VOL. 11. NO. 1 AUGUST 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v11i1.6549

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

FINE-GRAINED SENTIMENT ANALYSIS ON BIG DATA FROM MULTI-PLATFORM IN INDONESIA

Ronsen Purba[1]; Frans Mikael Sinaga[1*]; Sio Jurnalis Pipin[1]; Kelvin[1]

Informatics[1]
Universitas Mikroskil, Medan, Indonesia [1]
https://mikroskil.ac.id/[1]
ronsen@mikroskil.ac.id*, frans.sinaga@mikroskil.ac.id*, sio.pipin@mikroskil.ac.id,
kelvin.chen@mikroskil.ac.id[4]

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

Abstract— *Sentiment analysis on multi-platform big data in Indonesia presents a complex challenge, particularly in optimizing sentiment classification with higher granularity levels. This study aims to develop and optimize a sentiment classification model for analyzing public opinion on ChatGPT using a Fine-Grained Sentiment Analysis approach based on Indonesian Bidirectional Encoder Representations from Transformers (IndoBERT). The method is applied to big data collected from various social media platforms to improve accuracy and precision in identifying a broader spectrum of sentiments, including highly positive, positive, neutral, negative, and highly negative categories. A comparative analysis was conducted on different base models, including BERT, RoBERTa, and IndoBERT, to determine the most effective model. Experimental results show that the optimized IndoBERT model achieves an accuracy of 96% and outperforms other models in terms of precision and F1-score across all sentiment categories. Additionally, this study evaluates the model's computational efficiency and adaptability to diverse data. Thus, the developed model can serve as a more effective solution for gaining deeper insights into public opinion across various digital platforms in Indonesia.*

Keywords: *big data, chatgpt, fine-grained sentiment, indobert, sentiment analysis.*

Intisari— *Analisis sentimen pada big data multi-platform di Indonesia menjadi tantangan yang kompleks, terutama dalam mengoptimalkan klasifikasi sentimen dengan tingkat granularitas yang lebih tinggi. Penelitian ini bertujuan untuk mengembangkan dan mengoptimalkan model klasifikasi analisis sentimen mengenai opini publik terhadap ChatGPT dengan metode Fine-Grained Sentiment Analysis berbasis Indonesian Bidirectional Encoder Representations from Transformers (IndoBERT). Metode ini diterapkan pada big data yang dikumpulkan dari berbagai platform media sosial untuk meningkatkan akurasi dan ketepatan dalam mengidentifikasi spektrum sentimen yang lebih luas, termasuk kategori sangat positif, positif, netral, negatif, dan sangat negatif. Analisis perbandingan dilakukan terhadap base model yang berbeda, termasuk BERT, RoBERTa, dan IndoBERT, dilakukan untuk menentukan model yang paling efektif. Hasil eksperimen menunjukkan bahwa model IndoBERT yang dioptimalkan mampu mencapai akurasi sebesar 96%, serta unggul dalam hal presisi dan F1-score di semua kategori sentimen dibandingkan model lainnya. Selain itu, penelitian ini juga mengevaluasi efisiensi komputasi dan adaptabilitas model terhadap data yang beragam. Dengan demikian, model yang dikembangkan dapat menjadi solusi yang lebih efektif dalam memahami opini publik secara lebih mendalam di berbagai platform digital di Indonesia.*

Kata Kunci: *data besar, chatgpt, sentimen terperinci, indobert, analisis sentimen.*

INTRODUCTION

Sentiment analysis on big data across multiple platforms in Indonesia presents significant challenges, particularly in optimizing classification with higher granularity levels [1]. The growing volume of unstructured text data from social media requires advanced techniques to accurately capture

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.v11i1.6549

sentiment variations and contextual nuances [2], [3]. Traditional sentiment analysis models often fail to provide precise classifications due to inadequate preprocessing, ineffective vectorization, and limited context understanding [4], [5], [6]. Recent advancements in deep learning, particularly transformer-based architectures such as BERT (Bidirectional Encoder Representations from Transformers), have significantly improved sentiment classification performance [7].

BERT eliminates the need for extensive preprocessing [8] and enhances classification accuracy through contextual embedding [9], [10]. IndoBERT, a variant of BERT pre-trained on Indonesian text corpora, is specifically designed to handle linguistic characteristics unique to the Indonesian language [11], [12]. Compared to standard BERT, IndoBERT provides better contextual representation for Indonesian text by addressing unique morphology, syntax, and word semantics unique to the language, leading to improved classification accuracy in sentiment analysis tasks [13], [14], [15]. Fine-grained sentiment analysis extends traditional sentiment classification by moving beyond basic polarities (positive, negative, neutral) to include more detailed sentiment categories such as very positive, positive, neutral, negative, and very negative [16]. This approach enables a deeper understanding of sentiment intensity, emotional variations, and nuanced expressions in text [17].

