VOL. 11. NO. 1 AUGUST 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v11i1.6392

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

ENHANCING SENTIMENT ANALYSIS ACCURACY WITH BERT AND SILHOUETTE METHOD OPTIMIZATION

Kelvin[1]; Frans Mikael Sinaga[1]; Wulan Sri Lestari[1]; Sunaryo Winardi[1]; Khairul Hawani Rambe[2]; Ronsen Purba[1]

Informatics[1]
Universitas Mikroskil, Medan, Indonesia[1]
https://mikroskil.ac.id/[1,]
kelvin.chen@mikroskil.ac.id; frans.sinaga@mikroskil.ac.id*; sunaryo.winardi@mikroskil.ac.id; khairulhawani@pnb.ac.id; ronsen@mikroskil.ac.id

Digital Business[2]
Politeknik Negeri Bali, Bali, Indonesia[2]
https://www.pnb.ac.id/[2]
wulan.lestari@mikroskil.ac.id

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *This research is based on the emergence of ChatGPT technology, which has significant implications in various fields. This research aims to design a model that improves sentiment analysis classification accuracy. The methods applied include the use of the Silhouette Coefficient to determine the best cluster parameters before performing data grouping with the Self-Organizing Map (SOM) method. Additionally, the Bidirectional Encoder Representations from Transformers (BERT) model is utilized to perform precise and convergent sentiment classification. The research methodology encompasses several phases, including data preprocessing through natural language processing techniques. Textual data is converted into vector representations, which are then processed using the Silhouette Coefficient to identify the optimal cluster parameters. These parameters are subsequently applied in the Self-Organizing Map method to cluster data, while the Bidirectional Encoder Representations from Transformers model determines public sentiment, categorized as positive, negative, or neutral. The findings of this study indicate that the best cluster parameter is 9, using a batch size of 64 and a maximum sequence length of 128. The highest accuracy achieved using the confusion matrix is 92.06%. Further tests with varying parameters confirm that the Silhouette Coefficient method significantly enhances the convergence and accuracy of classification outcomes. The conclusion of this research is that integrating the Silhouette Coefficient and Bidirectional Encoder Representations from Transformers is effective in optimizing sentiment analysis on large datasets, achieving both accurate and reliable results.*

Keywords: *BERT, big data, sentiment analysis, silhouette coefficient, SOM*

Intisari— *Penelitian ini didasarkan pada kemunculan teknologi ChatGPT, yang memiliki dampak signifikan di berbagai bidang. Penelitian ini bertujuan untuk merancang sebuah model yang meningkatkan akurasi klasifikasi analisis sentimen. Metode yang diterapkan mencakup penggunaan Silhouette Coefficient untuk menentukan parameter klaster terbaik sebelum melakukan pengelompokan data dengan metode Self-Organizing Map (SOM). Selain itu, model Bidirectional Encoder Representations from Transformers digunakan untuk mencapai klasifikasi sentimen yang akurat dan konvergen. Metode penelitian terdiri dari beberapa tahapan, termasuk praproses data menggunakan teknik natural language processing (NLP). Data tekstual diubah menjadi representasi vektor, yang kemudian diproses menggunakan Silhouette Coefficient untuk mengidentifikasi parameter klaster optimal. Parameter ini kemudian diterapkan dalam metode Self-Organizing Map untuk mengelompokkan data, sementara model Bidirectional Encoder Representations from*

VOL. 11. NO. 1 AUGUST 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i1.6392

Transformers menentukan sentimen publik yang dikategorikan sebagai positif, negatif, atau netral. Hasil penelitian menunjukkan bahwa parameter klaster terbaik adalah 9, dengan ukuran batch sebesar 64 dan panjang urutan maksimum 128. Akurasi tertinggi yang dicapai menggunakan matriks kebingungan (confusion matrix) adalah 92,06%. Pengujian lebih lanjut dengan parameter yang bervariasi mengonfirmasi bahwa metode Silhouette Coefficient secara signifikan meningkatkan konvergensi dan akurasi hasil klasifikasi. Kesimpulan dari penelitian ini adalah bahwa integrasi antara Silhouette Coefficient dan Bidirectional Encoder Representations from Transformers efektif dalam mengoptimalkan analisis sentimen pada dataset besar, menghasilkan hasil yang akurat dan andal.

***Kata Kunci**: BERT, data besar, analisis sentimen, koefisien siluet, SOM.*

INTRODUCTION

ChatGPT technology brings substantial advantages, it also presents potential for misuse, including copyright infringement, unlicensed utilization of journalistic content, and challenges to ethical accountability. Within the educational sector, its use raises concerns about the erosion of academic integrity [1] and the decline of critical and higher-order cognitive skills [2], and hinders student progress, potentially affecting education quality [3]. Sentiment analysis has become a crucial tool for understanding consumer opinions and emotions, particularly in social media and online reviews. The integration of advanced models such as Bidirectional Encoder Representations from Transformers (BERT) has significantly enhanced sentiment classification accuracy ([4],[5].

BERT has proven effective for sentiment analysis without extensive preprocessing [6] and shows improved performance with independent preprocessing methods [7]. Knowledge-enriched BERT models deliver detailed results, overcoming limitations in training data scope [6], and adapt well to bilingual and multilingual texts [8]. However, optimizing BERT's performance remains a challenge, and techniques such as the Silhouette method can be employed to enhance clustering and classification accuracy [9], [10],[11], [12]. Comparative analyses emphasize the importance of selecting appropriate methodologies to achieve high performance in sentiment classification tasks [13], [14].

