Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.3, June 2026, pp. 2155~2164
e-ISSN: 2548-6861

Enhancing Cyberbullying Sentiment Detection: A Comparative Study of IndoBERT and IndoBERTweet over SMOTE and Bernoulli Naive Bayes Approach

Eka Mardiana Putri Rahmawati [1]*, Mochammad Anshori [2]*, M. Syauqi Haris [3]*
* Informatics Study Program, Institut Teknologi Sains, dan Kesehatan RS. DR. Soepraoen Kesdam V/BRW
ekamardianaap@gmail.com [1], moanshori@itsk-soepraoen.ac.id [2], haris@itsk-soepraoen.ac.id [3]

Article Info

Article history:

Received 2026-03-05
Revised 2026-04-24
Accepted 2026-04-30

Keyword:

BERT,
Cyberbullying,
IndoBERT,
IndoBERTweet,
Natural Language Processing (NLP),
Sentiment Analysis

ABSTRACT

Cyberbullying has become a critical issue in social media use because it can negatively impact users' mental health and social interactions. The high volume of aggressive comments and hate speech on digital platforms highlights the need for an automatic detection system that can accurately and reliably identify cyberbullying content. This research compares the performance of Indonesian language transformer models, IndoBERT and IndoBERTweet, in detecting text-based cyberbullying. Before modeling, the dataset undergoes Exploratory Data Analysis to understand its characteristics, class distribution, comment length, and potential data imbalance. Next, text preprocessing and tokenization are performed before dividing the data using stratified holdout splitting to preserve class proportions in training and testing sets. Both models are then trained with the same hyperparameter settings to ensure an objective and fair performance comparison. Results show that IndoBERT achieved an accuracy of 0.8333, while IndoBERTweet performed better with an accuracy of 0.8409. The analysis of the confusion matrix and ROC curve confirms that IndoBERTweet is more effective at detecting cyberbullying across different classes. Compared to previous studies using the SMOTE method and Bernoulli Naïve Bayes algorithm, which achieved 84.00% accuracy, this study's findings are slightly higher at 84.09%. Notably, this was achieved without using synthetic oversampling techniques. This suggests that the approach employed in this research can deliver competitive performance even without data balancing with SMOTE. Overall, these findings indicate that a transformer-based approach, combined with a more representative dataset, can improve cyberbullying detection more efficiently and practically. Therefore, IndoBERTweet is a more suitable model for implementing a cyberbullying content moderation system in Indonesia.

I. INTRODUCTION

Cyberbullying is now viewed as a serious issue in the digital space, evident through the prevalence of hate speech, verbal harassment, and insults occurring online [1]. The United Nations Children's Fund (UNICEF) explains that cyberbullying is a form of harassment carried out by exploiting digital technology. This action can be placed through various media, including mobile phones, instant messaging services, and social networking platforms. This definition emphasizes that technological advancements also create opportunities for bullying through various digital communication channels [2]. According to a survey conducted by Polling Indonesia and the Indonesian Internet Service Providers Association (APJII), around 49% of internet users in Indonesia have experienced bullying on social media platforms [3]. The argument is further supported by the results of a survey conducted by the Center for Digital Society (CFDS), which involved 3,077 students at the junior high school and high school levels in Indonesia. The survey results show that 45.33% of respondents, equivalent to 1,182 students, admitted to having been involved as perpetrators of

cyberbullying [4]. The magnitude of these numbers indicates that cyberbullying is a widespread issue affecting mental health, including anxiety, depression, decreased self-esteem, and even the emergence of suicidal thoughts [5].

In the digital era, social pressure and exposure to negative content on social media increase the risk of psychological disorders, especially among the younger generation who are active in the online space [6]. Cyberbullying on social media platforms has been proven to have a wide-ranging impact on users' mental health [7]. In this context, social media, especially Instagram, serves as an interactive medium where users not only consume news but can also provide comments and opinions [8]. Comments laden with mockery, insults, and verbal harassment not only reflect the user's emotional state but also have the potential to worsen psychological conditions if not used wisely [9]. Thus, analyzing social media sentiment patterns becomes an important step for early detection of psychological conditions, which can be utilized in monitoring cyberbullying in society [10].

In this context, sentiment analysis is proposed as a way to understand users' perceptions and emotions about cyberbullying [11]. Sentiment analysis is a part of Natural Language Processing (NLP) that focuses on recognizing, interpreting, and classifying opinions or emotions in text [12], [13]. This technique helps researchers find emotional patterns in public conversations in online spaces [14]. However, social media text analysis is quite challenging due to its informal nature, full of noise, and often ambiguous, making it difficult for models to classify consistently [15]. This creates opportunities for the use of deep learning models in text analysis.

The development of deep learning technology, particularly transformers, has given a significant boost to sentiment analysis [16]. Transformers have become the main foundation in NLP, outperforming conventional models such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) in understanding long dependencies and complex sentence structures [17]. Transformers are capable of delivering better performance compared to conventional approaches in sentiment analysis across various datasets and languages [18], [19]. One of the prominent implementations is BERT (Bidirectional Encoder Representations from Transformers), a language representation model that has demonstrated superiority in various NLP tasks [17].

Recent literature emphasizes that the BERT model is very effective in handling various text classification tasks in the Indonesian context [20]. BERT has several variants, including IndoBERT and IndoBERTweet. IndoBERT has been proven to identify negative emotions and indications of psychological stress in Indonesian texts, making it effective in detecting online bullying behavior [21]. In addition, the IndoBERT model achieved very high performance with a precision of 0.9748, a recall of 0.9742, and an FI-Score of 0.9731 in detecting potential stress from user tweets [22]. On the other hand, IndoBERTweet is specifically designed to understand informal language, slang, and social media text characteristics, proving effective in identifying indications of depression on Twitter. The study noted an accuracy of around 82% in classifying potential depression [23]. Based on its reliability, IndoBERT and IndoBERTweet were chosen as the models in this research.

In addition to transformer-based approaches, several previous studies still rely on traditional classification methods such as Bernoulli Naïve Bayes combined with the class balancing technique Synthetic Minority Over-Sampling Technique (SMOTE) [24]. This approach is applied to optimize model performance by adding synthetic data to the class with fewer instances. In this way, the data composition becomes more balanced, thereby maintaining the distribution between classes. Furthermore, probabilistic models like Naïve Bayes operate with a relatively simple text representation approach, making them less optimal in capturing the contextual meaning relationships between words. Therefore, a method is needed that can understand the context of language more deeply, without relying on additional techniques to balance data distribution.

