ISSN 2355-0082

A Hierarchical Two Stage BERT Model for Cyberbullying Detection

Muhamad Syukron[1], Nayya Safitri Ramadani[2], Muhammad Ilham Firmansyah[3], Asi Emilia Putri[4]
Department of Statistics, Faculty of Sains and Data Analytics, Institut Teknologi Sepuluh Nopember
[1]muhamad.syukron@its.ac.id, [2]5003231040@student.its.ac.id,
[3]5003231150@students.its.ac.id, [4]5003231042@students.its.ac.id

Accepted 16 May 2026
Approved 28 June 2026

***Abstract*— Cyberbullying has become a significant social concern alongside the rapid expansion of social media usage, particularly in developing countries such as Indonesia. Online bullying transcends overt insults to include more nuanced manifestations, such as denigration and social exclusion, which are frequently subtle and context-dependent. This diversity poses substantial challenges for automated detection systems, particularly when cyberbullying is conceptualized as a singular, homogeneous category. This study addresses the challenge of detecting cyberbullying in the Indonesian language by proposing a two-stage classification framework utilizing the IndoBERT model. In the initial phase, a binary classification model is utilized to distinguish between content that constitutes bullying and that which does not. In the subsequent phase, tweets identified as cyberbullying are further classified into three distinct categories: harassment, denigration, and exclusion. The model is trained utilizing a curated dataset comprising Indonesian tweets, which have been collected from publicly accessible social media content. The findings demonstrate that the proposed framework exhibits robust and consistent performance, particularly in the domain of fine-grained cyberbullying classification, achieving an overall accuracy of approximately 92% in the second stage. The results indicate that hierarchical classification can achieve robust performance and significantly contribute to the identification of cyberbullying.**

***Index Terms*— Cyberbullying; IndoBERT; Two-Stage Classification; X (Twitter)**

I. INTRODUCTION

The swift progression of digital technology has led to a significant rise in social media usage worldwide, including in Indonesia. X (formerly known as Twitter) is one of the most widely utilized social media platforms globally, serving as a public forum for communication and interaction. According to data from Dataloka, Indonesia is ranked fourth globally in terms of the number of X users, with approximately 23.76 million active users [1]. The platform facilitates the sharing of text, images, and videos, and enables users to participate in open discussions through posts and replies. While this openness fosters freedom of expression, it concurrently enables adverse behaviors, such as cyberbullying [2].

Cyberbullying is defined as intentional and repeated behaviors aimed at causing harm to individuals through electronic or digital media. In such cases, the identity of the perpetrator may either be revealed or remain undisclosed [3]. Previous studies have identified seven distinct categories of cyberbullying: harassment, exclusion, flaming, cyberstalking, outing, impersonation, and denigration [4]. Exposure to cyberbullying has been closely linked to negative mental health outcomes among adolescents and young adults, including heightened risks of depression and anxiety [5].

The incidence of cyberbullying in Indonesia has been reported to increase on an annual basis. In 2021, a total of 1,283 cases of cyberbullying were documented [6]. Data from the Indonesian Child Protection Commission (KPAI) reveal that in 2023, over 2,000 incidents pertaining to digital ethics violations and cyberbullying among students were documented, with a consistent annual increase [7]. By 2024, reports identifying children as victims of cyberbullying were ranked third within the Special Child Protection cluster, highlighting the critical nature of this issue within the digital environment [8].

Given the rising occurrence of cyberbullying on social media platforms in Indonesia, there is a critical demand for accurate automated systems to detect cyberbullying tweets. This study presents a two-stage model for detecting cyberbullying, employing IndoBERTweet [9]. In the initial phase, textual data derived from social media posts are collected and undergo preprocessing through natural language processing techniques to cleanse and normalize the text. Subsequently, a binary classification model is utilized to distinguish between content that constitutes bullying and that which does not. In the subsequent phase, content identified as bullying is further classified into three specific categories of cyberbullying: harassment, denigration, and exclusion. The datasets in both stages are divided into 70% training data, 15% validation data, and 15% testing data through stratified splitting to maintain class distribution.

This study aims to address the increasing incidence of cyberbullying on social media platforms by

improving the accuracy of automated detection systems. The proposed model, utilizing a two-stage classification approach based on IndoBERTweet, is expected to enhance the detection of cyberbullying content and its specific categories. This advancement aims to enhance efforts to protect the mental health of users, with a particular focus on vulnerable groups such as adolescents and women.