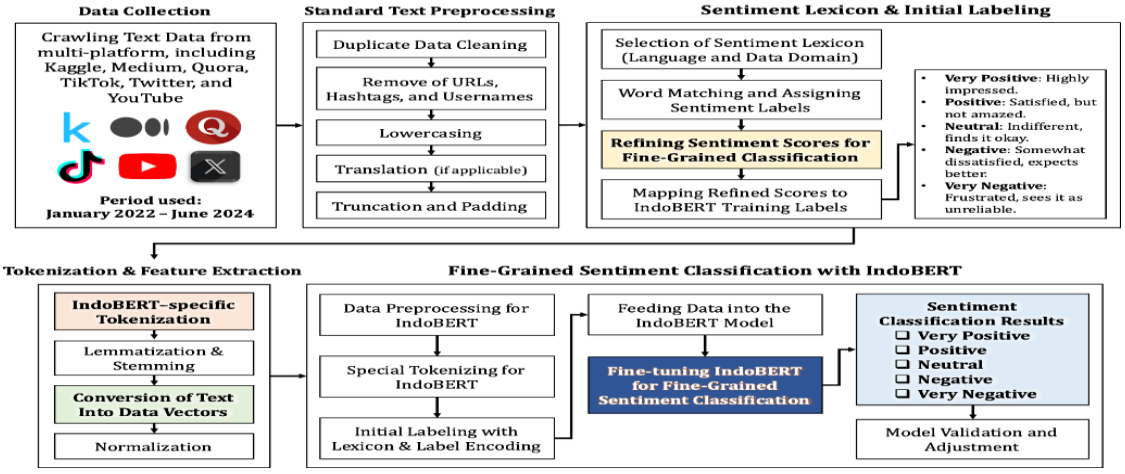
Fine-grained analysis is particularly crucial in Indonesian multi-platform data, where user-generated content often exhibits diverse linguistic styles, mixed sentiments, and implicit emotions [16], [17], [18]. IndoBERT's context-aware word representations allow for better sentiment classification by dynamically capturing long-range dependencies and meaning shifts in text [19], [20].

This study aims to develop and optimize a Fine-Grained Sentiment Analysis model using IndoBERT on multi-platform big data in Indonesia. The research focuses on improving classification accuracy through effective preprocessing, Lexicon-based sentiment labeling, and IndoBERT-based classification. By leveraging IndoBERT's capabilities, this study provides a robust methodology for understanding public sentiment in diverse digital environments, offering valuable insights into various societal and economic trends in Indonesia [12], [21]. While the application of transformer-based models like IndoBERT for Indonesian sentiment analysis is established, this study presents a novel and comprehensive approach by tackling the challenge of fine-grained sentiment analysis on large-scale, multi-platform big data in the Indonesian context. Our primary contribution lies in developing and optimizing a robust pipeline specifically for the noisy and diverse nature of user-generated content from platforms like Twitter, TikTok, and YouTube.

This environment is characterized by informal language, slang, code-switching, and platform-specific jargon, which pose significant challenges for standard models. We introduce a tailored preprocessing methodology combined with a lexicon-based labeling strategy to enhance the fine-tuning of IndoBERT for five distinct sentiment categories, a level of granularity often overlooked in previous studies. This research provides a more nuanced understanding of public opinion in Indonesia's complex digital ecosystem.

MATERIALS AND METHODS

The problem-solving stages, shown in Figure 1.



Source : (Research Results,2025)

Figure 1. The Steps in Problem Solving Approach.

In figure 1, include data collection, preprocessing, and fine-grained sentiment labeling. The sentiment labeling process utilizes a Sentiment Lexicon to categorize text data into five classes: very positive, positive, neutral, negative, and very negative. To improve classification accuracy, the IndoBERT model is applied to analyze the fine-grained sentiment, leveraging its deep contextual understanding to capture subtle variations in sentiment expression across multiple social media platforms.

A. Data Preparation

Data preparation involves crawling text data from multiple platforms, including Kaggle, Medium, Quora, TikTok, Twitter, and YouTube, using relevant keywords for this research [5].

Table 1. List of Dataset Keywords Used

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 multiple platforms as mentioned before. Additional datasets were incorporated with new keywords, expanding on previous research. Table 2 showcases sample rows from the collected raw dataset, where each entry contains several key fields. These fields include the URL, a direct link to the original post on the social media platform (e.g., x.com); the Date, which is the timestamp indicating when the content was posted; and Text Content, representing the text of the post. Furthermore, each entry is assigned a unique ID and conversation ID (Conv. ID), along with engagement metrics like Replies and Likes. The dataset also captures the Username of the account that published the post and a preliminary Label (e.g., Positive, Neutral, Negative) that was assigned during initial data collection or preprocessing

B. Pre-processing Data

In the preprocessing stage, a novel approach with IndoBERT prepares data. Sentiment Lexicon labels data using heuristic sentiment aggregation, while IndoBERT generates vector representations capturing text characteristics.

Table 2. Raw Data Entries from the Collected Dataset

URL	Date	Text Content	ID	Replies	Likes	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.	5935747428	0	0	1,61272E+18	InfoKomputer	Netral
https://x.com/naimalkalantani/status/1656498795914620932	Thu May 11 03:17:43 +0000 2023	Jemput baca tulisan terbaru saya mengenai ChatGPT. Impak paling besar yang saya dapat fikirkan sekarang ialah ChatGPT ini akan memberi kesan dalam pendidikan.	7168424458	0	8	1,6565E+18	naimalkalantani	Positif
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.	1184691158	11	132	1,61088E+18	AlphaARachman	Negatif

Source : (Research Results, 2025)

Fine-Grained Sentiment Labeling: To enhance sentiment classification, a fine-grained sentiment labeling approach is implemented. The Sentiment Lexicon assigns scores to textual data, categorizing them into five sentiment levels: very positive, positive, neutral, negative, and very negative. These fine-grained labels are then used for model training, allowing IndoBERT to capture nuanced sentiment expressions more accurately. Applying Sentiment Lexicon: This step is performed to label the data by combining sentiment values using heuristic rules described in Equation 1.