This research proposes the integration of BERT models, specifically IndoBERT and IndoBERTweet, to enhance sentiment analysis related to Indonesian language cyberbullying. IndoBERT's advantage lies in its training using an Indonesian language corpus containing over 220 million words, allowing it to identify language structures with higher precision [25]. Meanwhile, IndoBERTweet was trained using the Indonesian Twitter corpus with the application of domain-specific vocabulary initialization to handle informal expressions, making it more robust in social media language [26]. This integration is necessary because cyberbullying data contains a mix of formal language and non-standard emotional expressions, so the combination of both provides a more optimal analysis.

Thus, this research aims to design a BERT-based sentiment analysis model that is socially and linguistically adaptive to the Indonesian context, to produce a sentiment detection system that is more accurate, balanced, and sensitive to psychological expressions on social media. Through the implementation of this approach, it is expected to yield significant contributions, particularly in efforts for early detection and monitoring of cyberbullying cases. Additionally, the result obtained also has the potential to serve as a basis for formulating data-driven cyberbullying mitigation policies in Indonesia.

II. METHOD

This research takes a quantitative approach, focusing on the predictive ability of the BERT, which has been adapted to handle cyberbullying terms in Indonesian. An overview of the research is presented in Figure 1.

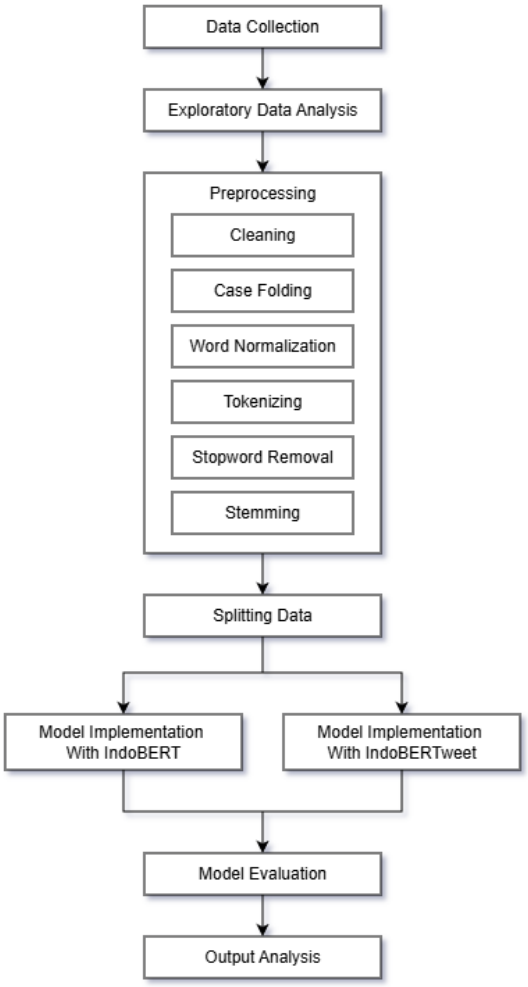


Figure 1. Research Workflow

Based on the research flow in Figure 1, the stages begin with data collection, exploratory data analysis (EDA), followed by preprocessing, data splitting, model implementation, and model performance evaluation.

A. Data Collection

In this study, cyberbullying is operationally defined as comments containing elements of insult, verbal harassment, hate speech, or negative expressions directed at individuals or groups directly on social media. Meanwhile, non-cyberbullying comments include texts that are neutral, informative, or do not contain elements of verbal aggression. This research uses the Indonesian Instagram Cyberbullying dataset obtained from the Mendeley Data platform [27]. This dataset is a collection of Indonesian language comments taken from the social media platform Instagram and has been annotated for cyberbullying purposes. In this study, three main attributes are used, namely text as the representation of the comment content, binary_label as the class label indicating the category of cyberbullying, and tokenized_comment as the result of tokenizing the previously processed text, thus facilitating the modelling stage.

B. Exploratory Data Analysis

The Exploratory Data Analysis stage is conducted to understand the initial characteristics of the dataset before preprocessing. This analysis includes the amount of data, data structure, label distribution, and general characteristics of the comment text. The findings at this EDA stage serve as the basis for monitoring the preprocessing stages applied. To standardize text representation and reduce noise before the data is used in the modeling stage [28].

C. Preprocessing

Raw data cannot be used directly because it contains irrelevant elements. This is consistent with recent research, which shows that user texts tend to be informal and require cleaning of ambiguous information [29]. The preprocessing stages carried out include:

1) *Cleaning*: This step is done to remove various non-linguistic components, such as URLs, hashtags, username mentions, emojis, numbers, and irregular characters, because these elements do not have semantic value and can potentially increase noise in the model training process [30].

2) *Case Folding*: At this stage, all text tokens are converted to lowercase (lowercase transformation). This is applied to standardize word representation and prevent feature duplication due to capitalization differences that do not carry semantic meaning [31].

3) *Word Normalization*: Normalization is carried out by mapping non-standard words, abbreviations, and slang vocabulary into standard forms. This step uses an informal dictionary and a standard word dictionary, supported by academic publications, providing scientific justification for its application [32].

4) *Tokenizing*: The text is segmented into tokens or word units (token-level segmentation) [33]. This process produces more structured text units, allowing words to be converted into the vector representations needed in the modelling stage [34]. This step is important to organize the input structure so that the machine learning model processes it well.

5) *Stopword Removal*: In this process, tokens with high occurrence frequency but low discriminative power against the target class are reduced. This aims to reduce feature redundancy and improve the signal-to-noise ratio in the input data, thereby enhancing the performance of the classification model [35].

6) *Stemming*: Applied to convert each token to its root form by removing morphological affixes [36]. This step can reduce feature sparsity and unify the model's generalization ability in detecting sentiment patterns.

D. Splitting Dataset

At the experimental stage, the dataset will be divided into two parts with an 80:20 ratio. 80% of the data will be allocated

for the model training process, while the remaining 20% will be used as test data. This separation process is carried out proportionally for each class, so that the composition of the test data still represents the class distribution present in the entire dataset. Thus, the division can maintain the balance and original characteristics of the data. In addition, such divisions are widely used in various studies because they have been proven to improve the efficiency of the computation process and ensure consistent model evaluation [37].