II. Related Work

The study conducted by [10] investigates the detection of cyberbullying on social media platforms, specifically X (Twitter), through the application of Machine Learning (ML) algorithms. The study employs a dataset consisting of 39,870 Twitter posts and comments, which are classified into five categories: religion-based, age-based, gender-based, ethnicity-based cyberbullying, and non-cyberbullying. Several machine learning algorithms, including Random Forest [11], Support Vector Machine (SVM) [12], Logistic Regression [13], Naïve Bayes, and K-Nearest Neighbor [14], were evaluated and compared. The experimental results indicate that the Random Forest classifier demonstrated the highest performance, achieving an accuracy of 94%. This was followed by the Support Vector Machine (SVM) and Logistic Regression models, while the K-Nearest Neighbor algorithm exhibited the lowest performance. Although this approach demonstrates relatively high accuracy, it relies on traditional NLP-based feature extraction methods and does not incorporate contextual representations from transformer-based language models. This indicates potential for further enhancement through the application of more advanced deep learning techniques.

In a similar vein, Perera and Fernando [15] conducted research on the identification of cyberbullying on social media platforms through the application of supervised machine learning techniques. The research employed three classification algorithms—Support Vector Machine (SVM), Naïve Bayes, and Logistic Regression—to systematically identify and categorize types of cyberbullying, with a particular emphasis on racial, sexual, and physical bullying. Although the system demonstrated reasonably good classification accuracy, it relied on old machine learning models. The authors advocate for further research that incorporates additional classifiers, integrates more recent natural language processing (NLP) models, and applies deep learning methodologies.

Mufva et al. [16] conducted a study examining deep learning techniques for the detection of hate speech on Indonesian social media platforms, specifically X (Twitter), employing contextual word embeddings. The study conducted a comparative analysis of various neural network architectures, including LSTM [17], BiLSTM [18], CNN [19], CNN-LSTM, and CNN-BiLSTM, with IndoBERTweet utilized as the embedding layer. The results indicated that the CNN-BiLSTM model demonstrated superior performance, achieving an accuracy of 86.75% and an F1-score of 84.39%. The CNN-LSTM model closely followed, achieving an accuracy of 86.28% and an F1-score of 83.50%. Furthermore, the CNN model independently outperformed the standalone LSTM and BiLSTM models, highlighting its effectiveness in identifying significant local patterns within brief social media texts.

Furthermore, Pekandi et al. [20] conducted a comparative analysis of three models for detecting fake news: an ensemble model, CNN-LSTM, and IndoBERT. The results demonstrated that IndoBERT attained the highest accuracy, registering at 92.24%, followed by the ensemble model with an accuracy of 91.18%. In contrast, the CNN-LSTM model exhibited the lowest F1-score, recorded at 73.53%. The findings highlight the effectiveness of transformer-based models in capturing contextual nuances within Indonesian-language text. Furthermore, [21] investigated the detection of clickbait utilizing IndoBERT and RoBERTa [22] across three dataset configurations: imbalanced, balanced, and augmented. Their experiments revealed that IndoBERT achieved superior performance, attaining an F1-score of up to 95% when trained on a balanced dataset. This underscores the importance of both model architecture and data distribution in text classification tasks.

The findings of this study suggest that the utilization of BERT-based models, such as IndoBERT, can produce superior outcomes. Nevertheless, existing research predominantly employs a single-stage classifier. As a result, when the dataset is imbalanced between instances of cyberbullying and non-cyberbullying, which is a frequent occurrence, it may lead to high accuracy but suboptimal recall and precision. Therefore, we propose a two-stage classification approach to detect cyberbullying tweets.

III. Methodology

A. Dataset

The dataset employed in this study was acquired through a web scraping process from the social media platform X (formerly known as Twitter). To ensure comprehensive coverage of relevant cyberbullying behaviors, the data collection was guided by keyword queries representing three predefined categories: harassment, denigration, and exclusion. The keywords were manually derived from the definitions of harassment, denigration, and exclusion in [4]. Keywords associated with harassment included terms such as "idiot," "stupid," "shut up," "loser," and "hate you." Denigration was represented by terms such as "fake," "liar," "fraud," "pathetic," and "disgusting." Exclusion was captured using expressions such as "you

are not welcome," "go away," "nobody likes you," "leave this group," and "you do not belong here." This targeted approach facilitated the retrieval of posts that reflect diverse manifestations of cyberbullying in online interactions.

