Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.4, August 2026, pp. 3837~3848
e-ISSN: 2548-6861

Indonesian Cyberbullying Detection Using IndoBERTweet-BiGRU Model on Class-Imbalanced X (Twitter) Data

Fajria Ulumin Nafiah[1*], Aviolla Terza Damaliana[2]*, Kartika Maulida Hindrayani[3*]
*Data Science, Universitas Pembangunan Nasional "Veteran" Jawa Timur
22083010081@student.upnjatim.ac.id[1], aviolla.terza.sada@upnjatim.ac.id[2] , kartika.maulida.ds@upnjatim.ac.id[3]

Article Info

Article history:

Received 2026-06-24
Revised 2026-08-10
Accepted 2026-08-12

Keyword:

Cyberbullying,
IndoBERTweet,
BiGRU,
Focal Loss,
X (Twitter)

ABSTRACT

Cyberbullying on social media platforms, particularly X (formerly Twitter), has become a serious issue that negatively affects users' mental health and well-being. Automatic cyberbullying detection in Indonesian remains challenging due to the widespread use of informal language, slang, abbreviations, and highly imbalanced class distributions. This study proposes a hybrid deep learning model that integrates IndoBERTweet with a Bidirectional Gated Recurrent Unit (BiGRU) to improve cyberbullying detection performance on Indonesian tweets. A dataset of Indonesian tweets was collected from X and annotated using a multi-stage dual large language model (LLM) labeling strategy to reduce the time and effort required for manual annotation while maintaining label consistency. To address class imbalance, this study investigates the effectiveness of Focal Loss and label distribution modification through multiple experimental scenarios. The proposed approach was evaluated using accuracy, precision, recall, and F1-score. The best performance was achieved by combining Focal Loss with a modified four-class label configuration consisting of Rude and Vulgar Words, Sexual Harassment, Body Shaming and Hate Speech, and Non-Cyberbullying. This configuration obtained an accuracy of 0.93, precision of 0.90, recall of 0.90, and F1-score of 0.90. These findings demonstrate that integrating contextual language representations with sequential modeling, supported by an efficient LLM-assisted labeling strategy and class imbalance handling, provides an effective approach for Indonesian cyberbullying detection and offers a practical solution for large-scale social media content moderation.

I. INTRODUCTION

Social media is already been an inseparable part of modern society. It has affected various aspects of human life, including the way people interact, communicate, and behave [1]. Social media serves as an effective communication tool that enabling people to communicate over long distances more easily. In addition, it offers several other advantages, such as expanding social connections, making online communities, and enhancing opportunities to establish new friendships [2]. These advantages have contributed to the significant growth social media usage, especially in Indonesia [3].

According to DataReportal, Indonesia had 143 million social media users at the beginning of 2025, representing 50.2% of the country's total populations [4]. This number has increased compared the previous year [5]. The growing use of social media has also led to various challenges, one of which is cyberbullying [6]. X (formerly Twitter) is one of the most popular social media platforms that allows the users to express and share their opinions. However, this could increase the likelihood of negative comments, offensive language, and hate speech, which may contribute to cyberbullying [7].

Cyberbullying can have serious consequences for both the physical and mental health to its victims[8]. Victims may experience depression, anxiety, psychological disorders, and post-traumatic stress symptoms [9]. In addition, cyberbullying can lead to behavioral changes, such as increased aggression, social withdrawal, and difficulties in forming new intrapersonal relationships [10]. In more severe

cases, cyberbullying may even contribute to suicidal thoughts or behaviors, particularly among adolescents [11]. Based on these concerns, it's important to develop an automated system that capable of detecting cyberbullying in Indonesian language text based social media content, which often contains informal expressions and slang. In this study, a Natural Language Processing (NLP) approach is employed to develop the proposed detection system. NLP has been widely applied to various tasks, including sentiment analysis and text classification [12]. Therefore, this study utilizes a text classification approach to automatically identify cyberbullying in Indonesian social media text. Several language models have been developed to effectively process Indonesian text data. One of the most widely used models is IndoBERT, which has demonstrated better performance than multilingual BERT (mBERT) on Indonesian language classification task such as sentiment analysis [13].

Previous studies have investigated Indonesian cyberbullying detection using various deep learning approaches. The first study focused on detecting cyberbullying in Indonesian YouTube comments using a Convolutional Neural Network with Long-Short Term Memory (CNN-LSTM) model. The proposed model employed a CNN layer for feature extraction and an LSTM layer to capture the contextual information within the text. The result of this experiment achieved an accuracy score of 96% [14]. Another study investigated the implementation of IndoBERT for cyberbullying detection using VADER Lexicon and a Generative Large Language Model (LLM) for multi-stage labelling. The fine-tuned IndoBERT model achieved an accuracy of 96.7%, outperforming the LSTM and Bi-LSTM models [15].

Those models have demonstrated promising performance on Indonesian social media text. However, a previous study [16] proposed IndoBERTweet, a pre-trained language model specifically designed for Indonesian social media text containing slang and informal expressions. IndoBERTweet was pre-trained on 26 million Indonesian tweets and incorporates 46% additional vocabulary relevant to informal language on social media. As a result, the model is better able to understand various forms of informal Indonesian, including abbreviations and slang. A previous study [17] evaluated the performance of IndoBERTweet by comparing it with IndoBERT and mBERT for multilabel classification of students' feedback. The study demonstrated that IndoBERTweet achieved the highest performance with an F1-score of 0.84, outperforming both IndoBERT and mBERT, which each obtained an F1-score of 0.82. A study about applying hybrid model on IndoBERTweet with BiLSTM for detecting Indonesian hatespeech on Twitter. Another study proposed a hybrid model that combines IndoBERTweet and BiLSTM for Indonesian hate speech detection on Twitter. The experimental results showed that hybrid IndoBERTweet-BiLSTM model achieved an accuracy of 93.3%, outperforming IndoBERTweet-CNN model, which achieved an accuracy of 92.2%. These findings indicate that IndoBERTweet performs effectively on Indonesian social media test dataset [18]. Based on several previous studies, this research proposes a hybrid model combining IndoBERTweet and Bidirectional Gated Recurrent Unit (BiGRU) for Indonesian cyberbullying detection.

Bi-Directional Gated Recurrent Unit (BiGRU) is a variant of Recurrent Neural Network architecture that processes input sequences in both forward and backward directions [19]. This ability enables the model to capture contextual information from both past and future tokens. BiGRU has a simpler architecture comparing other RNN-based architectures, but it's still capable to capture the contextual information effectively [20]. The combination of pre-trained language models and RNN architectures has been shown to improve the performance of various text classification tasks. A previous study demonstrated that hybrid models combining BERT and DistilBERT combined with BiGRU achieved better performance than the standalone BERT model for text classification tasks [21]. In addition, [22] shows that IndoBERTweet that combined with BiGRU for hate speech detection shows better performance than IndoBERTweet with LSTM.