$$S_{\text{very positive}} \sum_{i \in t}^{n} \text{very positive score}_i \quad (1)$$

$$S_{\text{positive}} \sum_{i \in t}^{n} \text{positive score}_i \quad (2)$$

$$S_{\text{neutral}} \sum_{i \in t}^{n} \text{neutral score}_i \quad (3)$$

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.v11i1.6549

$$S_{\text{negative}} \sum_{i\in t}^{n} \text{negative score}_i \quad (4)$$

$$S_{\text{very negative}} \sum_{i\in t}^{n} \text{very negative score}_i \quad (5)$$

The sentiment lexicon-based labeling is a crucial step for creating our fine-grained training labels. This process operates on a word-by-word basis. As outlined in Algorithm 1, the system first scans the input text. For each word, it checks for its presence in a predefined Indonesian sentiment lexicon. This lexicon is a dictionary where each word is assigned a polarity score (e.g., +1 for positive words like 'terbantu' (helped), -1 for negative words like 'galau' (upset), and 0 for neutral words).

A cumulative score for the entire text is calculated by aggregating the scores of the individual words. This aggregate score is then normalized, often by dividing by the number of words with sentiment scores, as shown in the corrected Equation 8. Finally, this continuous score is mapped to one of the five discrete, fine-grained sentiment categories (very positive, positive, neutral, negative, very negative) using a set of heuristic thresholds, as detailed in Algorithm 2. This method allows us to transform unstructured text into consistently labeled data suitable for training the IndoBERT model.

C. Classification using IndoBERT

BERT, an advanced transformer model, is used for sentiment analysis. Fine-tuning on fine-grained labeled data enables BERT to classify sentiment into five categories with high accuracy. Its deep contextual understanding allows it to capture subtle sentiment variations across different languages and domains. The stages of the IndoBERT model include:

1. Data Processing for IndoBERT: Each token needs to be tokenized or transformed into vector representations using embedding techniques.
$$V_t = Trans\ (W_t X_a + b_t) \quad (6)$$
2. Feeding Data into the IndoBERT Model: The IndoBERT model processes text and generates vector representations for each token in the text.
$$\mathrm{V_b} = \mathrm{Bert}(\mathrm{W_b X_b} + \mathrm{b_b}) \quad (7)$$
3. Initial Labeling with Sentiment Lexicon: Sentiment lexicon-based labeling is a widely used approach for sentiment classification, where words and phrases are assigned sentiment scores based on predefined polarity values [23], [24].
$$\sum_{i=1}^{k} \frac{(p-n)_i}{k} \quad (8)$$
Where k represents the total number of words in the text that are listed in the lexicon, and $p - n$ denotes the sentiment polarity score associated with each word in the lexicon, which is determined based on its occurrence in positive and negative categories [25].
4. Self-Attention to produce better representations of each word within the context of a sentence or text.
$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (9)$$
Where Q, K, V are matrices of query, key, and value associated with tokens in the input.
5. Feed-Forward Neural Networks generate complex token representations after the Self-Attention stage, enhancing the model's ability to handle nonlinear textual information and improve comprehension and generalization.
$$FFN(H) = ReLU(HW_1 + b_1)W_2 + b_2 \quad (10)$$
Where H is the context vector resulting from the self-attention mechanism, and W1, W2, b1, b2 are the parameters of the feed-forward network layers.
6. Fine-tuning IndoBERT involves retraining the model on domain-specific sentiment data, optimizing it for better classification performance while maintaining contextual integrity.
7. Sentiment Classification: The IndoBERT model will process the text and classify it into five fine-grained sentiment categories based on the generated vector representations.

D. Data Testing

The model's performance is evaluated using a confusion matrix, which assesses accuracy based on four components: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). TN represents correctly classified negative data, FP indicates negatives misclassified as positives, TP denotes correctly identified positives, and FN reflects positives misclassified as negatives. Accuracy testing, described in Equation 9, is utilized to measure the effectiveness of the classification method.

RESULTS AND DISCUSSION

The findings of this study include data collection, preprocessing, sentiment labeling using a Sentiment Lexicon, and fine-grained sentiment classification using the IndoBERT model. The

preprocessing phase involves text normalization, partial data translation of English terms into Indonesian, tokenization, truncation, and padding to optimize input representation for IndoBERT.

A. The Results of Crawling the Dataset

The extracted dataset consists approximately 100,000 entries collected from multiple platforms, including Kaggle, Medium, Quora, TikTok, Twitter, and YouTube, spanning from January 2022 to June 2024. However, for this study, subset data of 25,000 sample data was selected to ensure balanced representation and manageable computational processing while maintaining analytical depth.

Table 3. The Crawled Dataset Results

URL	Date	Sentiment	ID	Replies	Likes
https://x.com/ndorokakung/status/1674551525799088128	Thu Jun 29 22:52:50 +0000 2023	Sebagian guru mulai galau sejak kemunculan ChatGPT. Dengan kecerdasan buatan [AI] seperti ChatGPT guru merasa terbantu sekaligus terancam eksistensinya. Bagaimana sebenarnya peran AI dalam mendukung pendidikan dan pembelajaran? -- sebuah utas -- https://t.co/FJ3kz6ckWr	18598424	13	375
https://x.com/InfoKomputer/status/1612723925166743552	Tue Jan 10 08:12:01 +0000 2023	Cegah siswa nyontek begini cara institusi pendidikan siasati ChatGPT - https://t.co/VWOhHaVLC5 https://t.co/WiPpMf4t99	59357428	0	0
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 https://t.co/y6CcSGJH39	1,18E+08	11	132

Source : (Research Results, 2025)

B. The Result of Initial Labeling with Sentiment Lexicon

The sentiment lexicon-based labeling process assigns sentiment scores based on predefined word polarities and contextual modifiers [23]. Each sentence is evaluated using a lexicon that captures sentiment intensity, allowing classification into five sentiment categories: very positive, positive, neutral, negative, and very negative, as summarized in Table 4. This fine-grained labeling provides a more detailed sentiment distribution, enabling deeper insights into sentiment intensity and emotional variations across multi-platform data.