E. Model Implementation

The BERT model is a model based on the transformer architecture developed by Google. In its training process, BERT is designed to understand context bidirectionally through two main objectives. First, the Masked Language Model (MLM) is the task of predicting tokens that are deliberately hidden randomly within a sentence. Second, Next Sentence Prediction (NSP), which aims to study the relationship or connection between sentences. Through these two mechanisms, the model can capture a more comprehensive and contextual representation of language [38]. The BERT architecture is shown in Figure 2.

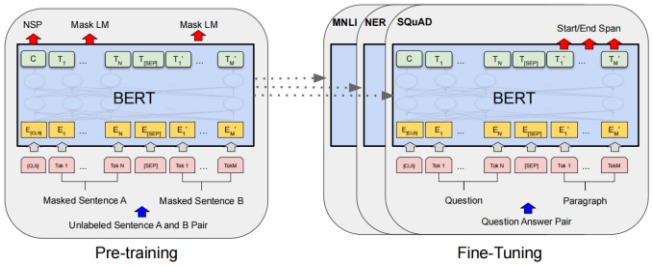


Figure 2. BERT Architecture [38]

The main stage of this research uses the application of IndoBERT and IndoBERTweet. To ensure a fair comparison, this study controls several key experimental factors. Both models were used in a base configuration with relatively comparable parameter scales, so the comparison was made at an equivalent model complexity level. Additionally, both models have undergone a pre-training process on a large-scale Indonesian language corpus, thus possessing similar foundational abilities in understanding language context.

The fine-tuning process is carried out using the same dataset, with identical data splitting schemes, number of epochs, and training procedures. Hyperparameter configurations such as learning rate, batch size, optimizer, and loss function are standardized to avoid bias caused by differences in training settings. This approach aims to ensure that the observed performance differences more accurately reflect the architectural characteristics and pre-training data of each model, rather than being the result of variations in experimental configuration.

Nevertheless, it should be noted that each transformer model has different characteristics and can potentially achieve optimal performance through a specific hyperparameter tuning process. Therefore, the use of the same configuration in this study is a deliberate strategy to maintain consistency and fairness in comparison (controlled experiment), but it also serves as a limitation of the research that can be further explored in future studies. From a technical perspective, both models utilize the self-attention mechanism to identify relationships between tokens in the text sequence [26], as formulated in Equation (1).

$$Attention\ (Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) V \quad (1)$$

In equation (1), the attention value is obtained by calculating the multiplication between the query (Q) and the key K^T, then dividing it by $\sqrt{d_k}$ to keep the score controlled. This result is then given to the softmax function to generate attention weights, which are subsequently used to combine the value (V) [39].

The BERT encoder architecture used in IndoBERT and IndoBERTweet processes tokens bidirectionally, so each token receives information from both the left and right sides simultaneously [40]. After passing through self-attention, the token representations are forwarded to the feed-forward network as shown in equation (2):

$$FFN(x) = max\ (0, x\ W_1 + b_1)\ W_2\ +\ b_2 \quad (2)$$

The FFN network functions to transform token representations non-linearly to enhance features before entering the next layer [41], [42]. Next, IndoBERT and IndoBERTweet function as transformer encoders optimized through pretraining based on Masked Language Modeling (MLM). This model incorporates new vocabulary using average subword embeddings, enabling it to comprehend informal language. The [CLS] representation is then used for classification, making IndoBERT and IndoBERTweet very effective for analyzing Indonesian language text.

F. Model Evaluation

After the training process is complete, the model is tested using testing data to assess its performance. Evaluation is conducted using the metrics of Accuracy, Precision, Recall, F1-Score, and Confusion Matrix. The purpose of the evaluation is to determine the level of success, and the model's success rate can be measured, both in terms of accuracy and consistency [43]. Mathematically, the calculation of evaluation metrics is formulated as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$F1-Score = 2\ x\ \frac{Precision\ x\ Recall}{Precision+Recall} \quad (6)$$

Through this confusion matrix, the distribution of the model's output can be analyzed based on four main components, namely True Positive (TP), False Positive (FP),

True Negative (TN), and False Negative (FN). This combination of metrics ensures that the model's performance assessment is conducted comprehensively, especially in handling sentiment variations in Indonesian-language social media texts.

III. Result and Discussion

This section presents the results and discussion based on all the methodological stages that have been designed, starting from data collection, exploratory data analysis, preprocessing, data splitting, implementation, evaluation, and output analysis. The findings obtained were analyzed to assess the performance of BERT in classifying cyberbullying sentiment on Instagram social media data.

A. Data Collection Results

Based on the data collection process that has been carried out, a total of 7,530 raw data points in the form of user comments will be used. The number of data labelled "yes" (cyberbullying) was 4,572, and labelled "no" (non-cyberbullying) was 2,958. The description of the dataset features used in the study is presented in Table 1, which includes the features text, binary_label, and tokenized_comment. For comments labeled as cyberbullying, it is associated with the emergence of words such as flaming, body shaming, harassment, and denigration.

TABLE I
SNAPSHOT OF THE DATASET

text	binary_label	tokenized_comment
Udah rusak penampilan, agama. Minimal akhlak lah	Yes	['sudah', 'rusak', 'tampil', 'agama', 'minimum', 'moral']
Beneran seee langsing enih, masya Allah	No	['benar', 'sih', 'langsing', 'masya', 'allah']
ko ga klarifikasi si smpe skrg? bener ya celingkuh? Wkwk	Yes	['tidak', 'klarifikasi', 'si', 'ya', 'celingkuh', 'wkwk']
Ini asli kan ya bukan editan badanya bagus sekali ily ðŸ˜	No	['asli', 'ya', 'tidak', 'edit', 'bada', 'bagus', 'ily']
Dihh Ogah banget dengerin lagu lu bang bikin sesek kupinggg!!	Yes	['dih', 'enggan', 'banget', 'dengar', 'lagu', 'bang', 'sek', 'kupingg']

B. Exploratory Data Analysis Results

TABLE 2
DESCRIPTIVE STATISTICS FOR WORD COUNT

Statistic	Word Count
Count	7530.000000
Mean	11.218459
Std	9.716459
Min	1.000000
25%	6.000000
50%	8.000000
75%	13.000000
max	157.000000

The Exploratory Data Analysis stage involves descriptive analysis to provide an understanding of the main characteristics of the dataset used. Summary statistics for the numeric variables are presented in Table 2. The table displays several descriptive measures that illustrate the characteristics of the data, including the mean, standard deviation, minimum value (min), percentage, and maximum value (max) [44]. Based on these results, the average comment length is 11,21 words with a standard deviation of 9,71, indicating a considerable variation in text length. Through this presentation, readers can obtain an overview of the distribution and trends of the analyzed data. A minimum of one word and a maximum of 157 words indicate that the dataset includes comments ranging from very short to relatively long. Meanwhile, the median value of 8 words indicates that most comments tend to be short.