This study highlights key issues in previous research: limited preprocessing led to the loss of critical text information, including meanings and phrases, and lacked effective vectorization [15]. SOM clustering struggled with optimal clusters due to random and unrepresentative centroid selection and inaccurate cluster parameter determination [16]. Additionally, limited context understanding and inadequate handling of complex text structures resulted in non-convergent and inaccurate sentiment analysis [16], [17]. Methods like K-Means and KNN are suboptimal for high-dimensional and nonlinear data, respectively [18], [19]. SOM struggles with optimal cluster determination and overfitting risks, leading to inconsistent results [16].

LSTM and FRCNN also face challenges with long, unstructured texts and sentiment variation [16], [17]. Previous research has demonstrated that integrating BERT with optimization techniques leads to improved sentiment analysis outcomes, as evidenced by studies highlighting the effectiveness of hybrid models in financial news sentiment analysis [20] and deep learning approaches in sentiment analysis [21]. Furthermore, feature extraction methods such as TF-IDF and GloVe play a crucial role in enhancing classification robustness [22]. SOM clustering struggles with optimal cluster determination and overfitting risks, leading to inconsistent results [16].

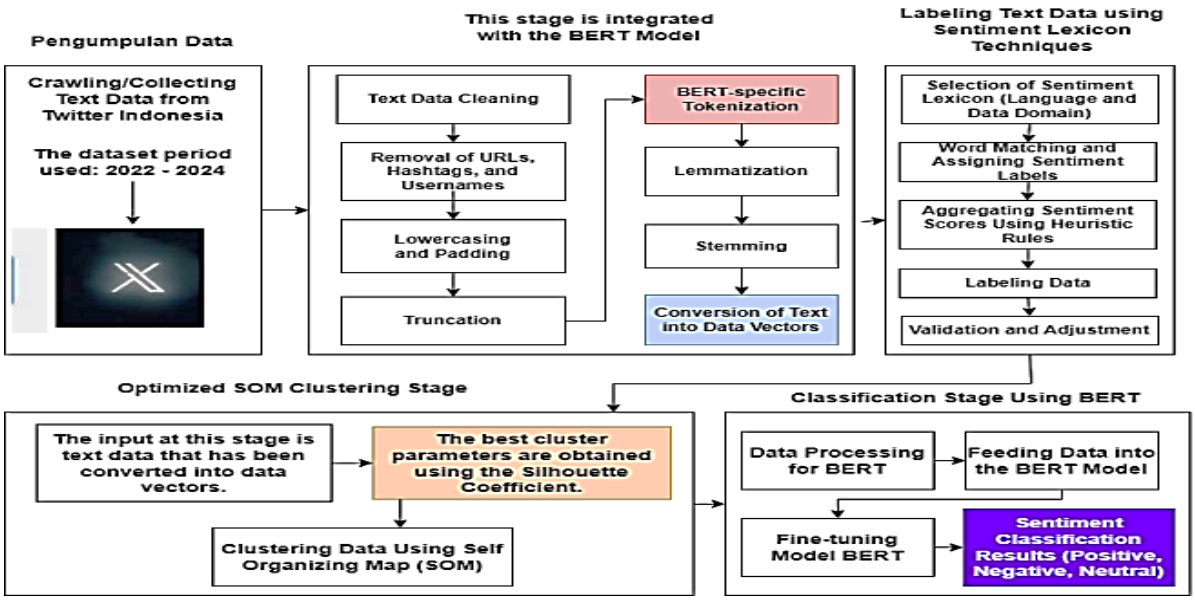
To address these gaps, advanced techniques like BERT can improve clustering and classification accuracy. This study emphasizes optimizing SOM clustering, mitigating overfitting risks, and leveraging BERT's contextual capabilities to enhance sentiment analysis outcomes [3], [7], [23]. Integrating these approaches provides a robust methodology for analyzing public sentiment towards ChatGPT and understanding its societal and educational impacts [24].

MATERIALS AND METHODS

The workflow illustrated in Figure 1 comprises several key phases: data acquisition, preprocessing, and the annotation of textual data using a sentiment lexicon-based approach. The Silhouette Coefficient identifies optimal cluster parameters, which SOM uses to group vector data based on similarity. Finally, the BERT model is applied for sentiment analysis classification.

VOL. 11. NO. 1 AUGUST 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v11i1.6392

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)



Source : (Research Results,2025)

Figure 1. The steps in problem solving approach

A. Data Preparation

Data preparation involves crawling text data from Indonesian Twitter users by employing relevant keywords aligned with the research objectives [16].

Table 1. List of dataset keywords used (in the Indonesian language)

No	Keyword Crowling Dataset	Total number of datasets
1	ChatGpt	
2	OpenAI	November 2022 – 2023
3	Pendidikan	5000 dataset
4	Penelitian	
5	ChatGpt Hak Cipta	
6	Plagiarisme	November 2022 – 2024
7	ChatGpt Disclaimer	7000 dataset
8	ChatGpt Berita Tanpa Izin	

Source : (Research Results,2025)

Table 1 lists the keywords used to collect text data from Indonesian Twitter users. Additional datasets were incorporated with new keywords, expanding on previous research.