Each word scores calculated with:

```
For each word in the text do
    if the word is in the positive list then
       sum = sum + 1
    else
       if the word is in the negative list
       then
           sum = sum - 1
       End if
    End if
End for
```

Algorithm 1. Pseudo-Algorithm to Label Each Word [26]

Each sentence scores then calculated with formula from Equation 8.

```
If score > 0.3 then
   label = 4
else
   if 0.1 < score <= 0.3 then
      label = 3
   else
      if -0.1 <= score <= 0.1 then
          label = 2
      else
         if -0.1 < score <= -0.3 then
            label = 1
         else
            if score < -0.3 then
               label = 0
            End if
         End if
      End if
   End if
End if
```

Algorithm 2. Pseudo-Algorithm to Fine-Grained Labeling Each Sentence

Where each label number represent the level of the sentiment label:

0 = very negative
1 = negative
2 = neutral
3 = positive
4 = very positive

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.v11i1.6549

Table 4. The Results of Labeling the Dataset using Lexicon

Lexicon Scores	Lexicon Result	Tweet	Date	Username	Sentimen Label			Final
-0.2	negative	Sebagian guru mulai galau sejak kemunculan ChatGPT. Dengan kecerdasan buatan [AI] seperti ChatGPT guru merasa terbantu sekaligus terancam eksistensinya. Bagaimana sebenarnya peran AI dalam mendukung pendidikan dan pembelajaran? -- sebuah utas -- https://t.co/FJ3kz6ckWr	Thu Jun 29 22:52:50 +0000 2023	ndorokakung	negative	negative	neutral	negative
0.3	positive	Jemput baca tulisan terbaru saya mengenai ChatGPT. Impak paling besar yang saya dapat fikirkan sekarang ialah ChatGPT ini akan memberi kesan dalam pendidikan. https://t.co/V0cuZIArfO https://t.co/vSYNFeEPo7	Thu May 11 03:17:43 +0000 2023	naimalkalantani	positive	positive	positive	positive
0.1	neutral	Dinas Pendidikan Kota New York jadi salah satu yang pertama memblokir akses ChatGPT. Chatbot milik OpenAI itu rentan disalahgunakan peserta didik. https://t.co/1sidoDr1cb https://t.co/1sBT02cFmr	Sat Jan 07 10:00:08 +0000 2023	VICE_ID	neutral	positive	neutral	neutral
0.4	positive	Thank @ditjendikti for another opportunity of research grant entitled Integrasi UTAUT2 TPB dan IS Success model: ChatGPT dalam pendidikan . We will work hard for the research & outcomes articles accepted in SSCI or Scopus indexed journals. https://t.co/cBMXpUSAfp	Tue Jun 06 05:53:39 +0000 2023	tjohor25	positive	positive	neutral	positive
0.8	very positive	In a world where advanced, artificial intelligence has become the norm. ChatGPT is the newest and most advanced chatbot on the market Capable of holding conversations with humans. ChatGPT is designed to help people with everyday tasks and make their lives easier. https://t.co/88LLi3nITV	Fri May 24 12:25:58 +0000 2024	user18587	positive	positive	positive	positive
-0.5	very negative	Pls don't share your personal information to chatgpt yang lagi rame. Sure, making it public can have it to learn lots of things, thus having better ai in a short time. But we never know if there are any backdoor yang bisa dipake company untuk ngelist your information for their use.	Sun Jul 16 03:54:19 +0000 2023	JusDooIt	neutral	negative	negative	negative

Source : (Research Results, 2025)

It is important to clarify that the sentiment analysis model does not have a limitation based on the number of sentences. The analysis is performed on the entire text content of a post (e.g., a full tweet).

However, like all transformer-based models, IndoBERT has a technical limitation regarding the maximum sequence length, which is the number of tokens it can process at once. For this study, texts longer than the model's maximum input size were truncated to ensure compatibility, as mentioned in the preprocessing stage. This is a standard practice and is based on token count, not sentence count. The classification result for each input text is a single sentiment category from the five defined classes (e.g., 'Negative'), which serves as the final label for that data point.

C. The Results of Data Preprocessing

Unstructured text is transformed and processed using techniques like punctuation removal, lowercasing, spelling normalization, filtering, and tokenization.

1. Data Cleaning results: This process removes unnecessary characters, such as punctuation and special characters. The results are shown in Table 5.

Table 5. Data Cleaning Results (in the Indonesia Language)

Data Cleaning Results	
Before cleaning text	After cleaning text
Sebagian guru mulai galau sejak kemunculan ChatGPT. Dengan kecerdasan buatan [AI] seperti ChatGPT guru merasa terbantu sekaligus terancam eksistensinya. Bagaimana sebenarnya peran AI dalam mendukung pendidikan dan pembelajaran? -- sebuah utas -- https://t.co/FJ3kz6ckWr	sebagian guru mulai galau sejak kemunculan chatgpt. dengan kecerdasan buatan ai seperti chatgpt guru merasa terbantu sekaligus terancam eksistensinya. bagaimana sebenarnya peran ai dalam mendukung pendidikan dan pembelajaran? sebuah utas https t.co fj kz ckwr

Source : (Research Results,2025)

2. Removal of URLs results: The step involves removing web links (URLs) from text, as they are typically irrelevant for most text analysis purposes. The results of this URL removal process are shown in Table 6.