Previous studies have explored several approaches for Indonesian cyberbullying and hate speech detection. CNN-LSTM has been applied to Indonesian cyberbullying detection, while IndoBERT has demonstrated strong performance for the same task [14][15]. However, those approaches do not specifically leverage a language model pretrained on Indonesian social media text. IndoBERTweet was developed specifically for Indonesian social media data and has demonstrated competitive performance compared with IndoBERT and mBERT in Indonesian text classification tasks [16][17]. Furthermore, hybrid architectures combining IndoBERTweet with recurrent neural networks have also shown promising results. For example, IndoBERTweet-BiLSTM has been applied to Indonesian hate speech detection on Twitter and outperformed the IndoBERTweet-CNN model [18]. Similarly, a previous study reported that an IndoBERTweet-BiGRU model performed better than BERT-BiGRU for hate speech detection [22]. Despite these findings, the application of IndoBERTweet-BiGRU specifically to Indonesian cyberbullying detection remains underexplored. Moreover, existing studies have primarily investigated either IndoBERT-based models or IndoBERTweet-based architectures using other recurrent or convolutional components, leaving the effectiveness of the IndoBERTweet-BiGRU combination for cyberbullying detection insufficiently investigated. Therefore, this study investigates an IndoBERTweet-BiGRU hybrid model for Indonesian cyberbullying detection.

This research also proposes an automated labeling approach, as the labeling is a crucial step in developing an accurate cyberbullying detection model. A previous study introduced an automated multi-stage labelling approach that combines a Generative Large Language Model (Gen LLM) with VADER lexicon. The results demonstrated that this approach improved the quality of the generated labels [15]. However, several lexicon-based sentiment analysis methods,

which have similar characteristics with VADER, have been reported to achieve lower accuracy than Gen LLM-based approaches [23]. Therefore, this research proposes a multi-stage labeling approach using two different Gen LLM models. This approach is expected to improve annotation consistency and reduce noisy labels by retaining only samples on which both LLMs agree.

Raw data collected from X (formerly Twitter) often exhibit an imbalanced class distribution after the labeling process. Class imbalance can make it difficult for a model to learn effectively due to the limited number of samples in the minority class [24]. Various approaches have been proposed to address this issue, one of which is Focal Loss. Focal Loss is designed to improve the model's sensitivity to minority classes by focusing the training process on difficult-to-classify samples. One of its main advantages is its ability to handle class imbalance without modifying the original dataset, unlike oversampling and undersampling techniques [25]. This method has been successfully applied in several studies. For example, study [26] employed Focal Loss to address severe class imbalance in a text classification task, where it achieved higher precision than the oversampling approach. Therefore, this research also adopts Focal Loss as the loss function during model training, with the expectation that it will enable the model to learn minority class representations more effectively and achieve better overall performance.

Although existing studies have demonstrated promising results in Indonesian cyberbullying and hate speech detection, several research gaps remain. First, while IndoBERT has been investigated for Indonesian cyberbullying detection and IndoBERTweet-BiGRU has shown promising performance for hate speech detection, the effectiveness of the IndoBERTweet-BiGRU architecture for Indonesian cyberbullying detection has not been sufficiently investigated. Second, existing automated labeling approaches for Indonesian cyberbullying datasets have relied on combinations of Gen LLMs and lexicon-based methods, whereas agreement-based annotation using two independent Gen LLMs remains underexplored. Third, the effectiveness of combining Focal Loss with label distribution modification for addressing severe class imbalance in Indonesian cyberbullying datasets has not been comprehensively evaluated. To address these gaps, this study contributes a novel framework that integrates an IndoBERTweet-BiGRU hybrid architecture, a dual-Gen-LLM agreement-based labeling strategy, and Focal Loss with label distribution modification for Indonesian cyberbullying detection.

II. METHODS

This section describes the methodology employed in this study. The research process consists of several stages, including data collection, data preprocessing, automated data labeling, model development, model training, and performance evaluation. The overall workflow of the proposed methodology is presented in Figure 1 below.

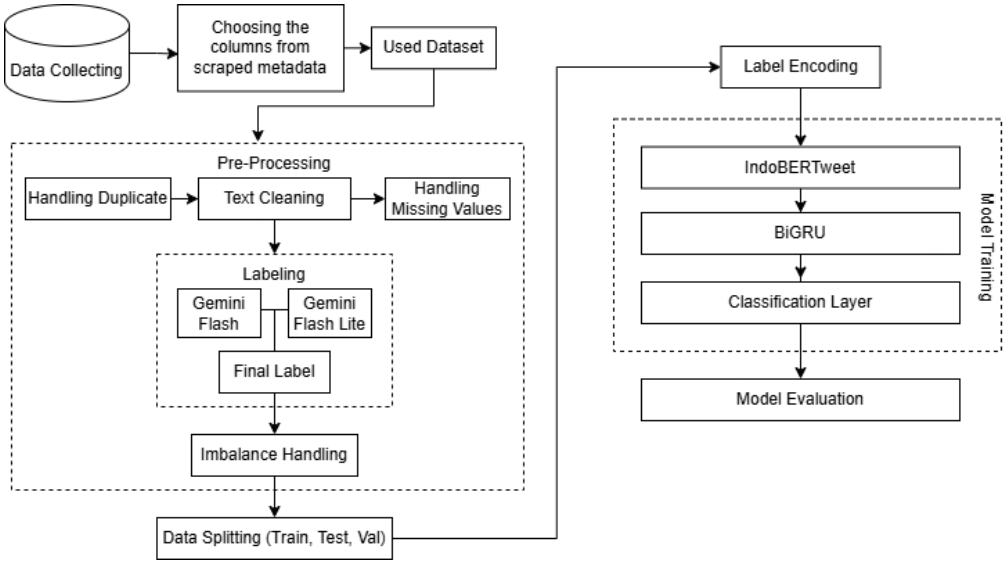

Figure 1. Research Methodology Flow

A. Data Collecting and Selecting Columns

The data were collected from X using the Scweet library available on GitHub [27]. Data collection was performed by scraping posts that contain several keywords commonly associated with cyberbullying. These keywords were grouped into several categories, including rude or offensive words (e.g., *"anjing", "bego", "bodoh"*), hate speech which targeting specific ethnicities, groups, or communities (e.g., *"LGBT hama", "kafir"*), and body shaming (e.g., *"gendut", "jelek", "item"*). The scraped dataset contained seven metadata fields, such as "tweet_id", "text", "username", "replies", "retweets", "likes", and "lang". A total of 27,634 posts were successfully collected and stored in CSV format.

Regarding ethical considerations, the study used publicly available posts collected from X solely for research purposes. Usernames and other personally identifiable information were not used as model inputs or reported in the analysis. The collected data were handled solely for research purposes and were not redistributed as part of the dataset.

B. Pre-Processing

In this stage, several preprocessing steps were performed to prepare the dataset for model training. First, duplicate entries were identified and removed. Next, a text cleaning process was applied, which included removing URLs, punctuation, emoji, and irrelevant symbols. Subsequently, all text was converted to lowercase. However, words written entirely in uppercase were preserved because they may indicate stronger emotional or aggressive expressions in social media posts. The @ symbol was also retained because user mentions were used to identify the target of cyberbullying during the noise-handling process. Usernames prefixed with @ were replaced with the token "USER". Finally, any remaining invalid characters were converted into missing values (NaN)

Unlike many text classification studies, this research did not perform stop-word removal or text normalizaton. Stop-word removal was omitted because stop words may contain contextual information that is important to identifying cyberbullying. Likewise, text normalization was not applied because IndoBERTweet is specifically designed to handle informal Indonesian social media language, including slang, abbreviations, and other non-standard expressions commonly found in X posts. A previous study reported that omitting the normalization step when using IndoBERTweet for Indonesian cyberbullying dataset achieved better accuracy than applying text normalization [28]. Finally, the remaining missing values were removed to ensure that only valid text data were used in subsequent stages.