B. Pre-processing Data

In the preprocessing stage, a novel approach with BERT prepares data. Sentiment Lexicon labels data using heuristic sentiment aggregation, while BERT encodes textual features into contextualized vector representations.

Applying Sentiment Lexicon: This step is implemented to generate data labels by the combination of sentiment indicators using heuristic rules described in Equation 1.

$$S_{Positive} \sum_{i \in t}^{n} positive\ score_i \quad (1)$$

$$S_{neutral} \sum_{i \in t}^{n} neutral\ score_i \quad (2)$$

$$S_{negative} \sum_{i \in t}^{n} negative\ score_i \quad (3)$$

Table 2. List of dataset keywords used [25] (in the Indonesian language)

URL	Date	Tweet	ID	Replies	Retweets	Likes	Quotes	Conv. ID	username	label
https://x.com/InfoKomputer/status/1612723925166743552	Tue Jan 10 08:12:01 +0000 2023	Cegah siswa nyontek begini cara institusi pendidikan siasati ChatGPT.	5935 7428	0	0	0	0	1,61272E+18	InfoKomputer	Netral
https://x.com/naimalkalantani/status/165	Thu May 11 03:17:	Jemput baca tulisan terbaru saya mengenai ChatGPT. Impak paling besar yang saya dapat	71 68 42	0	3	8	0	1,6565E+18	naimalkalantani	Positif

VOL. 11. NO. 1 AUGUST 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i1.6392

URL	Date	Tweet	ID	Replies	Retweets	Likes	Quotes	Conv. ID	username	label
64987959 14620932	43 +0000 2023	fikirkan sekarang ialah ChatGPT ini akan memberi kesan dalam pendidikan.	45 8							
https://x.com/AlphaARachman/status/1610876241258901512	Thu Jan 05 05:49:59 +0000 2023	Dinas Pendidikan New York City sudah mem-blok ChatGPT @ChatGPTUser @OpenAI di gadget dan jaringan internet sekolah..! □ Tapi kelihatannya akan sulit.	11 84 69 15 8	11	12	132	14	1,61088 E+18	AlphaARachman	Negatif

Source : (Sinaga,2020)

C. K Value using Silhouette Coefficient

Before clustering, the optimal cluster parameters are obtained using the Silhouette Coefficient technique [25], [26].

Input: vektordata = datasets.vektordata(), X = datateks.vektordata [:,2:]

Output: d,k

d = [];

For each k = 1,k in the range (1,9) perform the following calculation :

$$\mathrm{d} = \sum_{t=1}^{k} \sum d\ i\ s\ t(x_{:}, c_i^2) \quad (4)$$

Return d , k ;

D. Cluster data using SOM

SOM will cluster vector data into several clusters based on data similarity [16], [27].

1. Intialization
2. *Best Matching Unit* (*BMU*)

$$d_{ij}\sqrt{\sum_{k=1}^{n}(X_{ki} - X_{kj})^2} \quad (5)$$

Where:

d_{ij} = The distance or separation instances, labeled as i and j

X_{ki} = Two distinct instances, labeled as i and j

3. *Weight Update Formula*

$$\sigma\ (\mathrm{t}) = \sigma_0 \exp\left(-\frac{\mathrm{t}}{\mathrm{n}}.\mathrm{t}\right) = 1, 2, 3 \ldots \quad (6)$$

Where:

$\sigma\ (t)$ = radius at time t_0

t = current time step = 1. 2. 3

π = current time

The update formula for neurons is given by:

$$\mathrm{Wi'} = \mathrm{W2} + \sigma\ (\mathrm{t})\ \mathrm{Wi}\ (\mathrm{Xi} - \mathrm{Wi}) \quad (7)$$

Where:

Wi = weight

Xi = neuron

4. Iteration: The process is iteratively repeated until distinct clusters are formed.

E. Classification using BERT

BERT, a sophisticated model based on the transformer architecture, has been employed for analyzing sentiment in text. Pretrained on large datasets, it is fine-tuned on labeled data to identify sentiment patterns. Fine-tuning optimizes parameters, enabling accurate predictions through semantic and contextual understanding. This approach is effective across multiple languages and domains. The stages of the BERT model include:

1. BERT Data Processing: Every token must undergo tokenization and be converted into vector representations through embedding methods.

$$V_t = Trans\ (W_t X_a + b_t) \quad (8)$$

2. Inputting Data into BERT: The model analyzes textual input and produces contextualized vector embeddings for each token.

$$\mathrm{V_b} = \mathrm{Bert}(\mathrm{W_b X_b} + \mathrm{b_b}) \quad (9)$$

3. Through self-attention, the model captures improved contextual embeddings for each word based on its relationship with other words in the sentence.

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (10)$$

Where Q, K, and V refer to the query, key, and value vectors computed from the input embeddings.

4. Feed-Forward layers process the outputs of the Self-Attention stage to enrich token-level representations, enabling the model to manage non-linear aspects of language more effectively and support broader generalization.

$$FFN(H) = ReLU(HW_1 + b_1)W_2 + b_2 \quad (11)$$

Where H s the context vector produced by self-attention, whereas W1, W2, b1, dan b2 correspond to the parameters of the feed-forward layers.