Table 6. Removal URL Results (in the Indonesia Language)

Removal URL Results	
Before removal url text	After removal url text
Jemput baca tulisan terbaru saya mengenai ChatGPT. Impak paling besar yang saya dapat fikirkan sekarang ialah ChatGPT ini akan memberi kesan dalam pendidikan. https://t.co/V0cuZlArfO https://t.co/vSYNFeEPo7	jemput baca tulisan terbaru saya mengenai chatgpt. impak paling besar yang saya dapat fikirkan sekarang ialah chatgpt ini akan memberi kesan dalam pendidikan. https t.co v cuziarfo https t.co vsynfeepo

Source : (Research Results,2025)

3. Removal of hashtags results: The process involves removing hashtags (#) from text, as they are generally not relevant for most text analysis purposes. The results of this hashtag removal are presented in Table 7.

Table 7. Removal URL Results (in the Indonesia Language)

Before removal hashtags results	After removal hashtags results
Berawal dr organisasi nirlaba dan bikin aplikasi utk pendidikan skrg #ChatGPT #ArtificialIntelligence menyebar ke seluruh dunia dan dan digunakan dalam semua aspek kehidupan. #samaltman https://t.co/GyilKJHNMx	berawal dr organisasi nirlaba dan bikin aplikasi utk pendidikan skrg chatgpt artificialintelligence menyebar ke seluruh dunia dan dan digunakan dalam semua aspek kehidupan. samaltman https t.co gyilkjhnmx

Source : (Research Results,2025)

4. Removal of usernames results: Usernames (@) are removed, as they are usually irrelevant for text analysis. The results are shown in Table 8.

Table 8. Removal Username Results. (in the Indonesia Language)

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

Source : (Research Results,2025)

5. Lowercasing results: All characters are converted to lowercase to ensure uniformity and avoid case-sensitivity issues. The outcomes are shown in Table 9.

Table 9. Lowercasing Results (in the Indonesia Language)

Before removal lowercasing results	After removal lowercasing results
[CHEAT OR CHAT(?) Pengaruh Chat GPT Dalam Dunia Pendidikan] Halo sobat Oranger! Keberadaan ChatGPT perlu disikapi dengan bijak. Karena kemunculan chat GPT juga membawa tantangan yang perlu diatasi oleh pendidikan. Kemudahan yang didapatkan dengan bertanya kepada fitur ChatGPT - https://t.co/paN44YCMU8	[cheat or chat(?) pengaruh chat gpt dalam dunia pendidikan] halo sobat oranger! keberadaan chatgpt perlu disikapi dengan bijak. karena kemunculan chat gpt juga membawa tantangan yang perlu diatasi oleh pendidikan. kemudahan yang didapatkan dengan bertanya kepada fitur chatgpt - https://t.co/pan44ycmu8

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.v11i1.6549

6. Translation: Some data that originally written in English were translated into Indonesian to maintain consistency in the dataset and improving contextual understanding.

Table 10. Translation Result(in the Indonesia Language)

Before translation	After translation
chatgpt is simply amazing. could be a potential search replacement. i asked for article idea for bobhatamarathi and here is what it generated. time to dive into generativeai.	chatgpt sungguh menakjubkan. bisa menjadi pengganti pencarian yang potensial. saya meminta ide artikel untuk bobhatamarathi dan inilah yang dihasilkannya. saatnya menyelami ai generatif

Source : (Research Results,2025)

7. Truncation and Padding results: Long text was truncated for consistent processing. The results are shown in Table 11.

Table 11. Truncation and Padding Result (in the Indonesia Language)

Before padding and truncation	After padding and truncation
Mengasah Skill Menulis dengan ChatGPT: Cara Membuat Buku atau E-Book dengan Lebih Efektif https://t.co/IQxUKx21Ta #BeritaJakarta #BeritaUtama #Pendidikan #SosialBudaya #Teknologi https://t.co/GSYddGxwdB	mengasah skill menulis dengan chatgpt: cara membuat buku atasu e-book dengan lebih efektif

Source : (Research Results,2025)

8. Special Tokenization for IndoBERT results: Text is split into tokens based on the language structure used by IndoBERT. The outcomes are shown in Table 12.

Table 12. Tokenizing Result (in the Indonesia Language)

Before tokenizing text	After tokenizing text
Cegah siswa nyontek begini cara institusi pendidikan siasati ChatGPT	['Cegah' ,'siswa' ,'nyontek' ,'begini' ,'cara' ,'institusi' ,'pendidikan' ,'siasati' ,'ChatGPT']

Source : (Research Results,2025)

9. Lemmatization results: Words are converted to their base forms (e.g., 'running' to 'run'). The results are shown in Table 13.

Table 13. Lemmatization Result (in the Indonesia Language)

Before lemmatization	After lemmatization
ChatGPT Jadi Tantangan Baru Dunia Pendidikan Pasca Pandemi Covid-19	ChatGPT Jadi Tantang Baru Dunia Didik Pasca Pandemi Covid-19

Source : (Research Results,2025)

10. Stemming results: This method trims word endings to obtain base forms (e.g., 'running' to 'run'). The results are shown in Table 14.

Table 14. Stemming Result (in the Indonesia Language)

Before Stemming	After Stemming
KPT Tidak Menghalang Penggunaan ChatGPT Dalam Pendidikan Tinggi	KPT Tidak Halang Guna ChatGPT Dalam Didik Tinggi

Source : (Research Results,2025)

11. Conversion of Text into Data results: Text is transformed into numerical vector representations for analysis and classification. The results are shown in Table 15.