C. Labelling

In this study, an automated multi-stage labeling approach was proposed using Generative Large Language Models (Gen LLMs). Two different models, namely Gemini 2.5 Flash-lite and Gemini 2.5 Flash, were employed to generate the labels. First, the cleaned dataset was converted into JSON format by adding a column containing a predefined prompt. The prompt described the characteristics of each cyberbullying category, including hate speech, body shaming, sexual harassment, rude language, and non-cyberbullying. The definitions used in the prompts were adapted from previous studies, namely [29] for hate speech, [30] for sexual harassment, [31] for rude, vulgar, and offensive language, and [32] for body shaming. After the JSON file had been prepared, it was uploaded to Google Cloud Console, where Gemini 2.5 Flash-Lite and Gemini 2.5 Flash were used independently to label each post.

The labels generated by the two models were subsequently compared for each post. Only posts for which Gemini 2.5 Flash and Gemini 2.5 Flash-Lite assigned the same label were retained for the subsequent modeling stage. Posts with different labels between the two models were excluded from the final dataset to reduce potential labeling noise and improve annotation consistency. To validate the quality of the automatically generated labels, a subset of 500 posts was selected from the labeled dataset and manually annotated by a human annotator. The human annotations were then compared with the labels generated through the dual-LLM agreement process. Cohen's Kappa was used to measure the agreement between the automated labels and the human annotations.

D. Imbalanced Data Handling

To address the class imbalance problem, Focal Loss was adopted as the loss function during model training. In addition, several experiments were conducted to investigate the impact of different label configuration on model performance. These experiments were designed to evaluate whether modifying the label distribution could improve the model's ability to learn minority classes.

E. Data Splitting and Label Encoding

The dataset was first divided into 80% training data and 20% testing data. The testing portion was then split equally into validation and testing sets. Consequently, the final dataset distribution consisted of 80% training data, 10% validation data, and 10% testing data. After the data splitting process, label encoding was applied to convert the categorical class labels into numerical values that could be processed by the model.

G. IndoBERTweet

IndoBERTweet is a pre-trained language model developed form IndoBERT and further pre-trained on Indonesian X (Twitter) data. Compared with IndoBERT, IndoBERTweet incorporates 14,584 additional WordPiece vocabularies that specifically designed to represent informal language that commonly used on social media. Consequently, the model is capable of understanding various characteristics of Indonesian social media text, including abbreviations, slang, informal expressions, dan non-standard sentence structures [16]. In this research, IndoBERTweet was employed as the embedding layer to generate contextual word representations before they were passed to the BiGRU layer. Since IndoBERTweet is based on BERT architecture, the overall BERT architecture is illustrated in Figure 2 below.

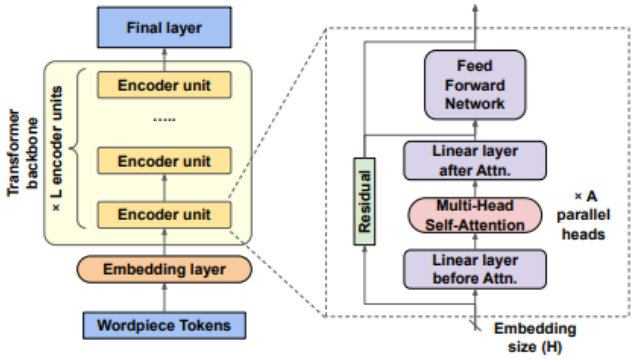

Figure 2. BERT Model Architecture

H. Bi-Directional Gated Recurrent Unit (BiGRU)

This study applies Bi-Directional Gated Recurrent Unit (BiGRU) to process the data after IndoBERTweet embedding layer. BiGRU consists GRU with backward and forward network. The GRU architecture consists of two gating mechanisms, which are the reset gate (r_t) and the update gate (z_t). These gates regulate how much information from the previous hidden state is retained or discarded, as well as how much new information is incorporated into the current hidden state. The mathematical formulations of the reset and update gates are presented in Equations (1) and (2). Then, the candidate hidden state (g_t) is calculated by the Eq. (3) using the reset gate and the hidden state of the previous time step. Finally, the hidden state at the current time step (h_t) is obtained by combining the candidate hidden state with the update gate.

$$r_t = \sigma\left(W_r x_t + U_r h_{t-1}\right) \quad (1)$$

$$z_t = \sigma\left(W_z x_t + U_z h_{t-1}\right) \quad (2)$$

$$g_t = \tanh\left(W_r x_t + U_r h_{t-1}\right) \quad (3)$$

$$h_t = z_t \otimes h_{t-1} + (1 - z_t) \otimes g_t \quad (4)$$

BiGRU employs two GRU networks to process the input sequence in both forward and backward directions. The outputs produced by these two networks are then concatenated to form the final representation, as defined in Equation (5). [33]

$$H_t = \left[\overrightarrow{h}_t, \overleftarrow{h}_t\right] \quad (5)$$

To provide a clearer understanding of the proposed sequence modeling architecture, the overall structure of the GRU and BiGRU network used in this study is illustrated in Figure 3 and Figure 4 below.

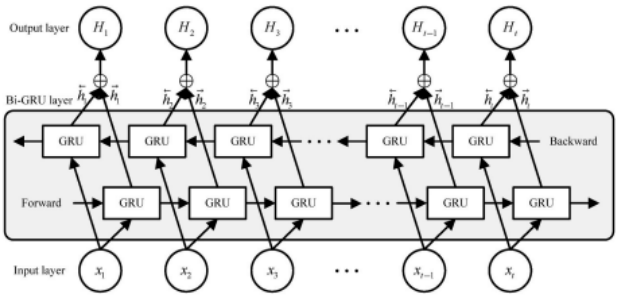

Figure 3. Bi-Directional Gated Recurrent Unit Architecture

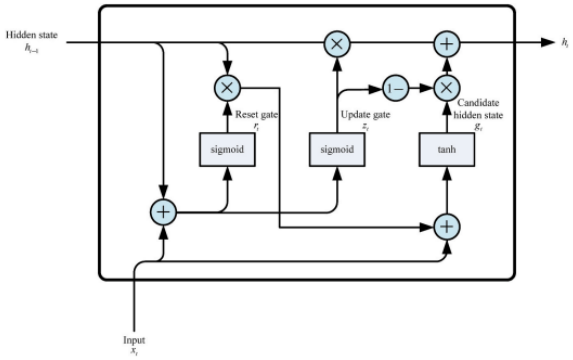

Figure 4. Gated Recurrent Unit Architecture

I. Focal Loss

Focal loss is a loss function that designed to address the class imbalance problem in classification tasks. Focal loss incorporates a modulating factor that reduces the contribution of samples that are already classified correctly with high confidence. The mathematical formulation of Focal Loss is presented in Eq. (6) below.

$$FL = -(1 - p_t)^{\gamma} \log(p_t) \quad (6)$$

The model places greater emphasis on misclassified samples, preventing the majority class from dominating the learning process. The degree of this effect is controlled by the focusing parameter, denoted as γ. Higher values of γ further decrease the influence of easy samples, so that the model can concentrate more on challenging examples during training [34].