5. Fine-Tuning the BERT Model: Enhancing sentiment classification performance by retraining the model on domain-specific data within BERT's contextual framework.
6. Sentiment Classification: BERT processes the input text and generates sentiment predictions by leveraging its contextualized vector representations.

F. Data Testing

In this study, the performance of the sentiment analysis model is rigorously evaluated using a confusion matrix, which serves as a critical tool for assessing the accuracy of the classification outcomes. The confusion matrix comprises four key components: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Specifically, True Negatives (TN) represent the instances of negative data that have been correctly classified, while False Positives (FP) indicate the negative instances that have been incorrectly classified as positive. Conversely, True Positives (TP) denote the instances that have been accurately identified as positive, and False Negatives (FN) reflect the positive instances that have been misclassified as negative. To quantify the model's accuracy, we employ a testing methodology that incorporates the aforementioned components. The performance of the classification method is evaluated using accuracy as the metric, which is mathematically represented by the equation:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (12)$$

This equation allows for a thorough assessment of the model's overall performance, ensuring that the results are both reliable and valid in the context of the main problems being studied. By analyzing these metrics, we can gain insights into the model's strengths and weaknesses, ultimately guiding further improvements in the sentiment analysis process.

RESULTS AND DISCUSSION

The outcomes of this study encompass data acquisition, text preprocessing, sentiment annotation using a lexicon-based approach, optimal cluster parameter identification via the Silhouette Coefficient, vector clustering using SOM based on data similarity, and sentiment classification with the BERT model.

A. The Results Of Crawling The Dataset

The extracted dataset from Indonesian Twitter spans January 2022 to June 2024, containing 7,000 entries collected using various keywords, as shown in Table 3

Table 3. The Crawled Dataset Results (in the Indonesian language)

URL	Date	Tweet	ID	Replies	Retweet	Likes	Quotes
https://x.com/ndorokakung/status/1674551525799088128	Thu Jun 29 22:52:50 +0000 2023	Kemunculan ChatGPT membuat sebagian guru mulai merasa galau. Kehadiran kecerdasan buatan [AI] seperti ChatGPT di satu sisi membantu, namun di sisi lain menimbulkan kekhawatiran terhadap eksistensi guru. Lantas, bagaimana sebenarnya peran AI dalam mendukung proses pendidikan dan pembelajaran? -- sebuah utas -- https://t.co/FJ3kz6ckWr	18598424	13	109	375	3
https://x.com/lia_liliy/status/1669140769398210560	Thu Jun 15 00:32:25 +0000 2023	Berawal dari sebuah organisasi nirlaba yang mengembangkan aplikasi untuk pendidikan, kini #ChatGPT dan #ArtificialIntelligence telah menyebar ke seluruh dunia dan digunakan dalam berbagai aspek kehidupan. #samaltman https://t.co/GyilKJHNMx	24141250	0	0	0	0
https://x.com/radyalabs/status/1805902516108083500	Wed Jun 26 09:54:46 +0000 2024	OpenAI baru saja meluncurkan ChatGPT Edu sebuah inovasi terbaru dalam dunia pendidikan. ChatGPT Edu dirancang untuk membantu proses belajar mengajar dengan menggunakan teknologi AI. Info lengkapnya cek gambar berikut #radyadigital #OpenAI #ChatGPT https://t.co/oeV511bhf4	2,25E+08	0	0	0	0

Source : (Research Results,2025)

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 1 AUGUST 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i1.6392

B. Dataset Labeling Results

Sentiment analysis of the Twitter dataset employed lexicon-based methods to classify sentiments into positive, negative, and neutral categories, as summarized in Table 4.

Table 4. Lexicon-Based Dataset Labeling Results (in the Indonesian language)

Lexicon Result	Tweet	Date	Username	Sentimen Label	
negative	Kemunculan ChatGPT membuat sebagian guru mulai merasa galau. Kehadiran kecerdasan buatan [AI] seperti ChatGPT di satu sisi membantu, namun di sisi lain menimbulkan kekhawatiran terhadap eksistensi guru. Lantas, bagaimana sebenarnya peran AI dalam mendukung proses pendidikan dan pembelajaran? -- sebuah utas -- https://t.co/FJ3kz6ckWr	Thu Jun 29 22:52:50 +0000 2023	ndorokakung	Negatif	Negatif	Netral
positive	Berawal dari sebuah organisasi nirlaba yang mengembangkan aplikasi untuk pendidikan, kini #ChatGPT dan #ArtificialIntelligence telah menyebar ke seluruh dunia dan digunakan dalam berbagai aspek kehidupan. #samaltman https://t.co/GyilKJHNMx	Thu Jun 15 00:32:25 +0000 2023	lia_liliy	Positif	Positif	Positif
neutral	KPT dilaporkan sedang menyediakan garis panduan penggunaan ChatGPT dalam sistem pendidikan negara. ChatGPT merupakan perkhidmatan AI yang boleh melakukan tugasan seperti menulis artikel kod program dan etc. Korang rasa perlu ke ChatGPT untuk sistem pembelajaran di Malaysia? https://t.co/wq6ZMvJqfy	Sat Mar 18 12:47:26 +0000 2023	tech_lagi	Netral	Netral	Netral

Source : (Research Results,2025)

C. Data Preprocessing Results

Raw textual data is preprocessed through a series of steps, including punctuation elimination, case normalization, correction of spelling variations, noise filtering, and tokenization.

