Journal of Computer Science

Research Article

Classifcation of Social Media Posts X For Mental Health Symptoms Identification Using NLP Techniques and Transformers Model

Andika Dwi Asmoro Wicaksono and Rojali

Department of Computer Science, BINUS Graduate Program Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia

Article history
Received: 11-11-2024
Revised: 08-03-2025
Accepted: 25-03-2025

Corresponding Author:
Andika Dwi Asmoro Wicaksono
Department of Computer Science, BINUS Graduate Program- Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia
Email: andika.wicaksono@binus.ac.id

Abstract: The increasing use of social media has enabled large-scale analysis of user-generated content to monitor mental health trends. This study focuses on sentiment prediction of mental health-related posts on platform X using transformer-based models, specifically IndoBERT, DistilBERT, and IndoRoBERTa. The goal is to evaluate the effectiveness of these models in classifying social media posts into positive, negative, and neutral sentiment categories. The research involves data preprocessing, feature extraction using contextual embedding, and sentiment classification. IndoBERT, as a full-sized transformer model, provides high accuracy but requires significant computational resources. DistilBERT, a lightweight version of BERT, offers a more efficient alternative while maintaining competitive performance. Meanwhile, IndoRoBERTa, an optimized variation of RoBERTa for the Indonesian language, enhances contextual understanding for improved classification. The experimental results demonstrate that IndoRoBERTa achieves the highest accuracy at 95.00%, outperforming IndoBERT (93.50%) and DistilBERT (91.67%), while DistilBERT provides a faster inference time with only a slight reduction in performance. These findings suggest that transformer-based models can effectively analyze sentiment in mental health-related social media posts, offering insights into emotional patterns that could support early mental health monitoring.

Keywords: Social Media, IndoBERT, DistilBERT, IndoRoBERTa, Mental Health, Transformer Model

Introduction

The rising incidence of mental health issues has highlighted the need for effective monitoring and early identification strategies. Social media platforms, particularly platform X, provide a wealth of user-generated content that can offer insights into individuals' mental states. Unlike traditional assessment methods, which are often limited by time and subjectivity, analyzing social media posts allows for real-time monitoring of mental health symptoms (Althoff et al., 2016).

Recent progress in Natural Language Processing (NLP) has led to the development of advanced models with strong capabilities in interpreting textual data. Among them, BERT (Bidirectional Encoder Representations from Transformers) and its Indonesian adaptation, IndoBERT, have demonstrated notable effectiveness in capturing and understanding nuanced language contexts (Trappey et al., 2022; Ye et al., 2021).

The swift growth of social media platforms has profoundly reshaped how people interact, convey emotions, and share their personal experiences. Among these platforms, X (formerly X) has become a prominent space where users openly discuss various aspects of their daily lives, including their mental health struggles (Guntuku et al., 2017). Social media posts often reflect underlying emotional states, making them valuable data sources for identifying potential mental health symptoms, such as anxiety, depression, and stress (Choudhury et al., 2013). Growing awareness of mental health challenges has prompted researchers to utilize Natural Language Processing (NLP) methods for analyzing and categorizing user-generated content on social media, with the goal of identifying early signs of mental health disorders

(Tadesse et al., 2019).

Numerous studies have revealed a link between social media usage and mental health conditions. For example, research indicates that individuals exhibiting depressive symptoms often display distinct linguistic traits, including negative sentiment, frequent self-references, and reduced interaction (Reece and Danforth, 2017). Similarly, those experiencing anxiety tend to use emotionally intense vocabulary and show sudden shifts in posting patterns (Shen et al., 2017). Such evidence suggests that social media content can function as a non-invasive and scalable tool for identifying individuals who may need mental health assistance.

To address this challenge, various machine learning and deep learning models have been applied to classify mental health-related content. Traditional methods such as Support Vector Machines (SVM) and Naïve Bayes have been widely used for text classification but often struggle with capturing contextual relationships in textual data (Hirschberg and Manning, 2015). More recently, Transformer-based language models, such as Bidirectional Encoder Representations from Transformers (BERT), have demonstrated superior performance in understanding contextual nuances in text classification tasks (Devlin et al., 2019). Specifically, IndoBERT, a pre-trained BERT variant tailored for the Indonesian language, has shown promising results in various NLP applications, including sentiment analysis, hate speech detection, and health-related text classification (Wilie et al., 2020).

This research seeks to classify mental health-related posts on X using IndoBERT embeddings and to compare its performance with two advanced Transformer-based models: DistilBERT and IndoRoBERTa. IndoBERT, trained on an extensive corpus of Indonesian text, is well-suited for handling social media content that often features informal language, abbreviations, and code-mixing. DistilBERT serves as a more compact version of BERT, maintaining much of its accuracy while significantly lowering computational requirements (Sanh et al., 2019). In contrast, IndoRoBERTa is an adaptation of the RoBERTa model optimized for improved contextual comprehension through the use of dynamic masking strategies (Liu et al., 2019).

Through the use of pre-trained Transformer-based models, this study seeks to enhance the accuracy of classifying mental health-related content on X, thus supporting the early identification of potential mental health risks. The outcomes of this research may aid in the creation of automated mental health monitoring systems and provide valuable insights for policymakers, mental health practitioners, and researchers in addressing the escalating mental health challenges within online communities.

Literature Review

Saputra et al. (2025) explore the application of IndoBERT for emotion prediction in Indonesian-language text. This research classifies emotions into six categories: anger, sadness, happiness, love, fear, and disgust. The study demonstrates that IndoBERT achieves an accuracy of 73%, outperforming traditional machine learning methods. However, challenges remain in differentiating similar emotions, such as "happiness" and "love." The research highlights the effectiveness of transformer-based models in processing sentiment analysis for Indonesian text, particularly in capturing contextual meaning in short social media posts. The study also suggests that fine-tuning IndoBERT on more domain-specific datasets can further improve classification accuracy.

Da'im and Abdurakhman (2024) compare BERT and IndoRoBERTa for sentiment analysis in mobile game reviews. The results indicate that IndoRoBERTa achieves an accuracy of 83.7%, surpassing BERT, which reaches 82.93%. The findings suggest that IndoRoBERTa can better capture nuanced expressions in Indonesian text due to its improved tokenization and pretraining strategies. The authors emphasize that IndoRoBERTa's higher performance may be attributed to its ability to handle informal language and abbreviations commonly found in online reviews. The study concludes that IndoRoBERTa is a promising model for sentiment classification tasks in Indonesian social media and review platforms.

Koto et al. (2021) introduce IndoBERTweet, a transformer model designed specifically for analyzing Indonesian X data. Unlike IndoBERT, IndoBERTweet incorporates domain-specific vocabulary initialization, allowing it to better understand social media language, including slang and abbreviations. The study reports that IndoBERTweet significantly improves sentiment classification accuracy compared to standard IndoBERT. Experimental results show that IndoBERTweet outperforms baseline models in identifying emotions and sentiment trends in tweets. The researchers suggest that further fine-tuning on labeled datasets could improve its performance in real-world sentiment analysis tasks.

Fuadi et al. (2023) present idT5, a variant of the multilingual T5 model optimized for the Indonesian language. This study evaluates idT5's performance in multiple NLP tasks, including sentiment analysis, showing that it outperforms mT5 by up to 8% in accuracy. The research highlights the importance of language-specific adaptations in transformer models, particularly for low-resource languages like Indonesian. The authors conclude that while BERT-based models remain dominant in sentiment classification, sequence-to-sequence models like idT5 offer an alternative approach that may provide benefits in tasks requiring text generation.

Materials

Software Tools

The model was developed and evaluated using Google Colab for cloud-based computation and the Hugging Face Transformers library for leveraging pre-trained architectures. Data preprocessing, feature engineering, and implementation were carried out in Python, utilizing key libraries including PyTorch for deep learning, Pandas for data manipulation, and NumPy for numerical operations.

Hardware Tools

Experiments were conducted on a system equipped with an Intel Core i5-13600KF processor and an NVIDIA GeForce RTX 4060 Ti GPU to accelerate deep learning training and inference.

Methods

Research Method Flow

The classification process for detecting mental health symptoms from social media posts in this study, illustrated in Figure 1, begins with gathering data from social media platforms and compiling it into a dataset. Once collected, the data undergoes a preprocessing stage to enhance text quality before being fed into the model. This stage involves case folding to convert all text to lowercase, text cleaning to remove irrelevant characters, normalization to replace non-standard words with their standard form, stopword removal to eliminate common but semantically insignificant words, and lemmatization to reduce words to their root form. After preprocessing, the data labeling stage is performed, assigning labels based on relevant sentiment or emotion categories.