J. Model Evaluation

Model evaluation is conducted to assess the effectiveness of the proposed model and compare its performance with other experimental configurations. For classification tasks, the most commonly used evaluation metrics are accuracy, precision, recall, and F1-score. These metrics provide a comprehensive assessment of the model's classification performance from different perspectives. The mathematical formulations of accuracy, precision, recall, and F1-score are presented in Equations (7), (8), (9), and (10), respectively [35].

$$\text{Accuracy} = \frac{\text{total of correct prediction}}{\text{total data}} \times 100 \quad (7)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (8)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (9)$$

$$\text{F1} - \text{Score} = \frac{2 \times P \times R}{P+R} \quad (10)$$

where TP, FN FP and TN denote True Positive, False Negative, False Positive, and True Negative as defined in the confusion matrix. The detailed structure of confusion matrix is illustrated in Figure 5.

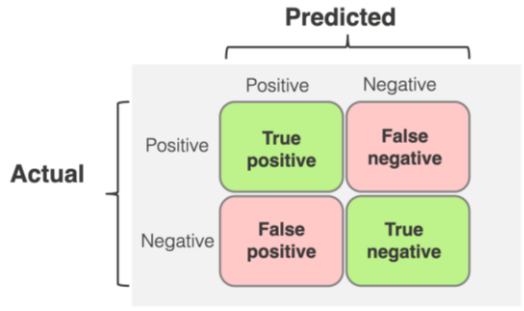

Figure 5. Confusion Matrix

III. Result and Discussion

A. Data Collecting and Selecting Columns

The data were collected from X (formerly Twitter) in May 2026 through a scraping process using several keywords related to rude and vulgar language, body shaming, hate speech, and sexual harassment. The scraping process was performed using the Scweet library available on GitHub, with a maximum of 1,700 posts collected for each keyword. Since the number of posts associated with each keyword varied, the temporal coverage of the collected posts also varied across keywords. A total of 27,634 posts were successfully collected and stored in CSV format. The dataset contained several metadata fields, including tweet_id, date, username, text, likes, retweets, replies, keyword, and lang. To ensure that the dataset contained only Indonesian-language posts, the data were filtered based on the lang field, retaining only posts with the language code "id". After the language filtering process, only the text column was retained as the input for subsequent preprocessing and training, while the other metadata fields were not used as model inputs. An example of the resulting dataset is presented in Table 1.

TABLE I
SELECTED DATASET COLUMN

	Text
1	*Gimanapun baiknya lo kalau orang emang ga seneng sama kita tuh emang susah tetap aja dia ga bakal seneng Jadi ya udah biarin aja nanti juga bakal mati sendiri*
2	*Wkwkwkwk mati aja lo semua a*****g. Jijik.*
3	*@r**purw***h Capek capek bikin negara nonblok, malah kek gini, a*g a*g presiden g****k bodoh mati aja lo prabowo*
4	*Perempuan sialan. Mati aja sono ikut mama lo.*
5	*@gongjunanaa @tanyarlfes yang harusnya digebukin itu kaum lgbt, jangan normalisasi penyimpangan t***l lu a*****g b**o*

NB: The (*) symbol representing censored inappropriate words and mentioned usernames.

B. Pre-processing

Pre-processing techniques were applied by removing duplicate entries and cleaning the text, which included removing URLs, punctuation (while retaining the @ symbol to preserve user mentions), emojis, and irrelevant symbols. Subsequently, all text was converted to lowercase, except for words written entirely in uppercase to preserve expressions of emphasis or anger. Next, usernames were replaced with the token "USER". Finally, the remaining missing values were removed. After the pre-processing steps were applied, the data has 19,688 rows. Table 2 presents a comparison of the dataset before and after pre-processing.

TABLE II
BEFORE AND AFTER PRE-PROCESSING

Before Pre-Processing	After Pre-Processing
*@th**8*8 Emaklo g****k idi*t..termul...mati aja lo*	*USER emaklo g****k id**t termul mati aja lo*
Gimanapun baiknya lo kalau orang emang ga seneng...	*gimanapun baiknya lo kalau orang emang ga seneng...*
㊗ pokoknya nentto wlw menghadapi hari senin https://...	*pokoknya nentto wlw menghadapi hari senin*
*@g*ng**n*a @t*nya***es yang harusnya digebukin itu kaum lgbt, jangan normalisasi penyimpangan t***l lu a***ng b**o*	*USER USER yang harusnya digebukin itu kaum lgbt jangan normalisasi penyimpangan tol*l lu a***ng b**o*

C. Labeling

An automated multi-stage labeling process was applied using two Generative Large Language Models (Gemini 2.5 Flash-Lite and Gemini 2.5 Flash). The cleaned dataset was converted into JSON format by adding a prompt column containing the labeling instructions. Each post was assigned to one of five categories that consists of Non-Cyberbullying, Rude and Vulgar Words, Sexual Harassment, Body Shaming, and Hate Speech. After the labeling process was completed, the resulting dataset consisted of two columns, which are "Text" and "Label". Sample entries from the labeled dataset are presented in Table 3.

TABLE III
TEXT AND LABEL EXAMPLES

Text	Label
*INDOSAT K****L B**I A****G USER USER HARI SINYAL...*	Rude and Vulgar Words
*USER yapp tiap hari gw s***ng k****l nya pagi...*	Sexual Harassment
gua pacul tuh si item gendut jelek	Body Shaming
USER kalo lo nyari yang kenceng ya vivo yd pro	Non-Cyberbulying
*USER USER yang harusnya digebukin itu kaum lgbt jangan normalisasi penyimpangan t***l lu a***ng b*go*	Hate Speech

Table 5 presents a comparison of the label distributions produced by Gemini 2.5 Flash-Lite, Gemini 2.5 Flash, and their intersection. Only instances for which both models

produced the same label were retained for the subsequent modeling stage. As a result, the final dataset used for model training consisted of 16,574 samples.

TABLE IV
LABELING DISTRIBUTION COMPARISON

Labeling Method	Label Distribution				
	RVW	SH	BS	HS	NCB
Flash	16888	1670	1339	1366	2770
Flash Lite	12825	2373	1061	1510	6255
Flash + Flash Lite	11769	1023	699	937	2146

To evaluate the reliability of the automated labeling process, 500 samples, representing approximately 3% of the final dataset, were selected using stratified random sampling based on the class labels. This sampling strategy was used to ensure that the validation subset maintained the class distribution of the final dataset. The selected samples were manually annotated by a human annotator and compared with the labels generated through the dual-LLM agreement process. The agreement between the automated and human annotations was evaluated using Cohen's Kappa, which resulted in a coefficient of 0.7951. According to the interpretation proposed by [36], a Cohen's Kappa value between 0.61 and 0.80 indicates substantial agreement. Therefore, the obtained coefficient of 0.7951 indicates substantial agreement between the automated labeling approach and the human annotations, supporting the reliability of the generated labels.