1. Data Cleaning results: This step eliminates extraneous elements, including punctuation marks and special symbols. The resulting cleaned data is presented in Table 5.

Table 5. Results of Data Cleaning (in the Indonesian language)

Results of Data Cleaning	
Before text cleaning	After text cleaning
Kemunculan ChatGPT membuat sebagian guru mulai merasa galau. Kehadiran kecerdasan buatan [AI] seperti ChatGPT di satu sisi membantu, namun di sisi lain menimbulkan kekhawatiran terhadap eksistensi guru. Lantas, bagaimana sebenarnya peran AI dalam mendukung proses pendidikan dan pembelajaran? -- sebuah utas -- https://t.co/FJ3kz6ckWr	kemunculan chatgpt membuat sebagian guru mulai merasa galau. kehadiran kecerdasan buatan [ai] seperti chatgpt di satu sisi membantu, namun di sisi lain menimbulkan kekhawatiran terhadap eksistensi guru. lantas, bagaimana sebenarnya peran ai dalam mendukung proses pendidikan dan pembelajaran? sebuah utas https t.co fj kz ckwr

Source : (Research Results,2025)

2. Removal of URLs results: This stage eliminates web links from the textual data, as they generally do not contribute meaningful information for text analysis. The results of this process are presented in Table 6.

Table 6. URL Removal Results (in the Indonesian language)

URL Removal Results	
Before text removal url	After text removal url
Silakan baca tulisan terbaru saya tentang ChatGPT. Saat ini, dampak paling besar yang terpikirkan oleh saya adalah bagaimana ChatGPT akan memengaruhi dunia pendidikan. https://t.co/V0cuZIArfO https://t.co/vSYNFeEPo7	silakan baca tulisan terbaru saya tentang chatgpt. saat ini, dampak paling besar yang terpikirkan oleh saya adalah bagaimana chatgpt akan memengaruhi dunia pendidikan. https t.co v cuziarfo https t.co vsynfeepo

Source : (Research Results,2025)

3. Hashtag Removal Results: This step involves eliminating hashtags (i.e., words prefixed with "#") from the text, as they typically do not contribute meaningful information in most text analysis tasks. The outcomes of this process are shown in Table 7.

Table 7. Hashtag Removal Results (in the Indonesian language)

Before Hashtag Removal	After Hashtag Removal
Berawal dari sebuah organisasi nirlaba yang mengembangkan aplikasi untuk pendidikan, kini #ChatGPT dan #ArtificialIntelligence telah menyebar ke seluruh dunia dan digunakan dalam berbagai aspek kehidupan. #samaltman https://t.co/GyilKJHNMx	berawal dari sebuah organisasi nirlaba yang mengembangkan aplikasi untuk pendidikan, kini chatgpt dan artificialintelligence telah menyebar ke seluruh dunia dan digunakan dalam berbagai aspek kehidupan. samaltman https t.co gyilkjhnmx

Source : (Research Results,2025)

4. Results of Username Removal: User mentions (@) are excluded from the text, as they generally hold little relevance for analytical purposes. The processed results are displayed in Table 8.

Table 8. Removal Username Results. (in the Indonesian language)

Before removal username results	After removal userma,e results
ChatGPT dalam Dunia Pendidikan: Manfaat dan Cara Penggunaannya https://t.co/hFGlTUCtp3 via @wikismartid #ChatGPT #chatgpt4	chatgpt dalam dunia pendidikan: manfaat dan cara penggunaannya via wikismartid chatgpt chatgpt

Source : (Research Results, 2025)

5. Results of Lowercasing: All text characters are transformed to lowercase to maintain consistency and eliminate case-sensitivity discrepancies. The results of this process are presented in Table 9.

Table 9. Results of lowercasing (in the Indonesian language)

Before lowercasing	After lowercasing
[CHEAT atau CHAT? Pengaruh ChatGPT dalam Dunia Pendidikan] Halo, Sobat Oranger! Kehadiran ChatGPT perlu disikapi secara bijak. Meskipun menawarkan berbagai kemudahan, kemunculannya juga menghadirkan tantangan baru yang perlu dihadapi oleh dunia pendidikan. Salah satu kemudahan tersebut adalah akses cepat untuk mendapatkan jawaban hanya dengan mengetik pertanyaan ke dalam fitur ChatGPT. - https://t.co/paN44YCMU8	[cheat atau chat? pengaruh chatgpt dalam dunia pendidikan] halo, sobat oranger! kehadiran chatgpt perlu disikapi secara bijak. meskipun menawarkan berbagai kemudahan, kemunculannya juga menghadirkan tantangan baru yang perlu dihadapi oleh dunia pendidikan. salah satu kemudahan tersebut adalah akses cepat untuk mendapatkan jawaban hanya dengan mengetik pertanyaan ke dalam fitur chatgpt. - https://t.co/pan44ycmu8

Source : (Research Results, 2025)

6. Truncation and Padding Results: To ensure uniformity in processing, lengthy texts were truncated. The outcomes of this procedure are presented in Table 10.