Table 15. Vector Data Result (in the Indonesia Language)

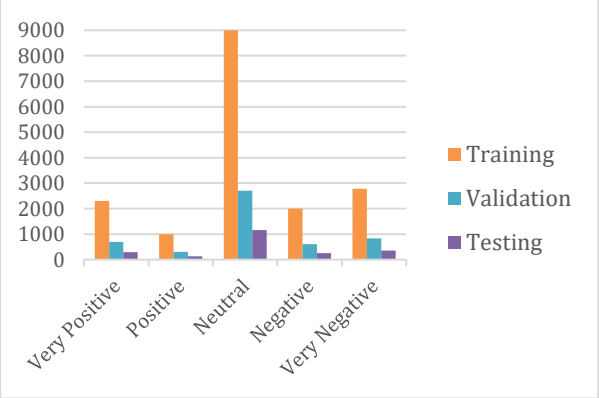
Before Text Data Get Vector	Vector Data Result
chatgpt lu membantu banget kegabu-tan gue. sekarang kayanya gue pengen jadi prompter AI aja deh	[0.8928571428571429, 0.9642857142857143, 1.0, 1.0, 1.0]

Source : (Research Results,2025)

D. The Result of the Fine-Tuned IndoBERT Model

Fine-tuning IndoBERT optimizes its parameters using task-specific training data, enhancing its ability to capture complex linguistic patterns. This process improves accuracy, precision, and recall by aligning the model's performance with task-specific nuances.

The dataset consists of approximately 24,000 samples with a distribution across fine-grained sentiment categories, as detailed in Table 16. Understanding the label distribution, as visualized in Figure 2, is crucial for evaluating IndoBERT's performance in handling imbalanced sentiment classes and addressing potential impacts on model training.



Source : (Research Results,2025)

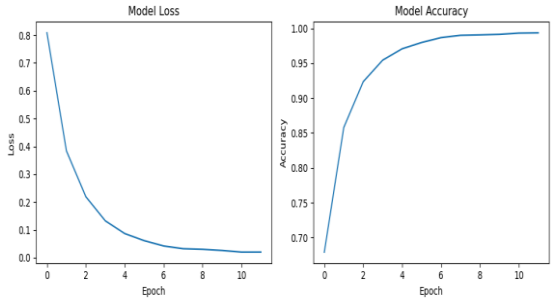
Figure 2. Distribution of Fine-Grained Sentiment Labels

Table 16. Fine-Grained Sentiment Label Distribution

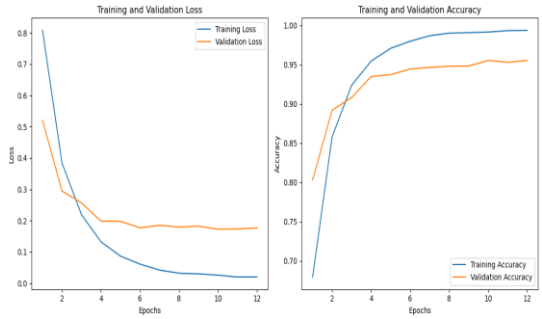
Label	Training	Validation	Testing
Very Positive	2,302	691	296
Positive	1,000	300	128
Neutral	8,999	2,700	1,158
Negative	2,004	601	258
Very Negative	2,786	836	358
Total Data	17,091	5,128	2,198

Source : (Research Results,2025)

Figures 3 and 4 further illustrate the model's convergence and learning progress over training epochs.



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



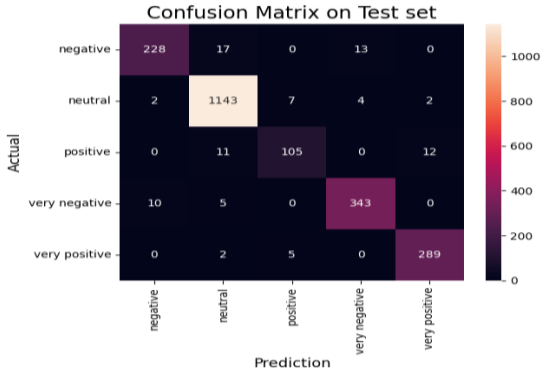
Source : (Research Results,2025)
Figure 4. Validation Data Graph of the IndoBERT Model

Table 17 presents the classification performance metrics, including accuracy, precision, recall and f1-score during testing mode, demonstrating the effectiveness of fine-tuned IndoBERT.

Table 17. Fine-Tuned IndoBERT Model Performance Metrics on Testing Mode

Label	Precision	Recall	F1-Score	Total Data
Very Positive	0.95	0.98	0.96	296
Positive	0.90	0.82	0.86	128
Neutral	0.97	0.99	0.98	1,158
Negative	0.95	0.88	0.92	258
Very Negative	0.95	0.98	0.96	296
Accuracy		95.905%		

Source : (Research Results,2025)



Source : (Research Results,2025)
Figure 5. Confusion Matrix on Testing Set

E. Analysis of Dataset Size on Model Performance

To evaluate the impact of training data size on model performance, experiments were conducted using different subsets of the dataset for training, validation, and testing with varying dataset sizes and label distributions. The results, summarized in Table 18 indicate that increasing the training dataset size generally improves classification accuracy.

Table 18. Performance Comparison Across Different Dataset Sizes

	Data Size			
Raw Data	850*	2,000	25,000	36,000
Training	577	980	17,091	21,867
Validation	159	294	5,128	6,747
Testing	88	126	2,198	2,624
Accuracy	100%	75%	96%	89%

*data collected from a single platform

Source : (Research Results,2025)

The perfect accuracy (100%) on the smallest dataset (850 raw data samples) suggests overfitting, as the model memorizes patterns instead of generalizing. Limited data reduces variation, making classification easier but less robust. However, beyond a certain threshold, additional data showed diminishing performance gains. These findings emphasize the need to balance dataset size, label distribution, and computational efficiency when fine-tuning IndoBERT for sentiment analysis.

F. Comparison of Different Base Models Performance

A performance comparison was conducted between BERT, IndoBERT, and RoBERTa for fine-grained sentiment classification. Each model was fine-tuned using the same dataset and hyperparameters to ensure a fair evaluation. The results, presented in Table 19, show that IndoBERT achieves the highest accuracy, outperforming the

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.v11i1.6549

other models in all sentiment categories. The hyperparameter settings used in the fine-tuning process for the three models can be seen in Table 20.

Table 19. Performance Comparison Across Different Base Models

Base Models	BERT	RoBERTa	IndoBERT
Data Size	12,000	24,000	24,000
Epoch	16	16	16
Learning Rate	9e-6	9e-6	9e-6
Training Accuracy	98,71%	95,99%	99,01%
Validation Accuracy	89,17%	89,37%	95,57%
Testing Accuracy	90,19%	89,58%	95,91%

Source : (Research Results,2025)

Table 20. Hyperparameter Settings for Model Fine-Tuning

Base Models	BERT	RoBERTa	IndoBERT
Pre-trained Model	bert-base-uncased	roberta-base	indobenchmark/indobert-base-p1
Max Sequence Length	256	256	256
Batch Size	16	16	16
Learning Rate	9,00E-06	9,00E-06	9,00E-06
Number of Epochs	16	16	16
Pre-trained Model	bert-base-uncased	roberta-base	indobenchmark/indobert-base-p1
Optimizer	AdamW	AdamW	AdamW
Adam Epsilon	1,00E-08	1,00E-08	1,00E-08
Weight Decay	0.01	0.01	0.01

Source : (Research Results,2025)

IndoBERT's superior performance can be attributed to its pretraining on Indonesian-language corpora, allowing better contextual understanding and word representations. In contrast, BERT (bert-base-uncased), which was pretrained on English text, struggles with tokenization and semantic understanding in Indonesian. RoBERTa (roberta-base), optimized for longer sequences and richer contextual representations, does not yield significant improvements in this task due to the lack of Indonesian-language adaptation.

The comparison highlights the importance of language-specific pretrained models in sentiment analysis, reinforcing IndoBERT as the optimal choice for handling fine-grained sentiment classification in Indonesian text.

G. Analysis and Discussion

Existing literature highlights IndoBERT's effectiveness in sentiment analysis with minimal preprocessing, further enhanced by integrating knowledge-based methods such as Sentiment Lexicon. Fine-grained sentiment analysis, which classifies text into multiple sentiment levels, benefits significantly from Sentiment Lexicon integration, improving classification granularity.

Compared to traditional classifiers such as Naïve Bayes, LSTM, and SVM, transformer-based models like IndoBERT offer superior performance in handling sentiment variations, particularly in complex and multilingual datasets. IndoBERT's self-attention mechanisms allow it to retain long-range dependencies, making it highly effective for processing unstructured text [19], [20]. Moreover, Sentiment Lexicon-based labeling enhances the model's ability to capture nuanced emotions [2], [3], complementing IndoBERT's contextual understanding [14], [15]. While LSTM and CNN-based models face challenges in handling lengthy and context-dependent texts [5], [6], IndoBERT's deep contextual embeddings ensure more accurate sentiment classification by dynamically adjusting word representations based on surrounding text [13].

Our results demonstrate that the fine-tuned IndoBERT model achieves superior performance with 96% accuracy compared to other transformer-based models like BERT and RoBERTa. This superiority is largely attributable to IndoBERT's pre-training on a large Indonesian corpus, which allows it to better capture the language's unique syntax and semantics. When contextualized with traditional machine learning and deep learning approaches, the advantages of our model become even clearer. For instance, models like Naïve Bayes and SVM, while effective for basic text classification, often fail to grasp complex contextual nuances. Similarly, sequential models like LSTMs can struggle with long-range dependencies in text, a limitation that the transformer architecture's self-attention mechanism effectively overcomes. By integrating a sentiment lexicon for fine-grained labeling with IndoBERT's powerful contextual embeddings, our approach provides a more robust and accurate solution than what is typically achievable with these older methods in the complex domain of multi-platform social media analysis. The combination of IndoBERT and Sentiment Lexicon provides a robust framework for fine-grained sentiment analysis across multiple platforms.

CONCLUSION

The conclusion of this study is that the integration of Sentiment Lexicon with fine-grained sentiment labeling significantly enhances sentiment classification performance. By categorizing text into

five sentiment classes—very positive, positive, neutral, negative, and very negative—this approach provides a more detailed sentiment representation compared to traditional classification methods.

Additionally, IndoBERT's contextual learning capabilities, combined with lexicon-based labeling, improve sentiment analysis accuracy, allowing for better detection of nuanced sentiment expressions across diverse social media data in Indonesian. The preprocessing techniques used in this study further aid in text normalization, ensuring cleaner inputs for the IndoBERT model.