D. Imbalance Handling

As shown in Table 4, the Body Shaming and Hate Speech categories contain significantly fewer samples than the other categories in the intersection of the Flash-Lite and Flash model outputs that indicating severe class imbalance in the dataset. To address this issue, Focal Loss was adopted as the loss function during the model training process. In addition, several experiments were conducted by modifying the label configuration. First, the vulgar expressions were separated from the Rude and Vulgar Words category and merged into a new category named Sexual Harassment and Vulgar Words. Furthermore, the Body Shaming and Hate Speech categories were merged into a single category to reduce the degree of class imbalance. Although Body Shaming and Hate Speech represent different types of harmful language and have distinct linguistic characteristics, both categories contained relatively few samples compared with the majority classes. Therefore, their combination was evaluated as an experimental label configuration to increase the representation of minority instances and examine whether reducing class imbalance through label merging could improve the model's ability to learn minority-class representations. The resulting class distributions after these modifications are presented in Tables 5 and 6.

TABLE V
MODIFIED CLASS DISTRIBUTION (EXPERIMENT 1)

Label	Counts
Rude Words	1077
Sexual Harassment and Vulgar Words	2615
Non Cyberbullying	2146
Hate Speech	937
Body Shaming	699

TABLE VI
MODIFIED CLASS DISTRIBUTION (EXPERIMENT 2)

Label	Counts
Rude Words	1077
Sexual Harassment and Vulgar Words	2615
Non Cyberbullying	2146
Body Shaming and Hate Speech	1636

E. Model Training

Several hyperparameter configurations were evaluated to determine the optimal settings for the proposed model. The configuration that achieved the best performance was selected for the final model training. The selected hyperparameters are summarized in Table 7.

TABLE VII
HYPERPARAMETER CONFIGURATION

Hyperparameters	Value / Type
Learning Rate	2e-5
Batch Size	16
BiGRU Hidden Size	128
Epoch	4
Focal Loss' Gamma	2.0
Drop Out Layer	0.3
Optimizer	Adam

Using those selected hyperparameter settings, the first experiment was conducted using the baseline labeling configuration consisting of five classes that consist of *Rude and Vulgar Words*, *Body Shaming*, *Hate Speech*, *Sexual Harassment*, and *Non-Cyberbullying*. Focal Loss was employed as the loss function during model training. Subsequently, two additional experiments were performed using the modified labeling configurations presented in Tables 5 and 6. The training, validation, and testing accuracies obtained from each experiment are summarized in Table 8 below.

TABLE VIII
COMPARISON OF IMBALANCED HANDLING PERFORMANCE DURING TRAINING

Labeling	Result (Accuracy)		
	Train	Validation	Test
Baseline	0.97	0.91	0.92
Experiment 1	0.97	0.87	0.86
Experiment 2	0.98	0.93	0.93

Based on Table 8 above, the second experiment achieved the best overall performance in terms of the gap between the training, validation, and testing accuracies. The validation and testing accuracies were only 0.05 lower than the training accuracy, indicating that the model obtained from the second experiment generalized well while maintaining consistent performance. The baseline imbalanced handling method also demonstrated competitive performance, although its accuracy was slightly lower than that of the second experiment. In contrast, the first experiment exhibited a considerably larger gap of approximately 0.10 between the training accuracy and both validation and testing accuracies. This result suggests that the model generalized less effectively to unseen data and may have experienced a higher degree of overfitting than the other experimental imbalanced data handling method.

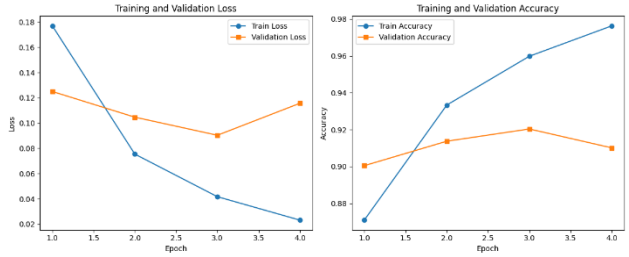

Figure 6. Training and Validation Performance during Training (Baseline)

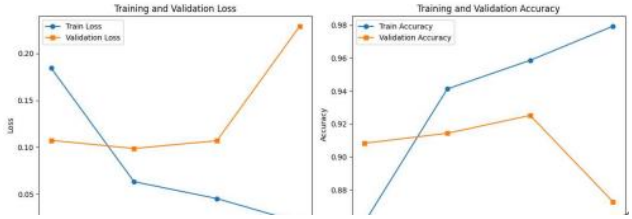

Figure 7. Training and Validation Performance during Training (Exp. 1)

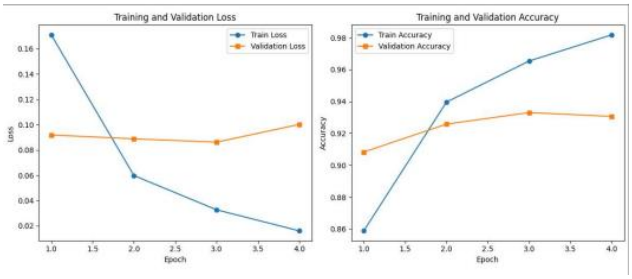

Figure 8. Training and Validation Performance during Training (Exp. 2)

Figures 6, 7, and 8 present the training and validation accuracy and loss curves for each experimental configuration. As shown in Figure 6, both the training and validation losses decreased from the first to the third epoch. However, at the fourth epoch, the validation loss increased slightly while the training loss continued to decrease. This trend suggests the onset of slight overfitting, although the gap between the two curves remained relatively small.

Figure 7 shows that the training loss continuously decreased from the first to the fourth epoch. In contrast, the validation loss began to increase slightly after the second epoch and rose more noticeably at the fourth epoch. A similar trend can also be observed in the validation accuracy, which dropped considerably during the final epoch. These results indicate a higher degree of overfitting than that observed in Figure 6, suggesting that the model generalized less effectively to unseen data.

Figure 8 shows that the training loss continued to decrease throughout the training process, whereas the validation loss increased only slightly between the third and fourth epochs. Likewise, the training accuracy consistently improved, while the validation accuracy remained relatively stable with only a slight decrease at the final epoch. The gaps between the training and validation loss, as well as between the training and validation accuracy, remained relatively small throughout the training process. These results indicate that the third experimental configuration achieved the most stable learningbehavior and the best generalization performance among all experiments.

F. Evaluations of Experiments

After the model training process was completed, the model was evaluated using a confusion matrix to analyze the distribution of correctly classified and misclassified instances for each class across all experimental configurations.