Table 10. Truncation and Padding results (in the Indonesian language)

Before padding and truncation	After padding and truncation
Meningkatkan Kemampuan Menulis: Panduan Efektif Membuat Buku dan E-Book dengan ChatGPT https://t.co/IQxUKx21Ta #BeritaJakarta #BeritaUtama #Pendidikan #SosialBudaya #Teknologi https://t.co/GSYddGxwdB	meningkatkan kemampuan menulis: panduan efektif membuat buku dan e-book dengan chatgpt

Source : (Research Results, 2025)

7. BERT-Specific Tokenization Results: The text is segmented into subword tokens according to the linguistic rules applied by the BERT model. The resulting tokenized output is presented in Table 11.

Table 11. Tokenizing results (in the Indonesian language)

Before text tokenizing	After text tokenizing
Strategi Institusi Pendidikan Mencegah Kecurangan di Era ChatGPT	['Strategi' ,'Institusi' ,'Pendidikan' ,'Mencegah' ,'Kecurangan' ,'di' ,'Era' ,'ChatGPT']

Source : (Research Results, 2025)

8. Lemmatization results: Words are normalized to their root or dictionary form (e.g., 'running' becomes 'run') to enhance linguistic consistency. The results are presented in Table 12.

Table 12. Lemmatization result (in the Indonesian language)

Before Lemmatization	After Lemmatization
Dunia pendidikan menghadapi tantangan baru pasca pandemi Covid-19 dengan hadirnya ChatGPT	Dunia Didik Hadap Tantang Baru Pasca Pandemi Covid-19 Dengan Hadir Chatgpt

Source : (Research Results,2025)

9. Stemming Results: This method transforms words into their root forms by eliminating suffixes; for example, ‘running’ is reduced to ‘run’. The results of this process are summarized in Table 13.

Table 13. Stemming Results (in the Indonesian language)

Before Stemming	After Stemming
KPT Mendukung Pemanfaatan ChatGPT dalam Pendidikan Tinggi	KPT Dukung Manfaat ChatGPT dalam Didik Tinggi

Source : (Research Results, 2025)

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 1 AUGUST 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i1.6392

10. Conversion of Text into Data Results: The input text was transformed into numerical vector representations to enable computational processing and classification. The results of this transformation are summarized in Table 14.

Table 24. Vector Data Results (in the Indonesian language)

Before Data Teks Get Vector	Vector Data Results
ChatGPT sangat membantu saya ketika sedang bingung. Sekarang saya mulai tertarik untuk menjadi ahli dalam membuat prompt AI	[0.8928571428571429, 0.9642857142857143, 1.0, 1.0, 1.0]

Source : (Research Results, 2025)

D. K Value Results Using Silhouette Coefficient

In this stage, the vector data is used as input to the Silhouette Coefficient model to obtain the optimal cluster value. This cluster value will be applied to the SOM model. Table 14 shows that the optimal number of clusters is 9.

Table 15. Silhouette Coefficient results

K Value	Average Silhouette Score
2	0.048
3	0.052
4	0.061
5	0.067
6	0.084
7	0.092
8	0.105
9	0.108

Source : (Research Results,2025)

The Silhouette Coefficient model indicates an optimal cluster value of 9, which will be used in the SOM method for data clustering.

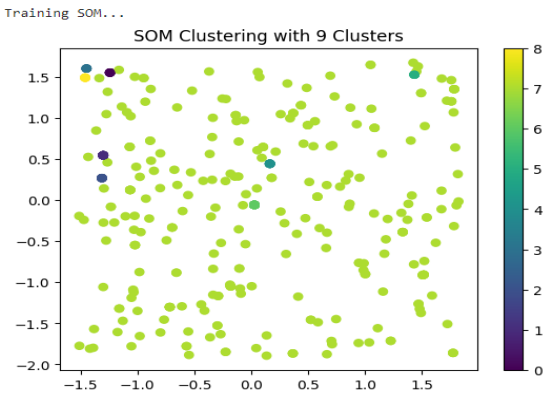
E. SOM Clustering Result

The Silhouette Coefficient determined the optimal cluster parameter, applied in the SOM stage. Text data was grouped into 9 clusters, as shown in Table 16 and Figure 2.

Table 16. Results of Data Clustering using SOM

cluster	Value
(2, 1)	[array([0.00021158, 0.00323711, -0.00163248, 0.00058241, -0.00990965])
(0, 0)	[array([8.5452683e-03, -5.8203419e-03, -7.6229847e-03, -1.3369779e-03])
(2, 0)	[array([-0.00057262, 0.00044173, 0.00503212, 0.0091439 , -0.00929612])
(1, 2)	[array([-0.0087137 , 0.00387609, 0.00522056, 0.00582366, 0.00751048])
(0, 1)	[array([3.34258475e-05, 3.20621976e-03, -6.81053149e-03, -1.29819987e-03])
(2, 2)	[array([7.60365045e-03, 9.99092311e-03, -8.20401311e-03, -3.05119320e-03])
(1, 0)	[array([-8.7579330e-03, 2.2427794e-03, -8.5118844e-04, -9.2808520e-03])

Source : (Research Results, 2025)



Source : (Research Results, 2025)
Figure 2. Clustering Results using SOM

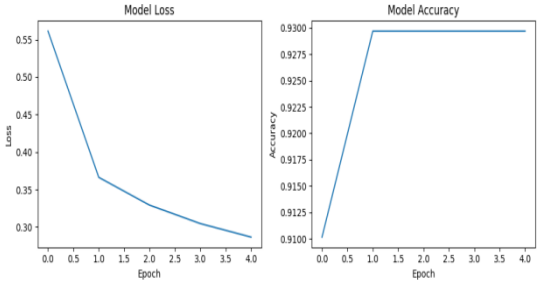
F. The Result Of The Fine-Tuned BERT Model

Fine-tuning the BERT model involves adjusting its parameters using data tailored to a specific task, thereby improving its capacity to interpret intricate language features. This improvement boosts key performance indicators—including accuracy, precision, and recall—by tailoring the model to better fit the contextual requirements of the target task. Results in Table 17 show enhanced sentiment classification accuracy and convergence, illustrated in Figures 3 and 4.