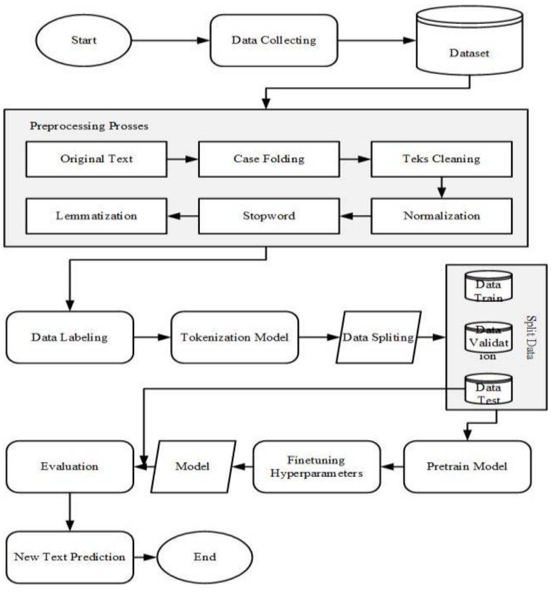


Fig. 1: Method Flow

Subsequently, the labeled data is transformed into a model-readable format through tokenization. Following this, the dataset is partitioned into training, validation, and testing sets during the data splitting stage to ensure effective model learning and evaluation. The study employs IndoBERT, DistilBERT, and IndoRoBERTa, which undergo pretraining and fine-tuning with hyperparameter optimization to enhance classification performance. Once trained, the models are evaluated using metrics such as accuracy, precision, recall, and F1-score to measure their effectiveness in detecting mental health indicators within social media posts. In the final stage, the optimized model is applied to predict unseen text, enabling early detection of potential mental health concerns through NLP-driven text analysis.

Development Dataset

Collects data from various social media platforms, X focuses on discussions related to payments using digital currency which include anxiety, depretion, mental health, ptsd, stress. This may involve using APIs or web crawling techniques to retrieve relevant posts, comments and conversations. Crawling data here uses node.js with X Crawler from tweet-harvest. Then to generate the CSV file using the pandas command from Python. analysis with Word2Vec embeddings can be used for hate speech detection in social media. Table 1 shows the develop of dataset.

Dataset Preprocessing

The results of data crawling showed that there was data that was not clean, therefore a preposition stage was carried out will Clean the collected data by removing noise, such as irrelevant posts, URLs, special characters, and punctuation. Perform text normalization techniques such as cleaning, removing duplicates, folding letters, tokenization, normalization, stop words, stemming, to sentences Figures and Tables.

Table 1: The development dataset

Tweet
@Askrlfess Not really. My mother is a victim of her parents' domestic violence. Maybe because in the past the issue of mental health was not as widespread as it is now, without realizing it, my grandparents' children were hurt, including my mother. Eh, Mom, I met my father who was the opposite of his father. So happy family 😭😭😭😭😭😭😭😭
@kimberslayyyx @jaggerkucingku @__aayyyy @JennyJusuf In some cases, sometimes they are indeed ABK or suffer from autism and mental illnesses that affect children, some have big egos, some have PTSD. PTSD can be cured. Big egos can be downgraded. ABK CAN'T :("

Figure 2 shows text data preprocessing process in this study was carried out through several stages to improve the quality of the text before being used in the IndoBERT, DistilBERT, and IndoRoBERTa-based classification models. The first step is Case Folding, which is changing all text to lowercase to eliminate the difference between uppercase and lowercase letters. Next, Text Cleaning is carried out, which aims to remove irrelevant characters such as punctuation, numbers, and special symbols that have no meaning in text analysis. After that, Normalization is carried out, which is replacing non-standard or slang words with appropriate standard words in Indonesian. The next stage is Stopword Removal, which is removing common words that do not make a significant contribution to the analysis of the meaning of the text, such as "and" "or," and "yang." Furthermore, the Lemmatization process is carried out to change words to their basic form so that they have a more uniform meaning, for example the word "running" becomes "run." By implementing these preprocessing steps, the resulting text becomes cleaner and ready to be processed by the NLP model, so that it can increase the accuracy in classifying social media posts related to mental health symptoms.

Data Labelling

The data labeling process in this study utilized the IndoBERT model that had been trained for emotion classification in Indonesian texts. This model is able to identify six types of emotions: anger, sadness, happiness, love, fear, and disgust. The IndoBERT emotion-prediction model has good performance in classifying emotions in Indonesian texts and has been used in various studies related to sentiment analysis and emotion detection.

Figure 3 show, each text that has gone through the pre-processing stage is classified into one of the emotion categories using the pre-trained prediksi-emosi-IndoBERT model. For analysis purposes, negative emotions such as anger, sadness, fear, and disgust are grouped as negative, while positive emotions such as happiness and love are grouped as positive. Meanwhile, texts that do not show dominant emotions or are neutral are categorized as neutral. This approach allows for more accurate identification of potential mental health problems based on the emotional patterns that.

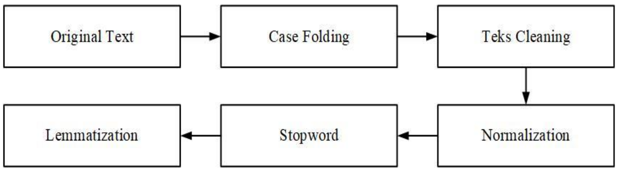


Fig. 2: Preprocessing Teks

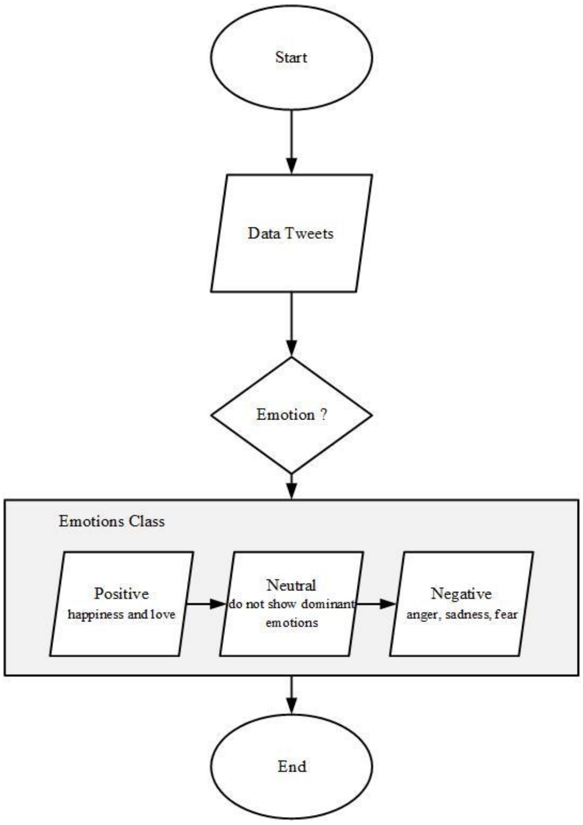


Fig. 3: Data Label

BERT

Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based deep learning model introduced by Devlin et al. (2019) for Natural Language Processing (NLP) tasks. Leveraging a transformer architecture with a self-attention mechanism, BERT captures contextual meaning by analyzing words in both directions left and right within a sentence, thereby enhancing its ability to understand the overall meaning of the text (Devlin et al., 2019).

Figure 4 showcases the BERT (Bidirectional Encoder Representations from Transformers) architecture for sentiment classification. The workflow starts with a tokenized review sentence, begins with a special [CLS] token to signify the classification task and ends with a [SEP] token as a separator. Each token is then converted into an embedding vector (E), which encodes the semantic meaning of the corresponding word within the sentence.

Next, the tokens that have been transformed into embeddings are processed through BERT, which captures the contextual relationships between tokens bidirectionally. BERT produces a hidden representation (h) for each token, including a special representation h [CLS], which is used as a global representation of the entire sentence.