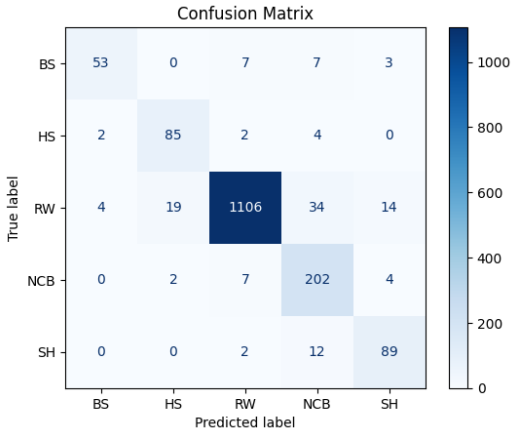

Figure 9. Baseline Labeling Confusion Matrix

Figure 9 shows the confusion matrix of the baseline labeling configuration, consisting of Rude and Vulgar Words (RW), Sexual Harassment (SH), Body Shaming (BS), Hate Speech (HS), and Non-Cyberbullying (NCB). Overall, most samples were correctly classified, as indicated by the dominant diagonal values. The Rude and Vulgar Words class achieved the highest number of correct predictions (1106 samples). Most misclassifications occurred between Body Shaming and Rude and Vulgar Words. This indicates that these categories share similar offensive lexical patterns. In addition, several Sexual Harassment samples were predicted as Non-Cyberbullying. This is suggesting that implicit sexual expressions remain could be difficult to identify.

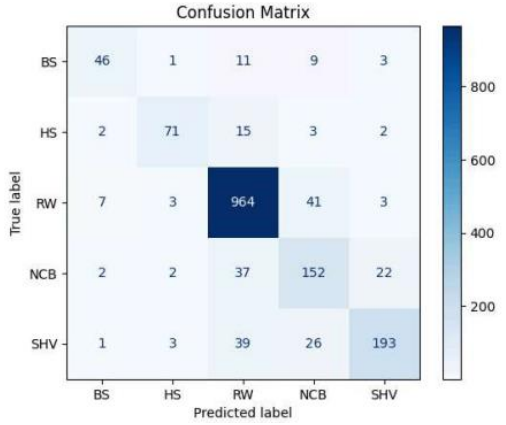

Figure 10. Experiment 1 Confusion Matrix

Figure 10 shows the confusion matrix of the first experiment labeling configuration that consist of Body Shaming (BS), Hate Speech (HS), Rude Words (RW), Non Cyberbullying (NCB), Sexual Harassment and Vulgar Words (SHV). Based on the figure, the diagonal values remain dominant. This indicates that most samples were correctly classified. The Rude Words class achieved the highest number of correct predictions (964 samples). However, more misclassifications were observed than in the baseline configuration, particularly between Rude Words and Non-Cyberbullying, as well as between Sexual Harassment and Vulgar Words and Rude Words. This suggests that separating vulgar words from the original Rude and Vulgar Words category increased the lexical similarity between these classes and making them more difficult for the model to distinguish.

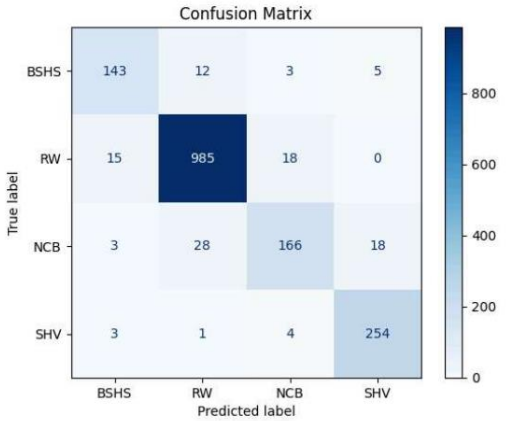

Figure 11. Experiment 2 Confusion Matrix

Figure 11 shows the confusion matrix of the second experimental labeling configuration, consisting of Body Shaming and Hate Speech (BSHS), Rude Words (RW), Non-Cyberbullying (NCB), and Sexual Harassment and Vulgar Words (SHV). Overall, the confusion matrix exhibits more dominant diagonal values than the previous experiments, indicating improved classification performance across all classes. The Rude Words and Sexual Harassment and Vulgar Words classes achieved the highest number of correct predictions, with 985 and 254 correctly classified samples, respectively. Most misclassifications occurred between Rude Words and Non-Cyberbullying. These results suggest that merging the Body Shaming and Hate Speech categories reduced class imbalance and enabled the model to learn minority class representations more effectively.

TABLE IX
CLASSIFICATION REPORT OF EXPERIMENTS

Labeling Method	Imbalance Handling	Classification Report			
		Acc	F1	Pre	Rec
Baseline	Focal Loss	0.93	0.87	0.85	0.88
Experiment 1		0.86	0.79	0.82	0.76
Experiment 2		**0.93**	**0.90**	**0.90**	**0.90**
Baseline	Cross Entropy	0.93	0.85	0.85	0.81
Experiment 1		0.91	0.90	0.90	0.89
Experiment 2		0.91	0.88	0.89	0.89
Baseline	Class Weights	0.88	0.80	0.76	0.85
Experiment 1		0.91	0.88	0.87	0.89
Experiment 2		0.92	0.89	0.89	0.87
Baseline	Random Over-sampling	0.93	0.85	0.88	0.83
Experiment 1		0.93	0.87	0.90	0.85
Experiment 2		0.93	0.89	0.90	0.89

Table 9 presents the classification performance of all labeling configurations across four imbalance-handling approaches, namely Focal Loss, Cross Entropy, Class Weights, and Random Oversampling, evaluated using accuracy, macro F1-score, precision, and recall. Overall, the effectiveness of each approach varied across labeling configurations, with the highest performance achieved by Experiment 2 using Focal Loss, which obtained an accuracy of 0.93 and a macro F1-score, precision, and recall of 0.90.

Focal Loss produced the lowest performance in Experiment 1, with an accuracy of 0.86 and a macro F1-score of 0.79. This result may be attributed to the highly imbalanced label distribution, where the Hate Speech and Body Shaming classes contained only hundreds of instances. The limited number of samples may have restricted the model's ability to learn representative patterns from these minority classes. In Experiment 2, combining these classes increased their number of samples to the thousands and resulted in a less imbalanced distribution, contributing to the improvement in macro F1-score from 0.79 to 0.90.

Class Weights and Random Oversampling also improved performance across the labeling configurations. Class Weights achieved its best result in Experiment 2, increasing the macro F1-score from 0.80 in the Baseline to 0.89. Random Oversampling produced relatively stable results, with Experiment 2 achieving an accuracy of 0.93 and a macro F1-score of 0.89. Overall, Experiment 2 consistently provided strong performance across the imbalance-handling approaches, with the combination of Experiment 2 and Focal Loss achieving the highest macro F1-score of 0.90. Therefore, this configuration was selected for the subsequent model comparison as presented in Table 10.

TABLE X
COMPARISON OF PERFORMANCE ACROSS DIFFERENT MODELS

Model	Classification Report			
	Acc	**F1**	**Pre**	**Rec**
IndoBERT	0.90	0.85	0.86	0.84
IndoRoBERTa	0.91	0.86	0.87	0.86
IndoBERTweet	0.93	0.87	0.89	0.88
IndoBERT + BiGRU	0.88	0.80	0.85	0.78
IndoRoBERTa + BiGRU	0.91	0.88	0.87	0.88
IndoBERTweet + BiGRU	**0.93**	**0.90**	**0.90**	**0.90**

Table 10 presents the classification performance of pretrained language models with and without BiGRU. The results show that the effect of BiGRU varied depending on the underlying pretrained language model. IndoBERT achieved an accuracy and macro F1-score of 0.90 and 0.85, respectively, but its performance decreased to 0.88 accuracy and 0.80 macro F1-score after incorporating BiGRU. In contrast, IndoRoBERTa improved from a macro F1-score of 0.86 to 0.88, with recall increasing from 0.86 to 0.88.