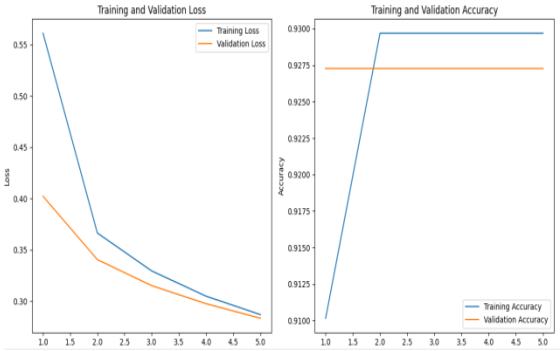
Table 17. Softmax Activation Function

	Softmax Activation Function					
No	Som_Size	Learning_rate	num_epoch_som	num_epoch_BERT	batch_size	Acuracy
1	9	0.01	100	128	16	91.80%
2	9	0.08	100	128	32	91.01%
3	9	0.001	100	128	64	92.06%

Source : (Research Results,2025)



Source : (Research Results, 2025)
Figure 3. BERT Model Performance Graph



Source : (Research Results, 2025)

Figure 4. Validation Results Graph for the BERT Model

E. Analysis and Discussion

The literature affirms that BERT remains robust in sentiment analysis, requiring only basic preprocessing, and its performance benefits from additional external strategies. Knowledge-enriched BERT models excel in aspect-based sentiment analysis [6], [7]. BERT outperforms Naïve Bayes, LSTM, and SVM in complex and multilingual data, while K-Means and KNN struggle with high-dimensional or nonlinear data [2], [28]. SOM faces challenges in optimal cluster selection, and LSTM and FRCNN show limited performance with lengthy, unstructured texts [16], [17].

CONCLUSION

This study concludes that the Silhouette Coefficient is a highly effective method for determining optimal cluster parameters, significantly enhancing the clustering results achieved through the SOM technique. The research findings indicate that utilizing the Silhouette Coefficient not only improves the clustering outcomes but also optimizes classification results, leading to accelerated convergence in the classification process. Furthermore, the integration of comprehensive preprocessing stages with the BERT model plays a crucial role in the effective processing of text data. The methodology employed for labeling text data using sentiment lexicons has been shown to enhance the accuracy of public sentiment analysis. Overall, the results of this research provide strong empirical support for the effectiveness of combining these advanced techniques in improving sentiment analysis on large datasets, thereby contributing valuable insights into the field.

REFERENCE

[1] L. Marron, “Exploring the potential of ChatGPT 3.5 in higher education: Benefits, limitations, and academic integrity,” in *Handbook of Research on Redesigning Teaching, Learning, and Assessment in the Digital Era*, IGI Global, 2023, pp. 326–349. doi: 10.4018/978-1-6684-8292-6.ch017.

[2] F. W. Putra, I. B. Rangka, S. Aminah, and M. H. R. Aditama, “ChatGPT in the higher education environment: Perspectives from the theory of high order thinking skills,” *Journal of Public Health (United Kingdom)*, vol. 45, no. 4, pp. e840–e841, Dec. 2023, doi: 10.1093/pubmed/fdad120.

[3] T. Adiguzel, M. H. Kaya, and F. K. Cansu, “Revolutionizing education with AI: Exploring the transformative potential of ChatGPT,” 2023, *Bastas*. doi: 10.30935/cedtech/13152.

[4] P. Shah, H. Patel, and P. Swaminarayan, “Multitask Sentiment Analysis and Topic Classification Using BERT,” *ICST Transactions on Scalable Information Systems*, vol. 11, Jul. 2024, doi: 10.4108/eetsis.5287.

[5] L. He, “Enhanced twitter sentiment analysis with dual joint classifier integrating RoBERTa and BERT architectures,” *Front Phys*, vol. 12, 2024, doi: 10.3389/fphy.2024.1477714.

[6] M. Pota, M. Ventura, R. Catelli, and M. Esposito, “An effective bert-based pipeline for twitter sentiment analysis: A case study in Italian,” *Sensors (Switzerland)*, vol. 21, no. 1, pp. 1–21, Jan. 2021, doi: 10.3390/s21010133.

[7] A. Zhao and Y. Yu, “Knowledge-enabled BERT for aspect-based sentiment analysis,” *Knowl Based Syst*, vol. 227, Sep. 2021, doi: 10.1016/j.knosys.2021.107220.

[8] Ms. M. P. Geetha and D. K. Renuka, “Improving the performance of aspect based sentiment analysis using fine-tuned Bert Base Uncased model,” *Int. J. Intell. Networks*, vol. 2, pp. 64–69, 2021, [Online]. Available: https://api.semanticscholar.org/CorpusID:238890573

[9] F. M. Sinaga, R. Purba, S. J. Pipin, W. S. Lestari, and S. Winardi, “Optimization of Sentiment Analysis Classification of

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 1 AUGUST 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i1.6392

ChatGPT on Big Data Twitter in Indonesia using BERT," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 8, no. 3, p. 1665, Jul. 2024, doi: 10.30865/mib.v8i3.7861.