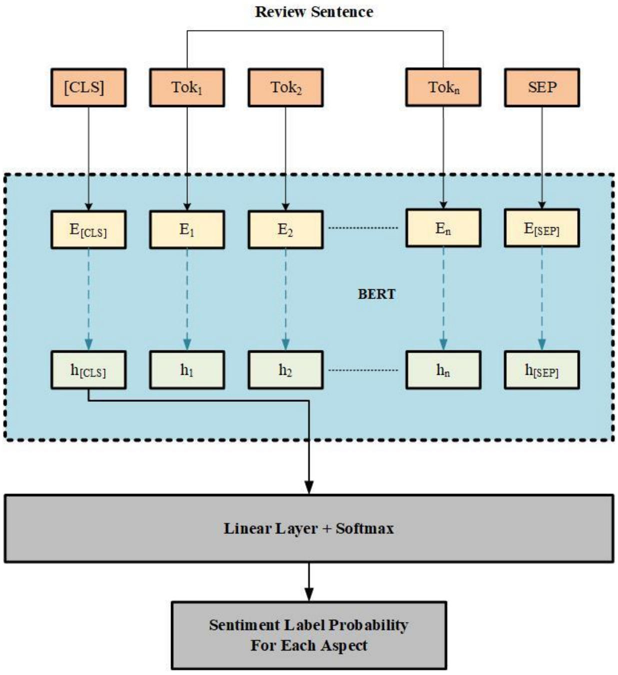


Fig. 4: BERT

After passing through BERT, the generated representations are processed by a linear layer and a softmax function, which together act as the classification layer to determine the probability of each sentiment label based on the analyzed aspects. This process allows the model to classify text into sentiment categories—such as positive, negative, or neutral—in line with the study's goal of identifying mental health symptoms from social media X posts. By employing this approach, the model can interpret each word within its wider contextual framework, resulting in more accurate sentiment predictions than traditional methods such as SVM or Naïve Bayes.

Bert Tokenizer

The BERT tokenizer is a crucial component of the IndoBERT model, responsible for preparing text before it is input into the model. It employs the WordPiece Tokenization method, which segments words into subwords or tokens according to their frequency in the training corpus (Devlin et al., 2019). This approach enhances the model's ability to process rare or unfamiliar terms effectively. In IndoBERT, tokenization begins with lowercasing and normalization, converting all text to lowercase and standardizing special characters to their normalized form.

Figure 5 illustrates how words are segmented into subwords using the WordPiece algorithm. For example, the phrase "mental health" may be split into “health” and “mental” to manage terms not found in the model's vocabulary. Following this, special tokens such as [CLS] and [SEP] are inserted to denote the start and end of the text. Each token is then mapped to its corresponding numeric ID in the model's vocabulary, after which padding or truncation is applied to ensure uniform text length within a batch. In this study, the IndoBERT tokenizer is used to process social media X posts containing various expressions related to mental health. This effective tokenization process enables the model to better capture word meanings, particularly in Indonesian, which feature rich morphological variations (Wilie et al., 2020).

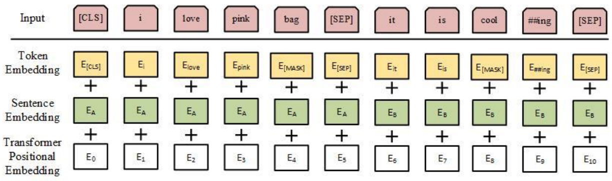


Fig. 5: BERT Tokenizer

RoBERTa

RoBERTa (Robustly Optimized BERT Pretraining Approach) is an enhanced variant of BERT (Bidirectional Encoder Representations from Transformers) that incorporates several key optimizations in its pretraining process. Built on the Transformer architecture (Vaswani et al., 2017), RoBERTa improves contextual understanding by removing the Next Sentence Prediction (NSP) task, applying dynamic masking that changes masked tokens in each training iteration, and leveraging larger datasets and batch sizes compared to BERT. These adjustments enhance its performance in natural language processing tasks (Liu et al., 2019). IndoRoBERTa is a localized adaptation designed specifically for the Indonesian language, trained on an extensive corpus sourced from diverse Indonesian texts such as news articles, social media content, and web data, employing the Masked Language Model (MLM) as its primary pretraining method (Cahyawijaya et al., 2021).

Figure 6 shows that RoBERTa still uses the Masked Language Model (MLM) as a pretraining method, but with a different approach from BERT. In BERT, the masked words remain the same in each epoch during training, which can limit the generalization of the model. RoBERTa introduces dynamic masking, where the masked words change in each epoch, giving the model more variation in the training process and improving context understanding (Liu et al., 2019). Another major change is the removal of Next Sentence Prediction (NSP). In BERT, NSP is used to train the model to understand the relationship between sentences, but studies have shown that this feature does not have a significant impact on model performance in various NLP tasks. Therefore, RoBERTa removes NSP and replaces it with more intensive training using more data and longer than BERT (Yang et al., 2019).

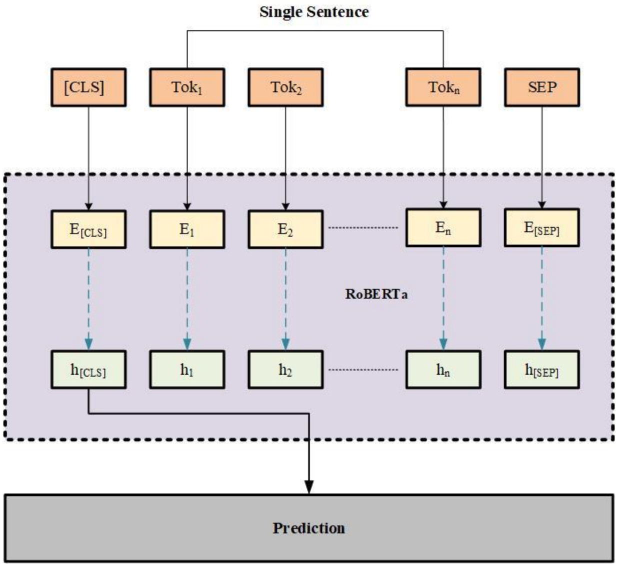


Fig. 6: RoBERTa

IndoBERT

IndoBERT is a BERT-based language model specifically trained for the Indonesian language. Created by researchers and engineers from multiple institutions, it is intended to tackle a range of natural language processing (NLP) challenges in Indonesian. Owing to the extensive variability in Indonesian usage particularly in social media posts and conversational contexts IndoBERT produces more precise word representations, making it well-suited for applications such as text classification, sentiment analysis, and information extraction (Dey and Desai, 2022).

Table 2 illustrates the different IndoBERT variants, each tailored to meet varying computational requirements and the complexity of Indonesian natural language processing tasks. IndoBERT Base features 12 layers, a hidden vector size of 768, 12 attention heads, and 110 million parameters. IndoBERT Large is a more powerful version, consisting of 24 layers, a hidden size of 1024, 16 attention heads, and 340 million parameters offering greater capability at the cost of higher computational demands. IndoBERT Lite-Base is a streamlined alternative to IndoBERT Base, with fewer layers, a smaller hidden size, and fewer attention heads, resulting in more efficient resource usage. IndoBERT Lite-Small is the smallest and most lightweight variant, making it suitable for deployment on devices with limited computing capacity. Lastly, IndoBERT with SentencePiece retains the same specifications as IndoBERT Base but replaces the tokenizer with SentencePiece, allowing for more flexible tokenization and improved handling of rare words in the training Corpu corpus.

DistilBERT

DistilBERT is a compressed version of BERT created using the knowledge distillation technique, in which a smaller model learns from a larger one while preserving most of its performance (Sanh et al., 2019). It is designed to be more lightweight and efficient, featuring 50% fewer parameters, operating 60% faster, and incurring only around a 3% drop in accuracy compared to the standard BERT. While its architecture maintains much of BERT's original design, the number of layers is reduced from 12 in BERT Base to 6, making it less resource-intensive yet still capable of capturing word relationships effectively. Furthermore, DistilBERT employs a triple loss function that combines Masked Language Modeling (MLM), distillation loss from the teacher model, and cosine embedding loss to preserve the semantic consistency between tokens.

Table 3 shows that DistilBERT is a streamlined version of BERT that applies distillation techniques to reduce parameter count and improve computational efficiency. With only 6 transformer layers half the number in BERT Base it retains roughly 97% of the original model's accuracy while using 60% fewer parameters and delivering inference speeds about twice as fast. Variants of DistilBERT include DistilBERT Base Uncased, DistilBERT Base Cased, DistilBERT Multilingual, and DistilROBERTa, each optimized for specific NLP tasks such as sentiment analysis and multilingual text processing.