The most notable improvement was observed in IndoBERTweet. Without BiGRU, IndoBERTweet achieved an accuracy of 0.93 and a macro F1-score of 0.87. After incorporating BiGRU, its macro F1-score increased to 0.90, while precision and recall both reached 0.90. This result indicates that the effect of BiGRU depends on the characteristics of the underlying pretrained model. The stronger performance of IndoBERTweet may also be attributed to its suitability for social media text, which is consistent with the characteristics of the X dataset used in this study. The addition of BiGRU may further improve the representation of sequential and contextual information in the text.

Overall, IndoBERTweet combined with BiGRU achieved the highest performance among the evaluated models, with an accuracy of 0.93 and a macro F1-score, precision, and recall of 0.90. Therefore, the combination of IndoBERTweet and BiGRU was selected as the best-performing architecture in this study.

G. Further Analysis

Although the evaluation metrics provide an overall assessment of model performance, analyzing individual prediction examples offers a deeper understanding of how each labeling configuration handles ambiguous expressions. Table 10 below presents several examples of model predictions on unseen texts that trained with IndoBERTweet and BiGRU using Focal Loss.

TABLE XI
COMPARISON OF PREDICTION RESULTS ON UNSEEN TEXTS

Method	Text	Pred	True
Baseline	*sebenernya anjing tuh makannya apa sih*	NCB	NCB
	*kafir si***n*	HS	HS
Exp. 1	*sebenernya anjing tuh makannya apa sih*	RW	NCB
	*kafir si***n*	NCB	HS
Exp. 2	*sebenernya anjing tuh makannya apa sih*	NCB	NCB
	*kafir si***n*	BSHS	BSHS

Based on Table 10, for the first example, the baseline configuration correctly classified the tweet as Hate Speech, whereas Experiment 1 misclassified it as Rude Words. This indicates that Experiment 1 still had difficulty distinguishing whether the word *"anjing"* was used as an offensive expression or simply referred to the animal. This highlights that the experiment 1 still difficult to understand contextual meaning. Next, the experiment 2 successfully classified the tweet correctly. This indicates that the experiment 2 is giving an improved ability to interpret ambiguous contexts. In the second example, both the Baseline and Experiment 2 correctly classified the tweet as Hate Speech and Body Shaming and Hate Speech (BSHS), respectively, where the latter represents the merged label used in Experiment 2. However, Experiment 1 incorrectly classified the tweet as Non-Cyberbullying. Overall, these examples demonstrate that the proposed labeling configuration in Experiment 2 reduced misclassifications in ambiguous cases and improved the model's contextual understanding.

IV. CONCLUSION

The proposed IndoBERTweet–BiGRU model with automated multi-stage labeling using Gemini 2.5 Flash and Gemini 2.5 Flash-Lite achieved the best overall performance when trained with Focal Loss and the modified label distribution from Experiment 2. This configuration, consisting of Rude Words, Sexual Harassment and Vulgar Words, Body Shaming and Hate Speech, and Non-Cyberbullying, achieved an accuracy of 0.93 and a macro F1-score, precision, and recall of 0.90. Among the evaluated imbalance-handling strategies, Focal Loss with Experiment 2 provided the highest overall performance, while Class Weights and Random Oversampling also improved classification performance compared with their respective baseline configurations. Furthermore, the comparison of pretrained language models showed that IndoBERTweet–BiGRU outperformed the other evaluated architectures, indicating its suitability for Indonesian social media text. These findings demonstrate the effectiveness of combining automated multi-stage labeling, appropriate imbalance handling, and a social-media-oriented pretrained language

model with BiGRU for multi-class Indonesian cyberbullying detection.

This study has several limitations. First, the proposed multi-stage labeling approach using Gemini 2.5 Flash and Gemini 2.5 Flash-Lite was validated by only one human annotator, which may introduce potential annotation bias. Future studies should involve multiple domain experts or benchmark datasets to strengthen labeling reliability. Second, the dataset was collected exclusively from X and consists only of Indonesian-language content, limiting the evaluation of cross-platform and multilingual generalization. Although the tweets span different dates, temporal generalization was not explicitly evaluated using a separate time period. Future research should therefore evaluate the model across different time periods and platforms, while exploring multilingual and multimodal approaches to improve its robustness and applicability

REFERENCES

[1] Haseeba, "The Impact of Social Media on Social Interactions: A Sociological Perspective," 2023.

[2] R. D. Setyaningsih and F. A. Nur, "Social revolution: The impact of social media in our lives," *Symposium of Literature, Culture, and Communication (SYLECTION) 2022*, vol. 3, no. 1, p. 59, Nov. 2023, doi: 10.12928/sylection.v3i1.13933.

[3] I. Putri, "Media Sosial Sebagai Media Pergeseran Interaksi Sosial Remaja," *Jurnal Ilmu Komunikasi Balayudha*, vol. 2, no. 2, p. 1, Dec. 2022.

[4] S. Kemp, "Digital 2025: Indonesia," Data Reportal. Accessed: Jun. 09, 2025. [Online]. Available: https://datareportal.com/reports/digital-2025-indonesia?rq=indonesia%202025

[5] S. Kemp, "Digital 2024: Indonesia," Data Reportal. Accessed: Jun. 09, 2025. [Online]. Available: https://datareportal.com/reports/digital-2024-indonesia?rq=indonesia%202024

[6] L. Fazry and N. C. Apsari, "Pengaruh Media Sosial Terhadap Perilaku Cyberbullying Di Kalangan Remaja," *Jurnal Penelitian dan Pengabdian Kepada Masyarakat (JPPM)*, vol. 2, no. 2, p. 272, Aug. 2021, doi: 10.24198/jppm.v2i2.34679.

[7] Y. Zhang, B. Zhou, Y. Hu, and K. Zhai, "From Individual Expression to Group Polarization: A Study on Twitter's Emotional Diffusion Patterns in the German Election," *Behavioral Sciences*, vol. 15, no. 3, p. 360, Mar. 2025, doi: 10.3390/bs15030360.

[8] M. E. Rao and D. M. Rao, "The Mental Health of High School Students During the COVID-19 Pandemic," *Front. Educ. (Lausanne).*, vol. 6, Jul. 2021, doi: 10.3389/feduc.2021.719539.

[9] S. Bansal, N. Garg, J. Singh, and F. Van Der Walt, "Cyberbullying and mental health: past, present and future," *Front. Psychol.*, vol. 14, Jan. 2024, doi: 10.3389/fpsyg.2023.1279234.

[10] M. T. Chamizo-Nieto and L. Rey, "Cybervictimization and suicidal ideation in adolescents: A prospective view through gratitude and life satisfaction," *J. Health Psychol.*, vol. 28, no. 7, pp. 620–632, Jun. 2023, doi: 10.1177/13591053221140259.

[11] I. Planellas Kirchner and C. Calderon Garrido, "Do Cybervictimizations Predict Suicide-Related Behaviors in Adolescents? Mediating Role of the 'Escaping' Coping Strategy," *J. Interpers. Violence*, vol. 40, no. 5–6, pp. 1015–1036, Mar. 2025, doi: 10.1177/08862605241256384.

[12] P. S. Reddy, "Natural Language Processing (NLP) and Understanding," *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, 2025, doi: 10.15662/IJAREEIE.2025.1402022.

[13] S. N. Afandy, K. M. Hindrayani, and A. T. Damaliana, "Comparison of the Effectiveness IndoBERT and mBERT for Sentiment Analysis of SME Customer Reviews," *bit-Tech*, vol. 8, no. 3, pp. 3383–3394, Apr. 2026, doi: 10.32877/bt.v8i3.3501.

[14] A. J. Andika, Y. Kristian, and E. I. Setiawan, "Deteksi Komentar Cyberbullying Pada YouTube Dengan Metode Convolutional Neural Network – Long Short-Term Memory Network (CNN-LSTM)," *Teknika*, vol. 12, no. 3, pp. 183–188, Oct. 2023, doi: 10.34148/teknika.v12i3.677.

[15] Y. D. Novandian *et al.*, "IndoBERT-based Indonesian Cyberbullying Detection with Multi-stage Labeling," in *2024 International Seminar on Application for Technology of Information and Communication (iSemantic)*, IEEE, Sep. 2024, pp. 515–521. doi: 10.1109/iSemantic63362.2024.10762553.

[16] F. Koto, J. H. Lau, and T. Baldwin, "IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 10660–10668. doi: 10.18653/v1/2021.emnlp-main.833.

[17] F. Indriani, R. A. Nugroho, M. R. Faisal, and D. Kartini, "Comparative Evaluation of IndoBERT, IndoBERTweet, and mBERT for Multilabel Student Feedback Classification," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 8, no. 6, pp. 748–757, Dec. 2024, doi: 10.29207/resti.v8i6.6100.

[18] J. F. Kusuma and A. Chowanda, "Indonesian Hate Speech Detection Using IndoBERTweet and BiLSTM on Twitter," *JOIV : International Journal on Informatics Visualization*, vol. 7, no. 3, pp. 773–780, Sep. 2023, doi: 10.30630/joiv.7.3.1035.

[19] M. R. R. Rana, A. Nawaz, S. U. Rehman, M. A. Abid, M. Garayevi, and J. Kajanová, "BERT-BiGRU-Senti-GCN: An Advanced NLP Framework for Analyzing Customer Sentiments in E-Commerce," *International Journal of Computational Intelligence Systems*, vol. 18, no. 1, p. 21, Feb. 2025, doi: 10.1007/s44196-025-00747-1.

[20] K. L. Tan, C. P. Lee, and K. M. Lim, "RoBERTa-GRU: A Hybrid Deep Learning Model for Enhanced Sentiment Analysis," *Applied Sciences*, vol. 13, no. 6, Mar. 2023, doi: 10.3390/app13063915.

[21] A. S. Talaat, "Hybrid Deep Learning Models for Text Classification: Performance Evaluation of TriDistilBERT and BiGRU Architectures," *Journal of Computer Science*, vol. 21, no. 9, pp. 1983–1992, Sep. 2025, doi: 10.3844/jcssp.2025.1983.1992.

[22] A. Muzakir, K. Adi, and R. Kusumaningrum, "Short Text Classification Based on Hybrid Semantic Expansion and Bidirectional GRU (BiGRU) based Method to Improve Hate Speech Detection," *Revue d'Intelligence Artificielle*, vol. 37, no. 6, pp. 1471–1481, Dec. 2023, doi: 10.18280/ria.370611.

[23] F. Koto, T. Beck, Z. Talat, I. Gurevych, and T. Baldwin, "Zero-shot Sentiment Analysis in Low-Resource Languages Using a Multilingual Sentiment Lexicon," Feb. 2024.

[24] C. Padurariu and M. E. Breaban, "Dealing with Data Imbalance in Text Classification," *Procedia Comput. Sci.*, vol. 159, pp. 736–745, 2019, doi: 10.1016/j.procs.2019.09.229.

[25] A. C. Zahro *et al.*, "Fairer Public Complaint Classification on LaporGub: Integrating XLM-RoBERTa with Focal Loss for Imbalance Data," *sinkron*, vol. 9, no. 4, pp. 1850–1862, Oct. 2025, doi: 10.33395/sinkron.v9i4.15260.

[26] D. Jiang and J. He, "Text Semantic Classification of Long Discourses Based on Neural Networks with Improved Focal Loss," *Comput. Intell. Neurosci.*, vol. 2021, no. 1, Jan. 2021, doi: 10.1155/2021/8845362.

[27] EriSetyawan166, "ScraperTwitter," GitHub. Accessed: Jul. 07, 2026. [Online]. Available: https://github.com/EriSetyawan166/ScraperTwitter

[28] M. Y. Pratiwi, "Analisis Pengaruh Data Normalisasi dan Non Normalisasi pada Deteksi Cyberbullying Berbahasa Indonesia dengan Metode IndoBERT," Unversitas Lambung Mangkurat, Banjarmasin, 2023.

[29] Y. Xu and P. Trzaskawka, "Towards Descriptive Adequacy of Cyberbullying: Interdisciplinary Studies on Features, Cases and

Legislative Concerns of Cyberbullying," *Int. J. Semiot. Law*, vol. 34, no. 4, pp. 929–943, Sep. 2021, doi: 10.1007/s11196-021-09856-4.

[30] A. C. Ehman and A. M. Gross, "Sexual cyberbullying: Review, critique, & future directions," *Aggress. Violent Behav.*, vol. 44, pp. 80–87, Jan. 2019, doi: 10.1016/j.avb.2018.11.001.

[31] Y. W. Riyayanatasya and R. Rahayu, "Involvement of Teenage-Students in Cyberbullying on WhatsApp," *Jurnal Komunikasi Indonesia*, vol. 9, no. 1, Jun. 2020, doi: 10.7454/jki.v9i1.11824.

[32] L. H. Collantes, F. A. Saputra, P. R. Solikhah, T. C. Laksana, N. K. Yakti, and J. Tipagau, "Cyberbullying Body-Shaming Levels in Adolescence," *Bulletin of Social Informatics Theory and Application*, vol. 6, no. 2, pp. 111–119, Dec. 2022, doi: 10.31763/businta.v6i2.603.

[33] H. S. Jo, S. H. Lee, and M. G. Na, "Prediction of small-scale leak flow rate in LOCA situations using bidirectional GRU," *Nuclear Engineering and Technology*, vol. 56, no. 9, pp. 3594–3601, Sep. 2024, doi: 10.1016/j.net.2024.04.009.

[34] Z. Labd, S. Bahassine, and K. Housni, "AFL-BERT : Enhancing Minority Class Detection in Multi-Label Text Classification with Adaptive Focal Loss and BERT," *International Journal of Advanced Computer Science and Applications*, vol. 16, no. 7, 2025, doi: 10.14569/IJACSA.2025.0160749.

[35] C. Miller, T. Portlock, D. M. Nyaga, and J. M. O'Sullivan, "A review of model evaluation metrics for machine learning in genetics and genomics," *Frontiers in Bioinformatics*, vol. 4, Sep. 2024, doi: 10.3389/fbinf.2024.1457619.

[36] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33, no. 1, p. 159, Mar. 1977, doi: 10.2307/2529310.