[10] A. Subakti, H. Murfi, and N. Hariadi, "The performance of BERT as data representation of text clustering," *J Big Data*, vol. 9, no. 1, Dec. 2022, doi: 10.1186/s40537-022-00564-9.

[11] A. Rajan and M. Manur, "Aspect based sentiment analysis using fine-tuned BERT model with deep context features," *IAES International Journal of Artificial Intelligence*, vol. 13, no. 2, pp. 1250–1261, Jun. 2024, doi: 10.11591/ijai.v13.i2.pp1250-1261.

[12] J. Ma, "Using the Bert model and the attention mechanism to obtain an accurate sentiment analysis model," 2024.

[13] P. Akter *et al.*, "Sentiment Analysis of Consumer Feedback and Its Impact on Business Strategies by Machine Learning," *The American Journal of Applied Sciences*, vol. 07, no. 01, pp. 6–16, Jan. 2025, doi: 10.37547/tajas/Volume07Issue01-02.

[14] Nikhil Sanjay Suryawanshi, "Sentiment analysis with machine learning and deep learning: A survey of techniques and applications," *International Journal of Science and Research Archive*, vol. 12, no. 2, pp. 005–015, Jul. 2024, doi: 10.30574/ijsra.2024.12.2.1205.

[15] S. Efendi and P. Sihombing, "Sentiment Analysis of Food Order Tweets to Find Out Demographic Customer Profile Using SVM," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 3, pp. 583–594, Jul. 2022, doi: 10.30812/matrik.v21i3.1898.

[16] F. M. Sinaga, S. J. Pipin, S. Winardi, K. M. Tarigan, and A. P. Brahmana, "Analyzing Sentiment with Self-Organizing Map and Long Short-Term Memory Algorithms," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 23, no. 1, pp. 131–142, Nov. 2023, doi: 10.30812/matrik.v23i1.3332.

[17] S. J. Pipin, F. M. Sinaga, S. Winardi, and M. N. Hakim, "Sentiment Analysis Classification of ChatGPT on Twitter Big Data in Indonesia Using Fast R-CNN," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 7, no. 4, p. 2137, Oct. 2023, doi: 10.30865/mib.v7i4.6816.

[18] M. Pota, M. Ventura, H. Fujita, and M. Esposito, "Multilingual evaluation of pre-processing for BERT-based sentiment analysis of tweets," *Expert Syst Appl*, vol. 181, Nov. 2021, doi: 10.1016/j.eswa.2021.115119.

[19] J. Lu and H. Gweon, "Random k conditional nearest neighbor for high-dimensional data," *PeerJ Comput Sci*, vol. 11, 2025, doi: 10.7717/PEERJ-CS.2497.

[20] O. Ndama, I. Bensassi, and E. M. En-Naimi, "The impact of BERT-infused deep learning models on sentiment analysis accuracy in financial news," *Bulletin of Electrical Engineering and Informatics*, vol. 14, no. 2, pp. 1231–1240, Apr. 2025, doi: 10.11591/eei.v14i2.8469.

[21] J. Sun, M. Wang, D. Ren, and D. Chen, "Research and Application of Text-Based Sentiment Analytics," in *Frontiers in Artificial Intelligence and Applications*, IOS Press BV, 2024, pp. 619–629. doi: 10.3233/FAIA241391.

[22] Z. Su, "Applications of BERT in sentimental analysis," *Applied and Computational Engineering*, vol. 92, no. 1, pp. 147–152, Oct. 2024, doi: 10.54254/2755-2721/92/20241711.

[23] O. Ndama, I. Bensassi, and E. M. En-Naimi, "The impact of BERT-infused deep learning models on sentiment analysis accuracy in financial news," *Bulletin of Electrical Engineering and Informatics*, vol. 14, no. 2, pp. 1231–1240, Apr. 2025, doi: 10.11591/eei.v14i2.8469.

[24] P. Tisna Putra, A. Anggrawan, and H. Hairani, "Comparison of Machine Learning Methods for Classifying User Satisfaction Opinions of the PeduliLindungi Application," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 22, no. 3, pp. 431–442, Jun. 2023, doi: 10.30812/matrik.v22i3.2860.

[25] F. Sinaga, S. Winardi, and Gunawan, "3SV-KNNC Optimization using SVR and LMKNN for Stock Price Prediction," Jan. 2022, pp. 1–6. doi: 10.1109/ICOSNIKOM56551.2022.10034892.

[26] H. Mulyani, R. A. Setiawan, and H. Fathi, "Optimization of K Value in Clustering Using Silhouette Score (Case Study: Mall Customers Data)," *Journal of Information*

Technology and Its Utilization, vol. 6, no. 2, pp. 45–50, Dec. 2023, doi: 10.56873/jitu.6.2.5243.

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

[28] Q. Hu, "A cross-language short text classification model based on BERT and multilayer collaborative convolutional neural network (MCNN)," *MCB Molecular and Cellular Biomechanics*, vol. 21, no. 3, 2024, doi: 10.62617/mcb739.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri