10.5121/csit.2024.140212.

[8] M. Bilal and A. A. Almazroi, "Effectiveness of Fine-tuned BERT Model in Classification of Helpful and Unhelpful Online Customer Reviews," *Electron. Commer. Res.*, vol. 23, no. 4, pp. 2737–2757, 2023, doi: 10.1007/s10660-022-09560-w.

[9] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," *Procedia Comput. Sci.*, vol. 181, no. 1, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.

[10] B. Richardson and A. Wicaksana, "Comparison of Indobert-Lite and Roberta in Text Mining for Indonesian Language Question Answering Application," *Int. J. Innov. Comput. Inf. Control*, vol. 18, no. 6, pp. 1719–1734, 2022, doi: 10.24507/ijicic.18.06.1719.

[11] M. A. Khder, "Web scraping or web crawling: State of art, techniques, approaches and application," *Int. J. Adv. Soft Comput. its Appl.*, vol. 13, no. 3, pp. 144–168, 2021, doi: 10.15849/ijasca.211128.11.

[12] A. Bello, S. C. Ng, and M. F. Leung, "A BERT Framework to Sentiment Analysis of Tweets," *Sensors*, vol. 23, no. 1, pp. 1–14, 2023, doi: 10.3390/s23010506.

[13] M. R. Akbar, I. Slamet, and S. S. Handajani, "Sentiment analysis using tweets data from twitter of indonesian's capital city changes using classification method support vector machine," *AIP Conf. Proc.*, vol. 2296, no. 1, pp. 1–8, 2020, doi: 10.1063/5.0030357.

[14] P. Chauhan, N. Sharma, and G. Sikka, "On the importance of pre-processing in small-scale analyses of twitter: a case study of the 2019 Indian general election," *Multimed. Tools Appl.*, vol. 83, no. 7, pp. 19219–19258, 2024, doi: 10.1007/s11042-023-16158-3.

[15] D. Cho, H. Lee, and S. Kang, "An empirical study of korean sentence representation with various tokenizations," *Electron.*, vol. 10, no. 7, pp. 1–12, 2021, doi: 10.3390/electronics10070845.

[16] K. Mehmood, D. Essam, K. Shafi, and M. K. Malik, "An unsupervised lexical normalization for Roman Hindi and Urdu sentiment analysis," *Inf. Process. Manag.*, vol. 57, no. 6, p. 102368, 2020, doi: https://doi.org/10.1016/j.ipm.2020.102368.

[17] Y. Sujana and H. Y. Kao, "LiDA: Language-Independent Data Augmentation for Text Classification," *IEEE Access*, vol. 11, no. 1, pp. 10894–10901, 2023, doi: 10.1109/ACCESS.2023.3234019.

[18] L. Alzubaidi *et al.*, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *J. Big Data*, vol. 8, no. 1, pp. 1–74, 2021, doi: 10.1186/s40537-021-00444-8.

[19] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, "Machine learning algorithm validation with a limited sample size," *PLoS One*, vol. 14, no. 11, pp. 1–20, 2019, doi: 10.1371/journal.pone.0224365.

[20] S. Abdel-Salam and A. Rafea, "Performance Study on Extractive Text Summarization Using BERT Models," *Inf.*, vol. 13, no. 2, pp. 1–10, 2022, doi: 10.3390/info13020067.

[21] X. Wang, F. Chao, and G. Yu, "Evaluating Rumor Debunking Effectiveness During the COVID-19 Pandemic Crisis: Utilizing User Stance in Comments on Sina Weibo," *Front. Public Heal.*, vol. 9, no. 1, pp. 1–18, 2021, doi: 10.3389/fpubh.2021.770111.

[22] Q. Meng *et al.*, "Electric Power Audit Text Classification With Multi-Grained Pre-Trained

Language Model," *IEEE Access*, vol. 11, no. 1, pp. 13510–13518, 2023, doi: 10.1109/ACCESS.2023.3240162.

[23] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proc. naacL-HLT*, vol. 1, no. 1, pp. 4171–4186, 2019, [Online]. Available: https://aclanthology.org/N19-1423.pdf

[24] D. Li, Y. Xiong, B. Hu, B. Tang, W. Peng, and Q. Chen, "Drug knowledge discovery via multi-task learning and pre-trained models," *BMC Med. Inform. Decis. Mak.*, vol. 21, no. 1, pp. 1–8, 2021, doi: 10.1186/s12911-021-01614-7.

[25] S. Lamsiyah, A. El Mahdaouy, S. E. A. Ouatik, and B. Espinasse, "Unsupervised extractive multi-document summarization method based on transfer learning from BERT multi-task fine-tuning," *J. Inf. Sci.*, vol. 49, no. 1, pp. 164–182, 2023, doi: 10.1177/0165551521990616.

[26] B. Ko and H. J. Choi, "Twice fine-tuning deep neural networks for paraphrase identification," *Electron. Lett.*, vol. 56, no. 9, pp. 449–450, 2020, doi: 10.1049/el.2019.4183.

[27] C. Liu, W. Zhu, X. Zhang, and Q. Zhai, "Sentence part-enhanced BERT with respect to downstream tasks," *Complex Intell. Syst.*, vol. 9, no. 1, pp. 463–474, 2023, doi: 10.1007/s40747-022-00819-1.

[28] M. P. Geetha and D. Karthika Renuka, "Improving the performance of aspect based sentiment analysis using fine-tuned Bert Base Uncased model," *Int. J. Intell. Networks*, vol. 2, no. 1, pp. 64–69, 2021, doi: 10.1016/j.ijin.2021.06.005.