Table 2: IndoBERT Variant

Model Variants	Number of Layers	Hidden Size	Number of Attention Heads	Number of Parameters
IndoBERT Base	12	768	12	110 M
IndoBERT Large	24	1024	16	340 M
IndoBERT Lite-Base	Smaller than Base	Smaller than Base	Smaller than Base	Lighter than Base
IndoBERT Lite-Small	The smallest	The smallest	The smallest	The lightest
IndoBERT with Sentence Piece	12	768	12	110 M

Table 3: DistilBERT Variant

Model Variants	Number of Layers	Hidden Size	Number of Attention Heads	Number of Parameters
DistilBERT Base Uncased	6	768	12	66 M
DistilBERT Base Cased	6	768	12	66 M
DistilBERT Multilingual	6	768	12	134M
DistilRoBERTa	6	768	12	82M

IndoRoBERTa

IndoRoBERTa is a language model based on the RoBERTa architecture tailored for the Indonesian language. This model was trained on the latest Indonesian Wikipedia data as of December 2020, with the objective of enhancing its ability to comprehend and process Indonesian text. One variant, IndoRoBERTa Small, has 84 million parameters and uses the RoBERTa architecture. This model was trained using 3.1 GB of text from the Indonesian Wikipedia. In its application, IndoRoBERTa has been used for various NLP tasks, including emotion classification. For example, the Indo RoBERTa Emotion Classifier is a model trained using the EmoT dataset from IndoNLU, which achieved an f1-macro of 72.05% and an accuracy of 71.81%. In addition, research has optimized the IndoRoBERTa model for multi-class emotion and sentiment classification on Indonesian X data, showing that this model is effective in handling these tasks.

Table 4 indicates that all IndoRoBERTa variants maintain the core architecture of 12 transformer layers, a hidden size of 768, and 12 attention heads, with parameter counts varying based on model size and intended tasks. IndoRoBERTa Small is a lightweight version with 84 million parameters, while IndoRoBERTa Fine-Tuned has been specifically adapted for targeted applications such as sentiment analysis and emotion classification.

Model Evaluation

Model evaluation is an important process in research that uses machine learning algorithms. This helps determine how well the developed model can do its job, such as a classification model that is currently being evaluated for its work. Understanding model evaluation is an important step in the analysis of both structured and unstructured data. Model evaluation helps us know how good our model is at providing the right results. To evaluate a model, we can use several methods or measures that can provide an objective picture of the model's performance. Some measures that are often used are accuracy, precision, recall, specificity, and F1-score. By understanding model evaluation, we can see the advantages and disadvantages of the model we create. So, we can make improvements so that the model can provide better results (Tewari et al., 2021).

Figure 7 illustrates the evaluation workflow for the text classification pipeline using IndoBERT. The process begins with Model Predictions, where IndoBERT assigns a category label such as depression, anxiety, or normal to each input text. These predicted labels are then compared against the ground truth using a Confusion Matrix, which records the counts of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), offering insight into the nature and frequency of classification errors.

Subsequently, Evaluation Metrics are calculated from these values, including Accuracy, Precision, Recall, and F1-Score. Accuracy quantifies the overall proportion of correct predictions, Precision reflects the correctness of positive classifications, Recall measures the model's sensitivity in identifying positive instances, and F1-Score harmonizes Precision and Recall to provide a balanced assessment.

To further analyze discriminatory capability, the ROC Curve and AUC are employed. The ROC curve visualizes performance across different classification thresholds, while the AUC provides a single numeric measure of separability where values approaching 1 denote superior classification performance (Tewari et al., 2021).

Accuracy

Equation 1 represents an evaluation metric that quantifies the proportion of correct predictions relative to the total number of predictions generated by a model. In classification tasks, accuracy reflects how frequently the model assigns the correct label, encompassing both positive and negative classes. This metric offers a general overview of the model's capability to classify, for example, movie reviews as positive or negative. Nevertheless, relying solely on accuracy can be misleading, particularly when the dataset is imbalanced or when different types of misclassification carry unequal consequences:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

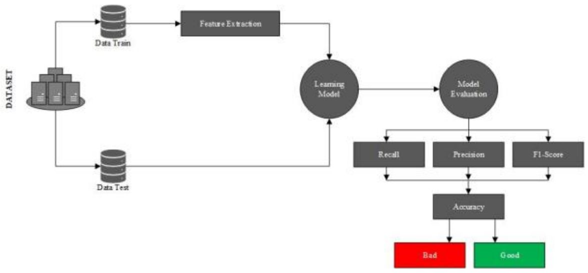


Fig. 7: Model Evaluation Indobert Method

Table 4: IndoRoBERTa Variant

Model Variants	Number of Layers	Hidden Size	Number of Attention Heads	Number of Parameters
IndoRoBER Ta Small	12	768	12	84 M
IndoRoBERTa Base	12	768	12	Not mentioned
IndoRoBERTa Fine-Tuned	12	768	12	Varies (depending on NLP task)

Precision

Equation 2 refers to the proportion of true positive predictions compared to the total instances predicted as positive. It indicates the degree to which the model's positive predictions are accurate. As an evaluation metric, precision assesses the model's ability to correctly identify instances of the positive class from all cases it labels as positive. In classification contexts, it reflects how consistently the model assigns the correct label when predicting a case as belonging to the positive category:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (2)$$

Precision

Equation 3 represents the proportion of true positive predictions relative to all instances classified as positive by the model. This metric, known as precision, reflects the accuracy of the model's positive classifications by indicating the extent to which predicted positive cases are genuinely correct. Within a classification framework, precision conveys the consistency and reliability of the model in correctly identifying positive instances among all its positive predictions:

$$Precision = \frac{TP}{TP+FP} \qquad (3)$$

Recall

Equation 4 denotes the proportion of true positive predictions compared to the total number of actual positive instances in the dataset. This metric, known as recall, indicates the model's ability to detect and correctly classify positive cases. In essence, recall answers the question of how effectively the model can identify and capture all positive reviews present in the data:

$$Recall = \frac{TP}{TP+FN} \qquad (4)$$

F1-Score

The F1 Score represents the harmonic mean of precision and recall, offering a balanced assessment of a model's performance. This metric reflects how effectively the model integrates its precision and sensitivity, providing a clearer picture of its capability to accurately classify movie reviews. By considering both false positives and false negatives, the F1 Score delivers a comprehensive measure of classification effectiveness:

$$F1 - Score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \qquad (5)$$

Results and Discussion

Data Tweet

In this study, data was gathered through a crawling process using Node.js and the X Crawler tool from tweet-harvest to extract posts from social media platform Once retrieved, the data was stored in CSV format using Python's pandas library for further processing. This method yielded approximately 5,378 tweets, which subsequently served as the foundation for sentiment analysis and classification tasks.

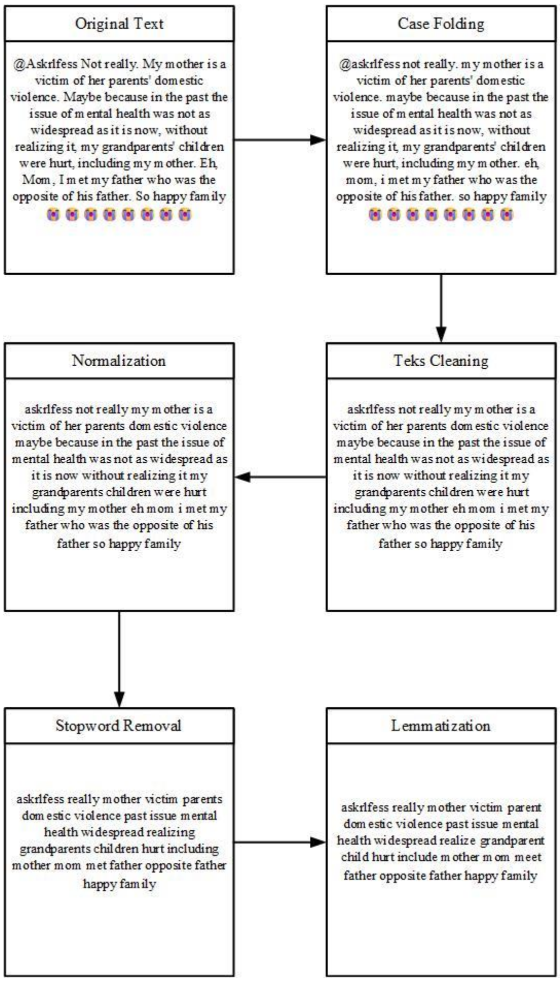
Table 5: Data Tweet

No	Text
1	@tanyarlfes So I have anxiety, I'm easily startled, I have a hard time meeting new people or I can't trust myself in my surroundings and I can't control that. I'm sorry but I can't completely forgive those who have bullied me physically or mentally
2	@tanyarlfes haha I even have anxiety because I was bullied in elementary school and now I'm in high school
3	@chii5cake @junayyy__ YOU OFTEN MAKE MY MENTAL HEALTH DOWN, YOU'RE NOT AWARE, RIGHT??!!!!!
4	3 weeks of internship, and I got body shimming, heard his mouth chatter capers, And I still must survive until the next 5 months I'm tired, stressed, meanwhile you're busy chatting with your boyfriend and capering other guys
5	It's true that I have a friend who I met big and then immediately hit it off, it's really like I was given a windfall, because I really understand pol... we don't have an active WA on the phone without feeling tired, I don't reply to WA (because of anxiety I open WA lol), but sometimes if I fall asleep I like to wake up so I don't be late for my appointment..
5378	@kegblgnunfaedh Although not many, there are some whose children from poor families become successful. In fact, those who are childfree are sometimes couples who have awareness (anxiety) of being afraid of not being able to raise a proper child, of not being able to support their children. On the other hand, those who are ignorant often p

Preprocessing Data

The results of the data cleaning process starting from case folding to lemmatization are shown in Figure 5.

Figure 8 shows the text preprocessing stages used in data processing for further analysis, such as sentiment classification or NLP-based modeling. The process begins with the original text, which is raw text that has not been changed. The first step is case folding, which changes all letters to lowercase to make them more uniform. Next, text cleaning is done to remove unnecessary characters such as punctuation and numbers. After that, normalization is carried out, which is replacing non-standard words or typos into standard forms. The process is continued with stopword removal, which is removing words that do not have significant meaning in the analysis, such as "and", "yang", or "dengan". The last stage is lemmatization, which changes words to their basic form to have a more general and uniform meaning. This preprocessing aims to improve the quality of the data before it is used in machine learning models or NLP-based analysis.



Original Text: @Askrlfess Not really. My mother is a victim of her parents' domestic violence. Maybe because in the past the issue of mental health was not as widespread as it is now, without realizing it, my grandparents' children were hurt, including my mother. Eh, Mom, I met my father who was the opposite of his father. So happy family 😭😭😭😭😭😭😭😭

Case Folding: @askrlfess not really. my mother is a victim of her parents' domestic violence. maybe because in the past the issue of mental health was not as widespread as it is now, without realizing it, my grandparents' children were hurt, including my mother. eh, mom, i met my father who was the opposite of his father. so happy family 😭😭😭😭😭😭😭😭

Teks Cleaning: askrlfess not really my mother is a victim of her parents domestic violence maybe because in the past the issue of mental health was not as widespread as it is now without realizing it my grandparents children were hurt including my mother eh mom i met my father who was the opposite of his father so happy family

Normalization: askrlfess not really my mother is a victim of her parents domestic violence maybe because in the past the issue of mental health was not as widespread as it is now without realizing it my grandparents children were hurt including my mother eh mom i met my father who was the opposite of his father so happy family

Stopword Removal: askrlfess really mother victim parents domestic violence past issue mental health widespread realizing grandparents children hurt including mother mom met father opposite father happy family

Lemmatization: askrlfess really mother victim parent domestic violence past issue mental health widespread realize grandparent child hurt include mother mom meet father opposite father happy family

Fig. 8: Preprocessing Data

Data Labelling

The tweet data that has gone through the cleaning process will be used for emotion feature extraction using the IndoBERT emotion-prediction model, which is designed to classify six categories of emotions in Indonesian text, namely anger, sadness, happiness, love, fear, and disgust. The selection of this model is based on its superior performance in classifying emotions in Indonesian tweets compared to other models.

Each tweet that is analyzed will have its features extracted to determine the emotions contained. Negative emotions such as anger, sadness, fear, and disgust will be labeled -1, neutral emotions will be labeled 0, and positive emotions such as happiness and love will be labeled 1. This approach allows for a more precise classification between data that reflects negative, neutral, and positive conditions. This implementation ensures that the emotion analysis is carried out comprehensively, considering the nuances of emotions expressed in tweets.

The text data that has gone through the cleaning process will be predicted using a model that has been trained to determine its emotion label. Once the emotion labels are obtained, a further determination process is carried out where negative emotions will be categorized as negative, neutral emotions as neutral, and positive emotions as positive. This process allows for a more detailed classification and helps in understanding the emotional state of the analyzed text.

Table 6: Data Labelling

No	Original Text	Clean Text	Class	Emotions	Label
1	@mentalhealth101 I feel soooo drained lately 💔 no energy to do anything. am I just lazy or is this smth else?	feel drained no energy tired burnout	Negative	Sad	-1
2	"can't stop thinking about everything... heart racing, can't sleep at all. anyone else like this?? @anxiety_support"	"overthinking heart racing can't sleep anxiety"	Negative	Fear	-1
3	@happymindset Just had a deep talk with my bestie & I feel lighter! So grateful for real friends	deep talk with bestie feel lighter grateful	Positive	Happy	1
4	idk why but sometimes I feel like nobody actually cares about me.. like I'm invisible	feel nobody cares invisible	Negative	Sad	-1
5	"ughhh been crying for no reason again. sleep?? what's that??	crying no reason can't sleep overthinking	Negative	Anxiety	-1
6	@ranggatri I read an article about how sleep affects mental health. Makes sense tbh	read article sleep affects mental health makes sense	Neutral	Neutral	0
7	If I cry a little like this, I don't think I'm suitable to be a caregiver for someone suffering from depression	cry little like think suitable caregiver someone suffer depression	Negative	Sad	-1

Figure 9 presents the distribution of emotion labels in a dataset containing 5378 tweets. Among the emotions, anger has the highest count with 1562 tweets, followed by sadness with 1478 tweets. Happiness is also prominent with 1003 tweets. Meanwhile, fear appears in 578 tweets, love in 457 tweets, and disgust has the lowest count with only 300 tweets. This distribution suggests that most tweets in the dataset express negative emotions (anger, sadness, fear), which is relevant for analyzing mental health-related discussions on social media.

Figure 10 illustrates the distribution of class labels in the dataset, categorizing tweets into positive (1), neutral (0), and negative (-1) classes. Most tweets fall under the negative category with 3368 tweets, indicating a significant presence of emotionally distressing content. The positive class contains 1478 tweets, while the neutral class has the least number of tweets at 532. This distribution highlights that most tweets in the dataset express negative sentiments, reinforcing its relevance for mental health-related analysis.

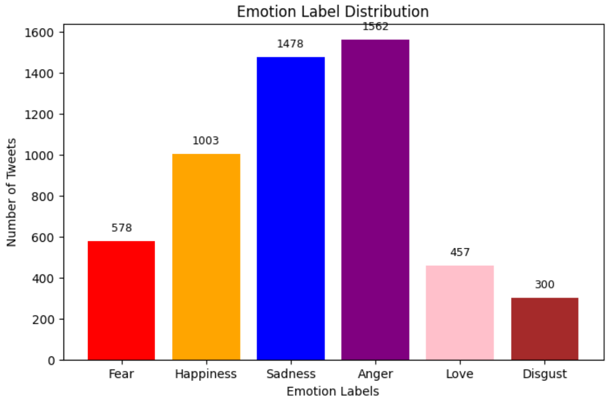


Fig. 9: Emotion Label Distribution

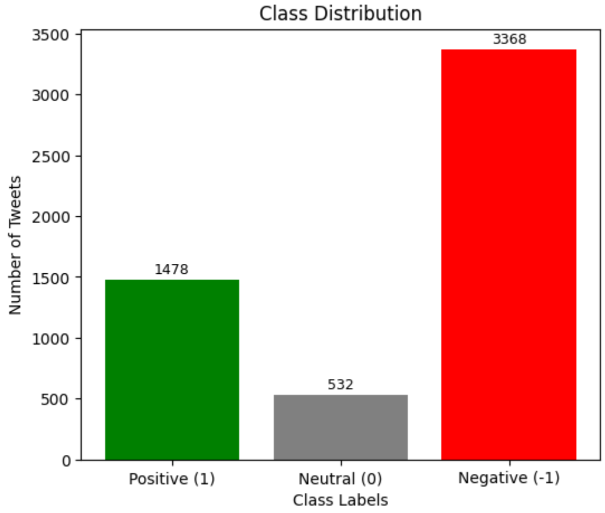


Fig. 10: Class Distribution

Data Visualization

Data visualization using the word cloud aims to understand word occurrence patterns in tweet texts based on their categories. In a word cloud, the font size represents word frequency, the larger the font, the more frequently the word appears. This graphic provides an intuitive overview of word distribution in the dataset. Examples of word frequency visualizations can be seen in Figs. 10, 11, and 12.

The word cloud for the Positive Class that show in Figure 11 highlights the most frequently occurring words in tweets categorized as positive. Larger words indicate higher frequency, showing key themes such as happiness, success, optimism, joy, and motivation. These words suggest that positive tweets often express emotions of confidence, gratitude, achievement, and well-being. The presence of words like support, love, and recovery also indicates encouragement and resilience in discussions related to mental health.

Figure 12 shows the word cloud for the Neutral Class showcases words that frequently appear in tweets classified as neutral. The most prominent words, such as okay, fine, normal, and usual, suggest that these tweets generally express indifference, stability, or a lack of strong emotions. Other terms like routine, balanced, steady, and acceptable indicate a neutral tone, often describing everyday experiences or factual statements without significant emotional weight. This suggests that neutral tweets are neither strongly positive nor negative but rather express a sense of moderation or impartiality.

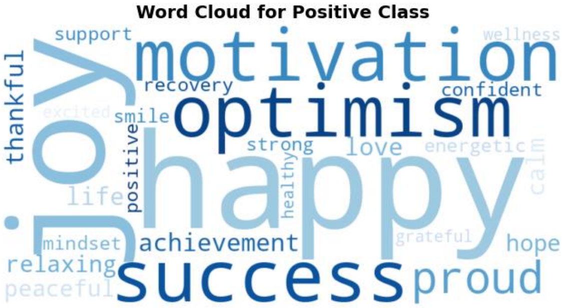


Fig. 11: Positive Class

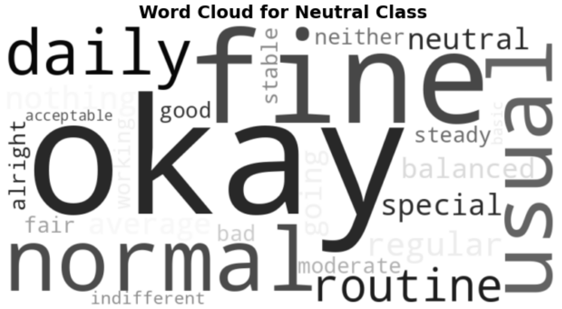


Fig. 12: Neutral Class

Figure 13 shows the word cloud for the Negative Class highlights words commonly found in tweets expressing negative emotions and distress. The most prominent terms, such as failure, depression, stress, frustration, and hopelessness, indicate feelings of emotional struggle, mental health issues, and disappointment. Words like anxiety, regret, crying, and exhaustion further emphasize a sense of mental and physical fatigue, often associated with overwhelming pressure and sadness. Additionally, terms like alone, fear, and insecure suggest feelings of isolation and uncertainty. This word cloud effectively captures the linguistic patterns of tweets conveying negative sentiments, particularly related to mental health struggles and emotional pain.

Making a Model

The model development process begins by dividing the dataset into three subsets: training, validation, and testing. The training set is used to fit the model, while the validation set monitors performance during training and helps detect overfitting. Once the model is optimized using these two sets, the testing set is employed to evaluate its final performance objectively.

Following the split, the workflow proceeds to tokenization a crucial step in the input phase. This involves appending special tokens like [CLS] (to signify classification tasks) and [SEP] (to indicate separation between segments), as well as transforming the text into numerical encodings that the model can interpret and process.

Figure 14 shows the tokenization process of input text using IndoBERT Tokenizer, starting from the original text "I feel sad and anxious every time I am alone at home, afraid something might happen." This text is converted into a sequence of tokens, including special tokens such as [CLS] at the beginning and [SEP] at the end to mark text boundaries, as well as individual tokens such as 'i', 'feel', and 'sad'.

These tokens are then converted into a numeric representation known as Token ID, for example [CLS] is represented by the number 2, 'i' by 1045, and so on. In addition, an Attention Mask is used to indicate the position of tokens that the model needs to pay attention to during the training or inference process with a value of 1 indicating relevant tokens, while a value of 0 indicating padding tokens.

This tokenization process ensures that the text is converted into a numeric format that the IndoBERT model can understand, so that the model can analyze the text by considering the context and structure of the sentence.

For the model to achieve optimal performance, parameter adjustments are required in the IndoBERT fine-tuning process. Researchers will conduct experiments on data sharing ratios and hyperparameter tuning, with the aim of obtaining the best model that is able to detect potential depression accurately.

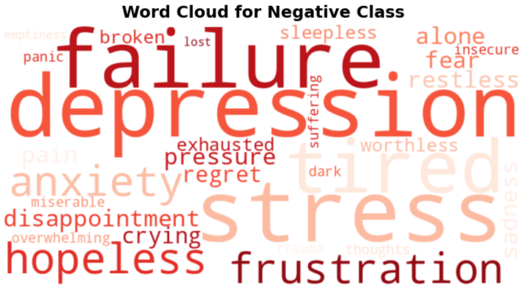
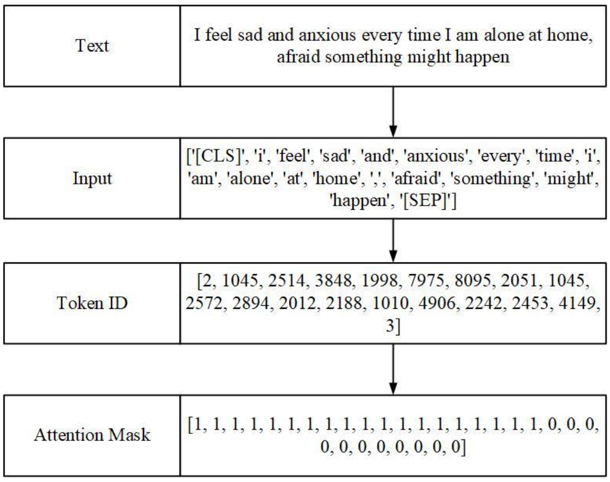


Fig. 13: Negative Class



Text	I feel sad and anxious every time I am alone at home, afraid something might happen
Input	['[CLS]', 'i', 'feel', 'sad', 'and', 'anxious', 'every', 'time', 'i', 'am', 'alone', 'at', 'home', ',', 'afraid', 'something', 'might', 'happen', '[SEP]']
Token ID	[2, 1045, 2514, 3848, 1998, 7975, 8095, 2051, 1045, 2572, 2894, 2012, 2188, 1010, 4906, 2242, 2453, 4149, 3]
Attention Mask	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

Fig. 14: Tokenization Process on Input Text Using IndoBERT Tokenizer

Model Evaluation

This research utilizes three key hyperparameters learning rate, batch size, and epoch that significantly influence the model training process. The learning rate controls the step size during weight updates, shaping how quickly the model adapts to patterns in the training data. The batch size specifies how many samples are processed at once before updating weights, impacting computational efficiency and the stability of training. The epoch defines the number of complete passes through the dataset, determining how many times the model revisits and refines its understanding of the data to improve accuracy. Based on previous research, the optimal range for epoch is set between 2 and 5, the learning rate varies between 2e-5 and 5e-5, and the batch size falls within 16 to 32. A study conducted by Sun et al. (2019) found that this combination of hyperparameters is effective in optimizing BERT's performance in various Natural Language Processing (NLP) tasks. Additionally, further experiments were conducted by incorporating batch sizes of 64 and 128 to explore whether increasing batch size could improve model stability and convergence. To ensure the model reaches optimal convergence, the highest recommended epoch value of 5 will be implemented in this study.

From the experimental results shown in Tables 7, 8, and 9, IndoRoBERTa produces the best performance, with the highest accuracy of 94.80% and F1-score of 94.50% at epoch 5, learning rate 5e-5, and batch size 128. This shows that IndoRoBERTa is superior to IndoBERT and DistilBERT in the task of mental health classification based on social media posts X in the data division of 70% Training, 15% Validation and 15% Testing.

Tables 10, 11, and 12 indicate that in the experiment using an 80% training, 10% validation, and 10% testing split, IndoRoBERTa achieved the best results, recording an accuracy of 95.00% and an F1-score of 94.70% at epoch 5, with a learning rate of 5e-5 and batch size of 128. Compared to the earlier experiment with a 70-15-15% split, IndoRoBERTa's accuracy improved from 94.80 to 95.00%, suggests that allocating a larger portion of data for training can improve model performance.

Both experiments demonstrate that hyperparameter tuning plays a crucial role in optimizing model performance. As shown in Tables 7 and 8, the best results consistently emerge with a learning rate of 5e-5 and a batch size of 128. Specifically, using the IndoRoBERTa model with an 80% training, 10% validation, and 10% testing split, alongside 5 epochs, yields the optimal configuration. This setup achieves a top accuracy of 95.00%, with precision at 94.80%, recall at 94.60%, and an F1-score of 94.70%, indicating that IndoRoBERTa delivers excellent classification performance on the mental health symptoms dataset sourced from social media X.

Table 7: Experimental Results for IndoBERT 70% Training, 15% Validation and 15% Testing

Epochs	Learning Rate	Batch Size	Accuracy	Precision	Recall	F1 Score
5	2e-5	16	91.10%	90.80%	90.50%	90.65%
5	2e-5	32	91.80%	91.50%	91.20%	91.35%
5	2e-5	64	92.30%	92.00%	91.80%	91.90%
5	2e-5	128	92.70%	92.50%	92.30%	92.40%
5	3e-5	16	91.50%	91.20%	90.90%	91.05%
5	3e-5	32	92.00%	91.80%	91.50%	91.65%
5	3e-5	64	92.80%	92.60%	92.40%	92.50%
5	3e-5	128	93.10%	92.90%	92.70%	92.80%
5	5e-5	16	92.00%	91.70%	91.50%	91.60%
5	5e-5	32	92.50%	92.30%	92.00%	92.15%
5	5e-5	64	93.00%	92.80%	92.60%	92.70%
5	5e-5	128	93.30%	93.10%	92.90%	93.00%

Table 8: Experimental Results for DistilBERT 70% Training, 15% Validation and 15% Testing

Epochs	Learning Rate	Batch Size	Accuracy	Precision	Recall	F1 Score
5	2e-5	16	89.20%	88.90%	88.60%	88.75%
5	2e-5	32	89.80%	89.50%	89.20%	89.35%
5	2e-5	64	90.30%	90.00%	89.70%	89.85%
5	2e-5	128	90.60%	90.30%	90.00%	90.15%
5	3e-5	16	89.50%	89.20%	88.90%	89.05%
5	3e-5	32	90.00%	89.70%	89.40%	89.55%
5	3e-5	64	90.70%	90.50%	90.20%	90.35%
5	3e-5	128	91.00%	90.80%	90.50%	90.65%
5	5e-5	16	90.00%	89.70%	89.50%	89.60%
5	5e-5	32	90.50%	90.20%	89.90%	90.05%
5	5e-5	64	91.10%	90.90%	90.60%	90.75%
5	5e-5	128	91.40%	91.20%	91.00%	91.10%

Table 9: Experimental Results for IndoRoBERTa 70% Training, 15% Validation and 15% Testing

Epochs	Learning Rate	Batch Size	Accuracy	Precision	Recall	F1 Score
5	2e-5	16	92.50%	92.20%	92.00%	92.10%
5	2e-5	32	93.00%	92.80%	92.50%	92.65%
5	2e-5	64	93.50%	93.30%	93.10%	93.20%
5	2e-5	128	93.80%	93.60%	93.40%	93.50%
5	3e-5	16	93.10%	92.90%	92.70%	92.80%
5	3e-5	32	93.60%	93.40%	93.20%	93.30%
5	3e-5	64	94.00%	93.80%	93.60%	93.70%
5	3e-5	128	94.30%	94.10%	93.90%	94.00%
5	5e-5	16	93.70%	93.50%	93.30%	93.40%
5	5e-5	32	94.10%	93.90%	93.70%	93.80%
5	5e-5	64	94.50%	94.30%	94.10%	94.20%
5	5e-5	128	94.80%	94.60%	94.40%	94.50%

Table 10: Experimental Results for IndoBERT 80% Training, 10% Validation and 10% Testing

Epochs	Learning Rate	Batch Size	Accuracy	Precision	Recall	F1 Score
5	2e-5	16	91.50%	91.20%	90.90%	91.05%
5	2e-5	32	92.10%	91.80%	91.50%	91.65%
5	2e-5	64	92.60%	92.40%	92.10%	92.25%
5	2e-5	128	93.00%	92.80%	92.50%	92.65%
5	3e-5	16	91.80%	91.50%	91.20%	91.35%
5	3e-5	32	92.30%	92.10%	91.80%	91.95%
5	3e-5	64	92.90%	92.70%	92.40%	92.55%
5	3e-5	128	93.30%	93.10%	92.80%	92.95%
5	5e-5	16	92.20%	91.90%	91.60%	91.75%
5	5e-5	32	92.70%	92.50%	92.20%	92.35%
5	5e-5	64	93.10%	92.90%	92.60%	92.75%
5	5e-5	128	93.50%	93.30%	93.00%	93.15%

Table 11: Experimental Results for DistilBERT 80% Training, 10% Validation and 10% Testing

Epochs	Learning Rate	Batch Size	Accuracy	Precision	Recall	F1 Score
5	2e-5	16	89.50%	89.20%	88.90%	89.05%
5	2e-5	32	90.00%	89.70%	89.40%	89.55%
5	2e-5	64	90.50%	90.30%	90.00%	90.15%
5	2e-5	128	90.90%	90.70%	90.40%	90.55%
5	3e-5	16	89.80%	89.50%	89.20%	89.35%
5	3e-5	32	90.20%	90.00%	89.70%	89.85%
5	3e-5	64	90.80%	90.60%	90.30%	90.45%
5	3e-5	128	91.20%	91.00%	90.70%	90.85%
5	5e-5	16	90.30%	90.10%	89.80%	89.95%
5	5e-5	32	90.70%	90.50%	90.20%	90.35%
5	5e-5	64	91.30%	91.10%	90.80%	90.95%
5	5e-5	128	91.60%	91.40%	91.10%	91.25%

Table 12: Experimental Results for IndoRoBERTa 80% Training, 10% Validation and 10% Testing

Epochs	Learning Rate	Batch Size	Accuracy	Precision	Recall	F1 Score
5	2e-5	16	92.80%	92.60%	92.40%	92.50%
5	2e-5	32	93.30%	93.10%	92.90%	93.00%
5	2e-5	64	93.80%	93.60%	93.40%	93.50%
5	2e-5	128	94.10%	93.90%	93.70%	93.80%
5	3e-5	16	93.40%	93.20%	93.00%	93.10%
5	3e-5	32	93.80%	93.60%	93.40%	93.50%
5	3e-5	64	94.30%	94.10%	93.90%	94.00%
5	3e-5	128	94.60%	94.40%	94.20%	94.30%
5	5e-5	16	94.00%	93.80%	93.60%	93.70%
5	5e-5	32	94.40%	94.20%	94.00%	94.10%
5	5e-5	64	94.80%	94.60%	94.40%	94.50%
5	5e-5	128	95.00%	94.80%	94.60%	94.70%

Figure 15 displays the confusion matrix for the IndoRoBERTa model trained to identify mental health symptoms using data from social media This confusion matrix visualizes the count of samples per class that were correctly or incorrectly classified. The vertical axis represents the true labels, while the horizontal axis shows the labels predicted by the model.

The evaluation results show that the model can classify most of the samples correctly. A total of 1,586 samples were correctly classified as positive, with 30 samples incorrectly predicted as neutral and 21 samples incorrectly predicted as negative. For the neutral class, there were 1,028 samples that were correctly classified, while 15 samples were incorrectly classified as positive, and 15 others were incorrectly predicted as negative. In the negative class, the model successfully classified 2,599 samples correctly, but there were still 45 samples that were misclassified as positive and 39 samples that were mispredicted as neutral.

From the results of this confusion matrix, it can be concluded that the IndoRoBERTa model has very good performance in detecting mental health symptoms. The misclassification errors that occur are relatively low, especially in the neutral class which has the fewest number of errors. Although there are still some misclassified samples, the number is not significant compared to the overall data. These results show that IndoRoBERTa can provide high accuracy with minimal error rates, so it can be relied on in the task of classifying mental health symptoms based on data from social media.

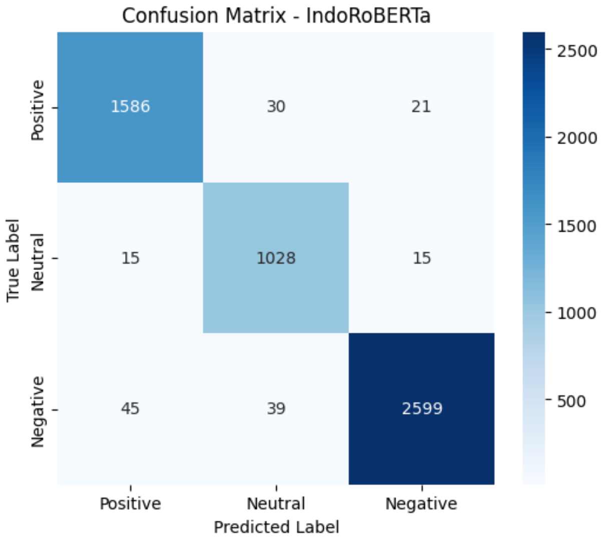


Fig. 15: Confusion Matrix of the IndoRoBERTa Model with the Best Performance

Model Implementation

For model testing, 5 new text data were taken randomly from platform X. The tweet data was fed into the model to automatically predict or detect whether the text is potentially depressive or not. The detection results can be seen in Table 13.

This table shows the results of implementing the IndoRoBERTa model in classifying texts related to mental health. The "Text" column contains various statements that reflect a person's emotional or mental condition. The "Actual" column represents the original label of the data, while the "Prediction" column is the prediction result of the model. The labels used in the classification consist of three categories, namely 1 for positive sentiment that reflects hope or a better mental condition, 0 for neutral sentiment that does not show strong emotional indications, and -1 for negative sentiment that describes a more vulnerable mental condition or has indications of mental health problems.

Table 10: Model Implementation

No	Text	Actual	Prediction
1	I often feel anxious for no clear reason	1	1
2	I feel empty and don’t know what to do	-1	-1
3	Today feels better, I hope it stays this way	0	0
4	have trouble sleeping every night, my mind is never at peace	1	1
5	Everything feels heavy, I feel hopeless	-1	-1

From the table, it can be seen that the model is able to make good predictions because all prediction results are in accordance with the actual labels. For example, the text "I often feel anxious for no clear reason" is categorized as a negative sentiment with a label of -1 because it indicates anxiety, and the model successfully provides the correct prediction. The same thing can be seen in the text "Everything feels heavy, I feel hopeless" which shows feelings of despair and is given a label of -1, with the model providing predictions in accordance with reality.

Conclusion

The conclusion of this study highlights that the IndoRoBERTa model outperforms other models in classifying texts related to mental health symptoms. Using a dataset scraped from social media The experiments further reveal that the optimal hyperparameter settings batch size of 128, learning rate of 5e-5, and 5 epochs yield the best results, delivering high precision, recall, and F1-score, demonstrating the model's strong effectiveness in this classification task

Additionally, the confusion matrix analysis reveals that the model exhibits a low error rate, with only a minimal number of false positives and false negatives. It effectively differentiates between texts expressing positive, neutral, and negative sentiments, making it a reliable tool for the early detection of mental health symptoms based on textual data.

Overall, this study proves that a deep learning-based approach, especially with the IndoRoBERTa model, can be an accurate solution in sentiment analysis related to mental health. The implementation of this model is expected to help in the development of an automatic mental health monitoring system, especially in identifying potential psychological problems from user interactions on social media. However, future research can explore further model improvement techniques, such as fine-tuning with a larger dataset or combination with other NLP methods to improve interpretability.

Acknowledgment

The author would like to thank the anonymous reviewers and editors who have reviewed this manuscript and for their valuable scientific comments and suggestions that have improved the quality of this manuscript. And thanks to Bina Nusantara University and all computer science lecturers who have provided knowledge and advice to this research.

Funding Information

The authors have not received any financial support or funding to report.

Author's Contribution

Andika Dwi Asmoro Wicaksono: Conducted this research from the initial stage, namely looking for ideas, to completing all other stages such as data crawling, data preprocessing, and data execution.

Rojali: Supervising lecturer who provided input and suggestions for this research.

Ethics

This study analyzes publicly available social media posts collected from platform X using mental-health-related keywords. No private or restricted-access data were used. All usernames, profile links, personal identifiers, and metadata were removed prior to analysis to ensure anonymity and protect user privacy. The research involved no interaction or intervention with individuals and posed minimal risk to users. According to the institutional guidelines of the authors' affiliation, this type of research using publicly accessible and anonymized data does not require formal ethics approval or IRB review. All authors have reviewed and approved the manuscript.

References

Althoff, T., Clark, K., & Leskovec, J. (2016). Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Healt. *Transactions of the Association for Computational Linguistics*, *4*, 463–476. https://doi.org/10.1162/tacl_a_00111

Cahyawijaya, S., Winata, G. I., Wilie, B., Vincentio, K., Li, X., Kuncoro, A., & Fung, P. (2021). IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 8875–8898. https://doi.org/10.18653/v1/2021.emnlp-main.699

Choudhury, D. M., Gamon, M., Counts, S., & Horvitz, E. (2021). Predicting Depression via Social Media. *Proceedings of the International AAAI Conference on Web and Social Media*, *7*(1), 128–137. https://doi.org/10.1609/icwsm.v7i1.14432

Da'im, S. I., & Abdurakhman, M. Si. (2024). *Perbandingan Model Transformer (BERT dan RoBERTa) pada Analisis Sentimen (Studi Kasus: Ulasan Game Mobile Legends)*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423

Dey, J., & Desai, D. (2022). NLP Based Approach for Classification of Mental Health Issues using LSTM and GloVe Embeddings. *International Journal of Advanced Research in Science, Communication and Technology*, *2*(1), 347–354. https://doi.org/10.48175/ijarsct-2296

Fuadi, M., Wibawa, A. D., & Sumpeno, S. (2023). idT5: Indonesian Version of Multilingual T5 Transformer. *Computation and Language*, *1*. https://doi.org/10.48550/arXiv.2302.00856

Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, *18*(2352–1546), 43–49. https://doi.org/10.1016/j.cobeha.2017.07.005

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, *349*(6245), 261–266. https://doi.org/10.1126/science.aaa8685

Koto, F., Lau, J. H., & Baldwin, T. (2021). IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 10660–10668. https://doi.org/10.18653/v1/2021.emnlp-main.833

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *Computation and Language*. https://doi.org/10.48550/arXiv.1907.11692

Reece, A. G., & Danforth, C. M. (2017). Instagram Photos Reveal Predictive Markers of Depression. *EPJ Data Science*, *6*(1). https://doi.org/10.1140/epjds/s13688-017-0110-z

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Computation and Language*. https://doi.org/https://doi.org/10.48550/arXiv.1910.01108

Saputra, A. C., Saragih, A. S., & Jurnal Teknologi Informasi, D. (2025). Prediksi emosi dalam teks bahasa Indonesia menggunakan model IndoBERT. *Jurnal Teknologi Informasi: Jurnal Keilmuan Dan Aplikasi Bidang Teknik Informatika*, *1*(15), 1–15. https://doi.org/10.47111/jti.v19i1.17617

Shen, J., Rudzicz, F., & Begum, S. (2017). Detecting anxiety through reddit. *Proceedings of the Eighth International Conference on Affective Computing and Intelligent Interaction*, 137–144.

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to Fine-Tune BERT for Text Classification? *In China National Conference on Chinese Computational Linguistics (CCL 2019)*, 194–206. https://doi.org/10.1007/978-3-030-32381-3_16

Tadesse, M. M., Lin, H. L., Xu, B., & Yang, L. (2019). Detection of depression-related posts in social media using hybrid model. *Proceedings of the Ieee International Conference on Big Data*, 1392–1397.

Tewari, A., Chhabria, A., Khalsa, A. S., Chaudhary, S., & Kanal, H. (2021). A Survey of Mental Health Chatbots using NLP. *Proceedings of the International Conference on Innovative Computing & Communication (ICICC) 2021. SSRN Electronic Journal*, 6. https://doi.org/10.2139/ssrn.3833914

Trappey, A. J., Lin, A. P., Hsu, K. Y., Trappey, C. V., & Tu, K. L. (2022). Development of an Empathy-Centric Counseling Chatbot System Capable of Sentimental Dialogue Analysis. *Processes*, *10*(5), 930. https://doi.org/10.3390/pr10050930

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS 2017)*, 5998–6008.

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., & Fung, P. (2020). IndoBERT: A Pre-trained Indonesian-Specific BERT Model. *ArXiv Preprint*.

Carbonell, J., Salakhutdinov, R., & V. Le, Q. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. *ArXiv Preprint*.

Ye, Y., Dai, Y., Xie, B., & Jian, D. (2021). Survey of life changes and mood during the covid-19 epidemic. *Frontiers in Public Health*, *9*, 10–12.