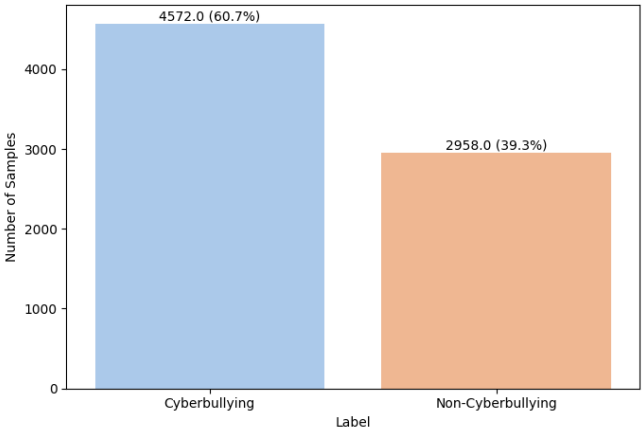


Figure 3. Sentiment Label Distribution

The EDA results on the label distribution in Figure 3 show a class imbalance, where the cyberbullying class has a significantly higher proportion compared to the non-cyberbullying class. Nevertheless, this research did not use data balancing techniques such as SMOTE in the main experiment. This is based on the consideration that transformer-based models like BERT have strong contextual representation capabilities, allowing them to capture language patterns even in imbalanced data distributions.

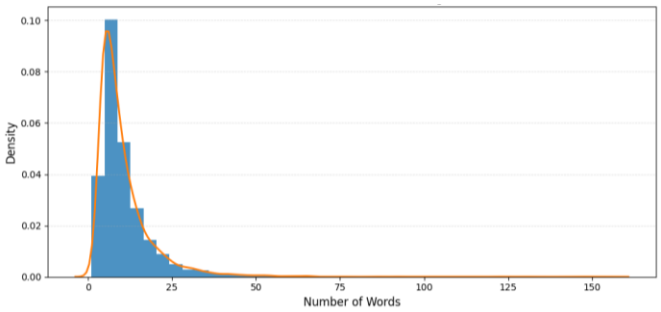


Figure 4. Distribution of Overall Text Length

Figure 4 illustrates the distribution of comment lengths in the dataset based on word count, providing an overview of text complexity. The distribution indicates that most comments are relatively short, with the highest frequency in texts containing fewer than 25 words. The data distribution is right-skewed, where only a few comments have very long text lengths. This finding suggests that Instagram comments tend to be brief and direct, necessitating special handling during the preprocessing stage, including noise removal, word normalization, and the selection of suitable text representation methods.

C. Data Preprocessing Results

This stage is applied to the entire raw data to generate standardized text. Changes in data at each stage are illustrated with example results in a table.

1) *Cleaning*: After the text cleaning process is completed, the resulting data only contains words that are considered relevant to the analysis needs. An illustration of the comparison of data before and after going through the cleaning at the stage in Table 3.

TABLE 3
DATA CLEANING RESULT

Input	Output
@spotifyid Please boycott Rio Clappy's song bunga abadi	Please boycott Rio Clappys song bunga abadi

2) *Case Folding*: After the case folding process, the entire text is converted to lowercase so that there are no differences in word representation due to capitalization variations. Example results from the stage in Table 4.

TABLE 4
CASE FOLDING RESULT

Input	Output
WKWK dalam salad penggemar garis keras Dura menjilat air liur itu sendiri	wkwk dalam salad penggemar garis keras dura menjilat air liur itu sendiri

3) *Word Normalization*: The normalization stage results in the transformation of non-standard words into standard forms, thereby reducing excessive lexical variation. An example of the application of this stage is in Table 5.

TABLE 5
WORD NORMALIZATION RESULT

Input	Output
Oh ini betina jadi selingkuhannya, wajar lah keturunan bule doyan begituan ðŸ˜,	oh ini betina jadi selingkuhannya wajar lah keturunan bule doyan begituan

4) *Tokenizing*: The tokenization process is carried out using the built-in tokenizers of each model, namely the IndoBERT and IndoBERTweet tokenizers based on subword (WordPiece/BPE). This approach allows the model to still recognize non-standard words, slang, or word forms not found in the standard dictionary by breaking them down into smaller subword units. An example of the results of this stage is in Table 6.

TABLE 6
TOKENIZING RESULT

Input	Output
Beneran nih Raffi penampung pencucian uang? Wahh parah sih kalo bener.	['benar', 'raffi', 'tampung', 'cuci', 'uang', 'kritis', 'sih']

5) *Stopword Removal*: After stopword removal, words with low discriminative power were successfully eliminated, reducing the number of irrelevant tokens. An example of the results of this stage is in Table 7.

TABLE 7
STOPWORD REMOVAL RESULT

Input	Output
Orang klo baru naek udah sombong amat pasti jatohnya juga cepet	['orang', 'sombong', 'jatohnya', 'cepat']

6) *Stemming*: At this stage, the token has been successfully converted to its base form, allowing for the morphological variations of the word to be unified in a single representation. An example of this stage is in Table 8.

TABLE 8
STEMMING RESULT

Input	Output
Bangga lokal kalau Nagita aja senengnya pake brand luar asal aneh nya ðŸ˜,	bangga lokal nagita neng pakai brand aneh

D. Data Splitting Results

The next step is to divide the dataset, which has a total of 7,531 records, into two parts training and testing, with an 80:20 composition. The data division and the number of records allocated at this stage are in Table 9.

TABLE 9
DATA SHARING

Data	Percentage	Amount
Training	80%	5971
Testing	20%	1493

In this process, a stratified holdout splitting ratio of 80:20 is used. The use of stratified holdout splitting in the data partitioning process ensures a balanced class distribution between the training and test data. This affects the results, considering that cyberbullying data tends to have a class imbalance, so the model not only focuses on the majority class but is also able to recognize patterns in the minority class.

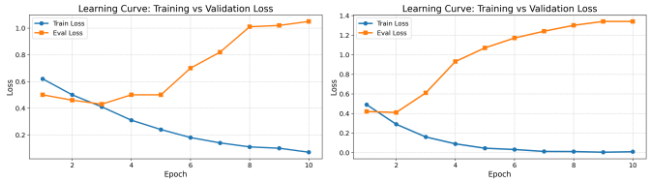
E. Model Implementation Results

At the implementation stage, the IndoBERT and IndoBERTweet Base models were fine-tuned using a cyberbullying dataset with key hyperparameters as shown in Table 10.

TABLE 10
MAIN HYPERPARAMETER

Parameter	Value
Number of Epochs	10
Batch Size	16
Learning Rate	2e-5
Optimizer	AdamW
Loss Function	Weight Cross-Entropy Loss
Evaluation Strategy	Per Epoch

After the key hyperparameters shown in Table 10 have been determined, the next step is to perform the fine-tuning process of the IndoBERT and IndoBERTweet models with the cyberbullying dataset. The fine-tuning process is carried out by updating all model weights using the pre-determined hyperparameter configuration to adapt to the language characteristics found in the cyberbullying-related data. The input text is tokenized using the built-in tokenizer of each model, with padding and truncation mechanisms to match the input length required by the model. This aims for the model to learn more specific contextual representations related to cyberbullying data, both formal and informal.



(a) (b)

Figure 5. (a) IndoBERT Learning Curve, (b) IndoBERTweet Learning Curve

Based on the learning curve graph, IndoBERT in Figure 5(a) shows the training loss consistently decreasing from the early epochs to the end, indicating an effective learning process. Meanwhile, the validation loss initially decreased at the beginning of the epoch but then increased in the subsequent epochs. Meanwhile, IndoBERTweet in Figure 5(b) shows a more stable and consistent validation accuracy throughout the epochs, thus maintaining its performance on the validation data. This shows that IndoBERT is more suitable for informal social media texts such as Instagram comments.

Overall, the implementation results of the IndoBERT model yielded an accuracy of 0.8333, while the IndoBERTweet model achieved a higher accuracy of 0.8409. The difference in results indicates that IndoBERTweet is more effective in handling informal social media language that frequently appears in the context of cyberbullying. This is in line with the characteristics of IndoBERTweet, which is pre-trained using Twitter data, making it better suited for language variations, abbreviations, and non-standard expressions that often appear in the context of cyberbullying.

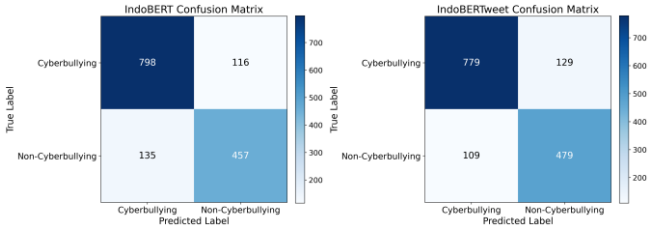
F. Model Evaluation Results

TABLE 11
EVALUATION RESULT SUMMARY OF INDOBERT AND INDOBERTWEET

Model	Evaluation Metrics			
	Acc	Prec	Recall	F1
IndoBERT	0.8333	0.8326	0.8333	0.8328
IndoBERTweet	0.8409	0.8421	0.8409	0.8413

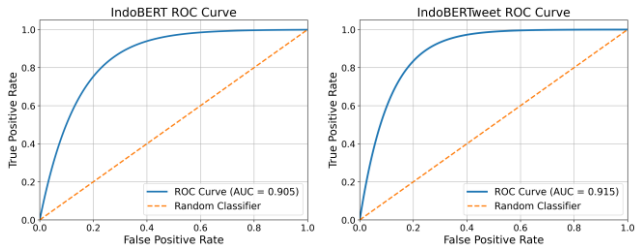
In the evaluation, several metrics were used, namely accuracy, precision, recall, and F1-score on the test data to provide a more comprehensive picture of the model's performance. Based on the results in Table 11, IndoBERTweet shows slightly higher performance compared to IndoBERT across all evaluation metrics. IndoBERT achieved an accuracy of 0.8333, while IndoBERTweet reached 0.8409, with consistent improvements also observed in precision, recall, and F1-score values. However, the performance difference between the two models is relatively small, thus not indicating a substantial difference in classification capability. This pattern of results indicates that both models have relatively comparable capacities in detecting cyberbullying. The observed superiority of IndoBERTweet tends to be marginal, yet consistent across various metrics, which can be attributed to its suitability for the characteristics of social media data that contain informal language.

Thus, the results of this study indicate that although IndoBERTweet tends to perform better, the difference is not significant enough to conclude a dominant advantage. Therefore, the selection of the model in the context of implementation can consider other factors such as efficiency, simplicity of the pipeline, and suitability with the data domain used.



(a) (b)

Figure 6. (a) IndoBERT Confusion Matrix, (b) IndoBERTweet Confusion Matrix



(a) (b)

Figure 7. (a) IndoBERT ROC Curve, (b) IndoBERTweet ROC Curve

The evaluation results of the IndoBERT model based on the confusion matrix in Figure 6(a) show that the model is capable of identifying data that is still misclassified as non-cyberbullying. For the non-cyberbullying class, the model successfully classified 457 data points accurately, with 135 data points misclassified as cyberbullying. This indicates that the model has adequate capabilities in identifying patterns in datasets containing elements of cyberbullying, although there are still classification errors generally caused by the similarity of language context between classes.

Although the IndoBERT model has shown quite good performance in detecting cyberbullying, the evaluation results still show classification errors, especially in data containing informal language and typical social media expression variations. Therefore, in the next stage, this research utilizes IndoBERTweet, a model specifically trained on Indonesian language Twitter data. The evaluation results of the IndoBERTweet model show improved performance in detecting cyberbullying cases compared to before. Referring to the confusion matrix in Figure 6(b), the model was able to correctly classify 779 cyberbullying data, while 129 cyberbullying data were still misclassified as non-cyberbullying. On the other hand, for the non-cyberbullying class, the model successfully identified 479 data correctly, with 109 data incorrectly predicted as cyberbullying.

Next, the ROC curve in Figure 7(a), the IndoBERT model achieved an Area Under Curve (AUC) value of 0.905, indicating the model's good discriminatory ability in classifying cyberbullying and non-cyberbullying classes. AUC value close to 1 indicates that the model has a high level of reliability in classification, as well as performance that is far better compared to random classification.

Furthermore, the ROC curve in Figure 7(b) shows an AUC value of 0.915, indicating the IndoBERTweet model has excellent discrimination ability in distinguishing between cyberbullying and non-cyberbullying classes. A higher AUC value compared to IndoBERT confirms that IndoBERTweet is more effective in capturing the characteristics of informal language and the context of social media conversations, thereby providing more optimal performance in the task of cyberbullying detection.

G. Interpretability Analysis

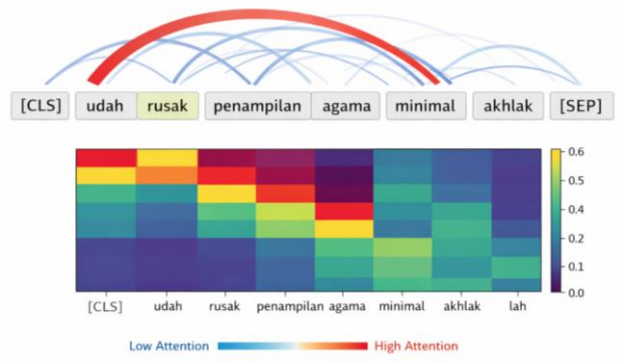


Figure 8. Attention Visualization (Heatmap)

Model interpretability is analyzed using attention weights on the transformer layer. The results show that tokens with negative connotations tend to have higher attention values compared to other tokens. This shows that the model is capable of identifying keywords that contribute to the classification of cyberbullying. To improve model transparency, interpretability analysis is conducted through attention visualization on BERT-based models. Figure 8 shows that the model pays the closest attention to tokens such as "damaged," "appearance," and "morality," which have negative connotations and are the main indicators in cyberbullying classification. On the other hand, tokens like "already" have a lower attention weight, indicating their minimal contribution to the model's decision. In addition, tokens like "religion" continue to receive attention due to their role in the overall context of the sentence. This pattern shows that the model does not only rely on a single word, but also considers the relationships between tokens in understanding meaning.

H. Comparative Analysis with Previous Research

TABLE 12
COMPARISON RESULT

Study	Dataset Size	Method	Data Split	Accuracy
Previous Study [24]	2.000	SMOTE + Naïve Bayes	70:30	84.00%
This Study	7.530	IndoBERTweet	80:20	84.09%

Based on the comparison with previous research [24], It can be seen in Table 12 that the previous study applied the SMOTE technique to address class imbalance in 2,000

Instagram comments, increasing the data to 2,204 with a 70:30 holdout split, and achieving an accuracy of 84.00%. Unlike that study, this research uses a larger and more diverse dataset, consisting of 7,530 Instagram comments without the need for artificial balancing intervention. Although only using the stratified holdout method with an 80:20 ratio, the results obtained show that the transformer-based model is able to effectively learn cyberbullying patterns from the original data distribution.

It should be noted that this study did not conduct a direct comparison between the use of SMOTE and without SMOTE in the same scenario. However, the approach without SMOTE shows efficiency from the pipeline side, as it does not require additional processes such as synthetic data generation and oversampling result validation. Additionally, transformer-based models are capable of learning data directly thru contextual representations, making them effective even when the data is unbalanced. Thus, efficiency in this research refers to the simplicity of the process and the reduction of computational complexity, not just performance improvement. This shows that a more representative dataset, combined with a transformer model, can be an effective alternative to oversampling techniques. Therefore, this research offers a simpler and more practical approach with competitive performance in detecting Indonesian language cyberbullying.

Based on these findings, this research not only focuses on performance but also emphasizes data efficiency and representativeness as important factors in model development. Therefore, the contributions of this research are formulated in three main aspects. First, evaluating and comparing the performance of IndoBERT and IndoBERTweet on a larger Indonesian language cyberbullying dataset. Second, demonstrating that transformer-based models can achieve competitive performance without oversampling techniques like SMOTE, thus better reflecting real data conditions. Third, analyzing the model's ability to handle informal social media language as a major challenge in cyberbullying detection.

IV. CONCLUSION

This research successfully developed a cyberbullying detection model for Indonesian-language social media by utilizing a transformer-based approach through a comparison of IndoBERT and IndoBERTweet. Based on the implementation and evaluation results, IndoBERT achieved an accuracy of 83.00%, while IndoBERTweet showed improved performance with an accuracy of 84.09%, and has better capabilities in handling informal language patterns commonly used in social media comments.

Compared to previous research that used Bernoulli Naïve Bayes with SMOTE and achieved an accuracy of 84.00%, this study attained a slightly higher result of 84.09% without oversampling. This demonstrates that a larger dataset and proper stratified data division can deliver competitive performance with a simpler approach. Therefore, this research offers a practical and stable approach. The model shows potential as an initial component in the social media content moderation system in Indonesia, but it still requires further evaluation in real-world scenarios. It should be noted that the exclusion of classical machine learning models as a baseline is a limitation of this study. This is due to fundamental differences in feature representation and preprocessing requirements compared to transformer-based models, making direct comparisons potentially unfair. Therefore, this research focuses on comparing models with similar architectures.

As a next step, further research is recommended to compare other models, such as IndoRoBERTa or DistilBERT, and to expand the dataset to make the model more robust against variations in language and the context of cyberbullying on social media. Additionally, further research can incorporate a multi-label classification approach and experiment with more effective fine-tuning strategies to enable the system to identify multiple types of cyberbullying.

REFERENCES

[1] G. Ray, C. D. McDermott, and M. Nicho, "Cyberbullying on Social Media: Definitions, Prevalence, and Impact Challenges," *Journal of Cybersecurity*, vol. 10, no. 1. Oxford University Press, 2024. doi: 10.1093/cybsec/tyae026.

[2] C. Murphy, "Cyberbullying among young people: Laws and policies in selected Member States," *EPRS | Eur. Parliam. Res. Serv.*, no. June 2024, pp. 1–12, 2024.

[3] N. Fiddiana and A. Bagus Priyambodo, "The Correlation Between Self-Control and Cyberbullying at Private High School X in Bogor," *KnE Soc. Sci.*, vol. 2021, no. ICoPsy 2021, pp. 255–266, 2022, doi: 10.18502/kss.v7i1.10216.

[4] D. Herlambang, R. Zildjianda, and D. Brajannoto, "Analysis of Cyberbullying Among Students: a Legal Perspective in Indonesia," *Entita J. Pendidik. Ilmu Pengetah. Sos. dan Ilmu-Ilmu Sos.*, no. 11, pp. 51–62, 2025, doi: 10.19105/ejpis.v1i.19134.

[5] N. Agustiningsih, A. Yusuf, A. Ahsan, and Q. Fanani, "The impact of bullying and cyberbullying on mental health: a systematic review," *Int. J. Public Heal. Sci.*, vol. 13, no. 2, p. 513, Jun. 2024, doi: 10.11591/ijphs.v13i2.23683.

[6] G. J. Hjetland *et al.*, "Digital self - presentation and adolescent mental health : Cross - sectional and longitudinal insights from the ' LifeOnSoMe '- study," *BMC Public Health*, 2024, doi: 10.1186/s12889-024-20052-4.

[7] K. Fitzgerald, L. Vandenbosch, and T. Tabruyn, "Adolescent Emotional Expression on Social Media: A Data Donation Study Across Three European Countries," *Affect. Sci.*, vol. 5, no. 4, pp. 436–448, Dec. 2024, doi: 10.1007/s42761-024-00259-9.

[8] N. Aulia Maharani and W. Wahyunengsih, "Social Media Instagram as a News Reference for Journalism Students," *J. Karya Abdi Masy.*, vol. 8, no. 1, pp. 41–48, 2024, doi: 10.22437/jkam.v8i1.26176.

[9] L. Ismail, U. Mukramin, A. Nursida, and U. Muhammadiyah Makassar, "The Impact of Hate Comments on Social Media Users' Self-Confidence: An Analysis Grounded in Social Comparison Theory and Cyberbullying Research," *PELNUS | Pen en Light Nat. Union Sci.*, vol. 3, no. 2, pp. 49–59, 2024, doi: https://doi.org/10.35335/psychologia.v3i2.57.

[10] M. Couto, A. Perez, J. Parapar, and D. E. Losada, "Temporal Word Embeddings for Early Detection of Psychological Disorders on Social Media," *J. Healthc. Informatics Res.*, 2025, doi: 10.1007/s41666-025-00186-9.

[11] F. Ramadan, E. Hassan, and F. A. Omara, "Cyberbullying detection in social media using natural language processing," *Sci. African,*

vol. 28, no. April, p. e02713, 2025, doi: 10.1016/j.sciaf.2025.e02713.
[12] A. S. Talaat, "Sentiment analysis classification system using hybrid BERT models," *J. Big Data*, 2023, doi: 10.1186/s40537-023-00781-w.
[13] "Matematika Pada Kecerdasan Buatan - Widyastuti Andriyani, Mochammad Anshori, Dwi Normawati, Risqy Siwi Pradini, Mohamad Zaenudin, Muhammad Iqbal Harisuddin, M. Syauqi Haris, Astuty, Anna Angela Sitinjak, Wahyu Teja Kusuma - Google Books." Accessed: Feb. 26, 2026. [Online]. Available: https://books.google.co.id/books?hl=en&lr=&id=cEYhEQAAQBAJ&oi=fnd&pg=PP1&ots=q4w9L__NSd&sig=p6pu2rV06VpfrZfkh6oJometj6k&redir_esc=y#v=onepage&q&f=false
[14] M. L. Jamil, S. Pais, J. Cordeiro, and G. Dias, "Detection of extreme sentiments on social networks with BERT," *Soc. Netw. Anal. Min.*, vol. 12, no. 1, pp. 1–16, Dec. 2022, doi: 10.1007/s13278-022-00882-z.
[15] U. States, "Deep Learning--based Text Classification: A Comprehensive Review," *ACM Comput. Surv.*, vol. 54, no. 3, 2026, doi: 10.1145/3439726.
[16] N. E. Zekaoui, S. Yousfi, M. Rhanoui, and M. Mikram, "Analysis of the evolution of advanced transformer-based language models : experiments on opinion mining," *IAES Int. J. Artif. Intell.*, vol. 12, no. 4, 2023.
[17] N. Mushtaq, G. Ali, D. Muhammad, K. Malik, and A. Bukhari, "BERT applications in natural language processing : a review," *Springer Nat.*, vol. 58, 2025, [Online]. Available: https://doi.org/10.1007/s10462-025-11162-5
[18] P. Xue and W. Bai, "A Fine-Grained Sentiment Analysis Method Using Transformer for Weibo Comment Text," *Int. J. Inf. Technol. Syst. Approach*, vol. 17, no. 1, pp. 1–24, Jul. 2024, doi: 10.4018/ijitsa.345397.
[19] A. Amrullah, "Sentiment Analysis in the Age of Transformers and Large Language Models : A Comprehensive Review of Recent Advances and Future," *Intellithings J.*, vol. 1, no. 1, 2025.
[20] K. Kamdan, M. P. Anugrah, M. J. Almutaali, R. Ramdani, and I. L. Kharisma, "Performance Analysis of IndoBERT for Detection of Online Gambling Promotion in YouTube Comments †," *Eng. Proc.*, vol. 107, no. 1, 2025, doi: 10.3390/engproc2025107066.
[21] A. A. Hafiza and E. B. Setiawan, "Enhancing Cyberbullying Detection on Platform 'X' Using IndoBERT and Hybrid CNN-LSTM Model," *J. Tek. Inform.*, vol. 6, no. 2, pp. 655–672, Apr. 2025, doi: 10.52436/1.jutif.2025.6.2.4321.
[22] M. S. M. Faza, U. Islam, S. Agung, and A. Azka, "Penerapan Model IndoBERT untuk Deteksi Potensi Sumber Stres dalam Teks Media Sosial," *J. Rekayasa Sist. danTeknologi*, vol. 3, no. 1, pp. 271–283, 2025, doi: https://doi.org/10.70248/jrsit.v3i1.3046.
[23] M. Fadhel and W. Maharani, "Depression Detection of Users in Social Media X using IndoBERTweet," *Sinkron*, vol. 8, no. 2, pp. 885–891, 2024, doi: 10.33395/sinkron.v9i2.13354.
[24] Akira Permata Ramadhani, Eka Dyar Wahyuni, and Amalia Anjani Arifiyanti, "Klasifikasi Cyberbullying pada Komentar Instagram dengan Menggunakan Supervised Learning," *Neptunus J. Ilmu Komput. Dan Teknol. Inf.*, vol. 2, no. 2, pp. 92–101, 2024, doi: 10.61132/neptunus.v2i2.108.
[25] I. R. Hidayat and W. Maharani, "General Depression Detection Analysis Using IndoBERT Method," *Int. J. Inf. Commun. Technol.*, vol. 8, no. 1, pp. 41–51, 2022, doi: 10.21108/ijoict.v8i1.634.
[26] F. K. J. H. L. T. Baldwin, "Indobertweet : A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization," *Assoc. Comput. Linguist.*, vol. https://ac, pp. 10660–10668, 2021, doi: https://doi.org/10.18653/v1/2021.emnlp-main.833.
[27] E. D. Wahyuni, "Indonesia Instagram cyberbullying," Mendeley Data. [Online]. Available: https://data.mendeley.com/datasets/xthb26ntc5/1
[28] I. O. Muraina, O. M. Adesanya, M. A. Agoi, and S. Onen, "The Necessity of Exploratory Data Analysis : How are preprocessing activities beneficial to Data Analysts and Professional Researchers in Academia ?," *Int. J. Sci. Res. Comput. Sci. Eng.*, vol. 11, no. 3, pp. 22–28, 2023.
[29] H. T. Duong and T. A. Nguyen-Thi, "A review: preprocessing techniques and data augmentation for sentiment analysis," *Comput. Soc. Networks*, vol. 8, no. 1, pp. 1–16, Dec. 2021, doi: 10.1186/s40649-020-00080-x.
[30] M. Kastrati, Z. Kastrati, A. Shariq Imran, and M. Biba, "Leveraging distant supervision and deep learning for twitter sentiment and emotion classification," *J. Intell. Inf. Syst.*, vol. 62, no. 4, pp. 1045–1070, Aug. 2024, doi: 10.1007/s10844-024-00845-0.
[31] M. Siino, I. Tinnirello, and M. La Cascia, "Is text preprocessing still worth the time ? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers," *Inf. Syst.*, vol. 121, no. July 2023, p. 102342, 2024, doi: 10.1016/j.is.2023.102342.
[32] N. A. Salsabila, Y. Ardhito, W. Ali, A. Septiandri, and A. Jamal, "Colloquial Indonesian Lexicon," *2018 Int. Conf. Asian Lang. Process.*, pp. 226–229, 2018.
[33] C. W. Schmidt *et al.*, "Tokenization Is More Than Compression," *EMNLP 2024 - 2024 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf.*, pp. 678–702, 2024, doi: 10.18653/v1/2024.emnlp-main.40.
[34] F. Qarah and T. Alsanoosy, "A Comprehensive Analysis of Various Tokenizers for Arabic Large Language Models," *Appl. Sci.*, vol. 14, no. 13, Jul. 2024, doi: 10.3390/app14135696.
[35] J. Xiao and M. Hong, "A Feature Selection Method Based on a Convolutional Neural Network for Text Classification," *Electron.*, vol. 14, no. 23, Dec. 2025, doi: 10.3390/electronics14234615.
[36] F. A. D. Aryanti, A. Luthfiarta, and D. A. I. Soeroso, "Aspect-Based Sentiment Analysis with LDA and IndoBERT Algorithm on Mental Health App: Riliv," *J. Appl. Informatics Comput.*, vol. 9, no. 2, pp. 361–375, 2025, doi: 10.30871/jaic.v9i2.8958.
[37] M. N. Anshory, M. I. Mazdadi, T. H. Saragih, and S. W. Saputro, "Application of Adaboost Algorithm with SMOTE and Optuna Techniques in Sleep Disorder Classification," *Indones. J. Electron. Electromed. Eng. Med. Informatics*, no. 2, pp. 415–426, 2025, doi: https://doi.org/10.35882/ijeeemi.v7i2.99.
[38] R. L. Mulianingrum and E. Y. Hidayat, "Comparative Performance of SVM and BERT-Base Using Hybrid Preprocessing for Fast Fashion Sentiment Analysis," *J. Appl. Informatics Comput.*, vol. 9, no. 6, pp. 3464–3478, 2025.
[39] A. Vaswani, "Attention Is All You Need," no. Nips, 2017.
[40] M. D. Deepa and A. Tamilarasi, "Bidirectional Encoder Representations from Transformers (BERT) Language Model for Sentiment Analysis task : Review," *Turkish J. Comput. Math. Educ.*, vol. 12, no. 7, pp. 1708–1721, 2021.
[41] A. V. L. Y. A. H. S. Pires, Telmo Pessoa, "One Wide Feedforward is All You Need," *Assoc. Comput. Linguist.*, vol. https://ac, pp. 1031–1044, 2023, doi: https://doi.org/10.18653/v1/2023.wmt-1.98.
[42] T. L. Models, "Layerwise Importance Analysis of Feed-Forward Networks in Transformer-based Language Models," *Publ. as a Conf. Pap. COLM 2025*, no. 2024, pp. 1–21, 2025, doi: https://doi.org/10.48550/arXiv.2508.17734.
[43] H. W. Wibowo, M. Hasbi, and M. Anshori, "Use Discriminant Analysis to Identify Eroticism-Related Terms in The Lyrics of Dangdut Songs," 2024.
[44] R. Premathilaka, "Context-Aware Detection of Deceptive Design Patterns in E-Commerce Websites Using Word Embedding Based Deep Learning Paradigms," *J. Ilmu Komput. dan Inf. (Journal Comput. Sci. Information)*, vol. 2, pp. 239–249, 2025.