The findings demonstrate that leveraging fine-grained sentiment analysis, instead of relying solely on binary or ternary sentiment categories, enhances sentiment classification robustness. Future work may explore further improvements in lexicon-based sentiment scoring and the potential integration of multi-modal sentiment analysis to expand classification performance.

REFERENCE

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

[2] A. S. George and T. Baskar, "Leveraging Big Data and Sentiment Analysis for Actionable Insights: A Review of Data Mining Approaches for Social Media," 2024, doi: 10.5281/zenodo.13623777.

[3] R. Nakka, T. S. Lakshmi, D. Priyanka, N. R. Sai, S. P. Praveen, and U. Sirisha, "LAMBDA: Lexicon and Aspect-Based Multimodal Data Analysis of Tweet," *Ingénierie des systèmes d information*, vol. 29, no. 3, pp. 1097–1106, Jun. 2024, doi: 10.18280/isi.290327.

[4] 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.

[5] 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.

[6] 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.

[7] M. D. Deepa and A. Tamilarasi, "Bidirectional Encoder Representations from Transformers (BERT) Language Model for Sentiment Analysis task: Review," 2021.

[8] 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.

[9] 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.

[10] 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

[11] F. Baharuddin and M. F. Naufal, "Fine-Tuning IndoBERT for Indonesian Exam Question Classification Based on Bloom's Taxonomy," *Journal of Information Systems Engineering and Business Intelligence*, vol. 9, no. 2, pp. 253–263, Nov. 2023, doi: 10.20473/jisebi.9.2.253-263.

[12] E. Yulianti and N. K. Nissa, "ABSA of Indonesian customer reviews using IndoBERT: single- sentence and sentence-pair classification approaches," *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 5, pp. 3579–3589, Oct. 2024, doi: 10.11591/eei.v13i5.8032.

[13] H. Imaduddin, F. Y. A'la, and Y. S. Nugroho, "Sentiment Analysis in Indonesian Healthcare Applications using IndoBERT Approach," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 8, 2023, doi: 10.14569/IJACSA.2023.0140813.

[14] Taufiq Dwi Purnomo and Joko Sutopo, "COMPARISON OF PRE-TRAINED BERT-BASED TRANSFORMER MODELS FOR REGIONAL LANGUAGE TEXT SENTIMENT ANALYSIS IN INDONESIA," *International Journal Science and Technology*, vol. 3, no. 3,

VOL. 11. NO. 1 AUGUST 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i1.6549

pp. 11–21, Nov. 2024, doi: 10.56127/ijst.v3i3.1739.

[15] H. Ahmadian, T. F. Abidin, H. Riza, and K. Muchtar, "Hybrid Models for Emotion Classification and Sentiment Analysis in Indonesian Language," *Applied Computational Intelligence and Soft Computing*, vol. 2024, no. 1, Jan. 2024, doi: 10.1155/2024/2826773.

[16] L. Zhu, Y. Xu, Z. Zhu, Y. Bao, and X. Kong, "Fine-Grained Sentiment-Controlled Text Generation Approach Based on Pre-Trained Language Model," *Applied Sciences*, vol. 13, no. 1, p. 264, Dec. 2022, doi: 10.3390/app13010264.

[17] W. Sofiya and E. B. Setiawan, "FINE-GRAINED SENTIMENT ANALYSIS IN SOCIAL MEDIA USING GATED RECURRENT UNIT WITH SUPPORT VECTOR MACHINE," *Jurnal Teknik Informatika (Jutif)*, vol. 4, no. 3, pp. 511–519, Jun. 2023, doi: 10.52436/1.jutif.2023.4.3.855.

[18] J. Wang, Y. Wang, Z. Zhang, J. Zeng, K. Wang, and Z. Chen, "SentiXRL: An advanced large language Model Framework for Multilingual Fine-Grained Emotion Classification in Complex Text Environment," 2024. [Online]. Available: https://arxiv.org/abs/2411.18162

[19] P. M. Gavali and S. K. Shiragave, "Text Representation for Sentiment Analysis: From Static to Dynamic," in *2023 3rd International Conference on Smart Data Intelligence (ICSMDI)*, IEEE, Mar. 2023, pp. 99–105. doi: 10.1109/ICSMDI57622.2023.00025.

[20] N. Alamsyah and N. Rijati, "Fine-Grained Sentiment Classification of Public Opinion on Electric Cars in Indonesia Using IndoBERT." *2024 International Seminar on Application for Technology of Information and Communication (iSemantic),* pp. 502-508, 2024, doi: 10.1109/iSemantic63362.2024.10762277

[21] 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.

[22] 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.

[23] M. Raees and S. Fazilat, "Lexicon-Based Sentiment Analysis on Text Polarities with Evaluation of Classification Models," Sep. 2024.

[24] A. Sathya and Dr. M.S Mythili, "Evaluating Sentiment Classification to Specify Polarity by Lexicon-Based and Machine Learning Approaches for COVID-19 Twitter Data Sets," *JOURNAL OF ADVANCED APPLIED SCIENTIFIC RESEARCH*, vol. 5, no. 4, pp. 12–27, Jul. 2023, doi: 10.46947/joaasr542023678.

[25] S. Consoli, L. Barbaglia, and S. Manzan, "Fine-grained, aspect-based sentiment analysis on economic and financial lexicon," *Knowl Based Syst*, vol. 247, p. 108781, Jul. 2022, doi: 10.1016/j.knosys.2022.108781.

[26] B. Y. Ziwei and H. N. Chua, "A Depression Diagnostic System using Lexicon-based Text Sentiment Analysis," 2022. [Online]. Available: https://www.who.int/teams/mental-health-and-substance-use/suicide-data