The authors annotated the collected data to provide high quality labels for model training and evaluation. The annotation process was conducted in two stages. First, each post was assigned a binary label, where 1 indicated the presence of cyberbullying content and 0 indicated its absence. Second, posts identified as containing cyberbullying content were further classified into one of three categories: harassment, denigration, or exclusion, encoded as 0, 1, and 2, respectively. The annotation was performed independently by all four authors. A label was accepted when at least three of the four annotators agreed on the assigned class. Posts that did not achieve this level of agreement were excluded from the dataset to minimize the risk of incorrect labeling. The annotated dataset was subsequently divided into training, validation, and testing subsets with proportions of 70%, 15%, and 15%, respectively.

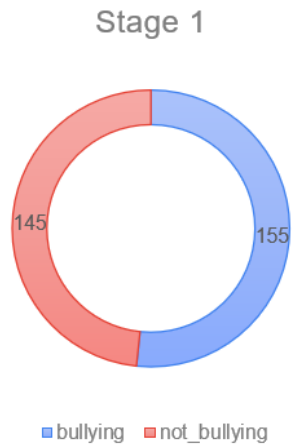

Fig 1. Distribution of Class in Stage 1 Classification

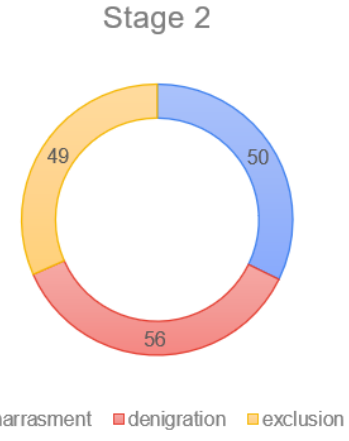

Fig 2. Distribution of Class in Stage 2 Classification

B. Data Pre-Processing

Data pre-processing entails transforming unstructured social media text into a structured format appropriate for model training. Social media data often contain noise, including informal language, redundant patterns, and non-textual elements, which can negatively affect model performance. To address these challenges, a series of Natural Language Processing techniques were employed to standardize and clean the dataset [23].

The preprocessing pipeline comprises multiple stages. Initially, all text is converted to lowercase to ensure a consistent representation. Non-semantic information, such as URLs, user mentions, and hashtags, is excluded to remove irrelevant elements. Non-alphanumeric characters, including symbols and emojis, are also omitted. Subsequently, text normalization is performed to address exaggerated character repetitions that frequently occur in informal communication. Ultimately, redundant whitespace is removed to ensure uniformity across all samples. These measures diminish extraneous noise while preserving crucial semantic content, thereby facilitating more effective representation learning with the IndoBERTweet model. An example of the preprocessing outcomes is presented in Table I.

TABLE I. EXAMPLE OF PRE-PROCESSING STEPS

Original Text	@Jenxie17 Style yg ini keren bangettt sih #JenniexChanelss26 #chanelss26 #JENNIE #ParisFashionWeek @CHANEL @jennierubyjane 😍 www.example.com
Lowercasing	@jenxie17 style yg ini keren bangettt sih #jenniexchanelss26 #chanelss26 #jennie #parisfashionweek @chanel @jennierubyjane 😍 www.example.com
Removing URL	@jenxie17 style yg ini keren bangettt sih #jenniexchanelss26 #chanelss26 #jennie #parisfashionweek @chanel @jennierubyjane 😍
Removing Mentions	style yg ini keren bangettt sih #jenniexchanelss26 #chanelss26 #jennie #parisfashionweek 😍
Removing Hashtag	style yg ini keren bangettt sih 😍
Removing Emoticon & Symbol	style yg ini keren bangettt sih
Normalization	style yg ini keren banget sih
Trim Extra Spaces	style yg ini keren banget sih
Final Text	style yg ini keren banget sih

Following preprocessing, IndoBERTweet is utilized to derive embeddings. Each preprocessed text

is tokenized using the IndoBERTweet tokenizer, which transforms the input into token IDs and attention masks. These tokenized inputs are then processed by the IndoBERTweet model to produce contextualized embedding representations with a dimensionality of 768. These embeddings encapsulate contextual and linguistic information from Indonesian social media text and are subsequently employed as input features for the proposed two-stage classification framework.

C. Two-Stage Classification

The proposed framework utilizes a hierarchical two-stage classification strategy to effectively address both the detection and categorization of cyberbullying content (See Fig. 3). This design allows the model to initially ascertain whether a post contains cyberbullying and subsequently classify the specific type of cyberbullying behavior. The architecture is constructed using IndoBERTweet embeddings, which are employed across both stages.

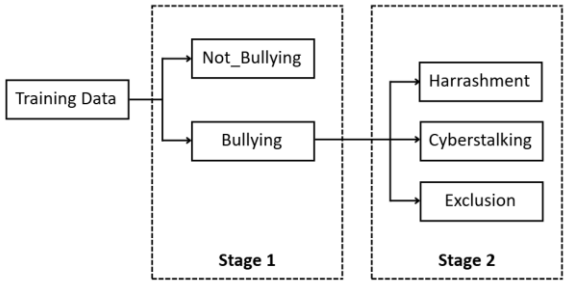

Figure 3. Two-Stage Classification

In the initial phase, the model performs a binary classification to distinguish between content that constitutes cyberbullying and that which does not. Let x denote a preprocessed input text, and let h represent the corresponding 768-dimensional IndoBERTweet embedding. The binary classifier maps 768-dimensional IndoBERTweet embedding h into a probability distribution over two classes using a fully connected layer followed by a softmax function:

$$\hat{y}_1 = softmax(W_1 h + b_1) \quad (1)$$

where W_1 and b_1 are learnable parameters. The model is optimized using binary cross-entropy loss:

$$\mathcal{L}_1 = -\big(y\ log(\hat{y}_1) + (1-y)\ log(1-\hat{y}_1)\big) \quad (2)$$

where $y \in \{0,1\}$ denotes the ground truth label. The output of the first stage determines whether an input instance is forwarded to the second stage for fine-grained classification.

During the second phase, instances identified as cyberbullying are further classified into three distinct categories: harassment, denigration, and exclusion. This stage involves performing multi-class classificatio n on the embedding representation h. The prediction is computed as:

$$\hat{y}_2 = softmax(W_2 h + b_2) \quad (3)$$

where W_2 and b_2 are learnable parameters. The model is optimized using categorical cross-entropy loss:

$$\mathcal{L}_2 = -\sum_{k=1}^{3} y_k log(\hat{y}_{2,k}) \quad (4)$$

where y_k represents the one-hot encoded ground truth label for class k.

Both stages are trained utilizing the Adam optimizer with a learning rate of 2×10−5. The batch size is set to 16, and training is conducted for a maximum of 100 epochs, incorporating early stopping with a patience of 10, based on validation loss. A dropout rate of 0.3 is applied to the classification layers to mitigate overfitting. Gradient clipping with a maximum norm of 1.0 is employed to stabilize the optimization process. The optimal model is selected based on the validation F1-score and subsequently evaluated on the test set.

D. Evaluation Metrics

To evaluate the effectiveness of the proposed two-stage classification framework, four standard metrics are employed: accuracy, precision, recall, and F1-score. These metrics are calculated based on the counts of true positives (TP_k), false positives (FP_k), false negatives (FN_k) for each class k.

Accuracy measures the overall percentage of instances correctly classified among all predictions.

$$Accuracy = \frac{\sum_{k=1}^{K} TP_k}{N} \quad (5)$$

where K is the number of classes and N is the number of samples.

Precision quantifies the proportion of true positive predictions relative to the total number of positive predictions made.

$$Precision = \frac{TP_k}{TP_k + FP_k} \quad (6)$$

Recall evaluates the model's ability to accurately identify all true positive instances.

$$Recall = \frac{TP_k}{TP_k + FN_k} \quad (7)$$

The F1-score, which represents the harmonic mean of precision and recall, provides a comprehensive assessment of model performance.

$$F1_k = \frac{2 \cdot Precision_k \cdot Recall_k}{Precision_k + Recall_k} \quad (8)$$

IV. Result and Discussion

A. Hierarchical Two Stage BERT Model

In the initial phase, the model is utilized to categorize social media posts into two distinct groups:

cyberbullying and non-cyberbullying. The assessment is performed on a test set comprising 45 samples, with a distribution of 22 non-cyberbullying posts and 23 cyberbullying posts.

TABLE II. EVALUATION METRICS ON STAGE 1

	Precision	Recall	F1-score	Support
Not Bullying	1.00	1.00	1.00	22
Bullying	1.00	1.00	1.00	23
Accuracy			1.00	45
Macro Avg	1.00	1.00	1.00	45
Weighted Avg	1.00	1.00	1.00	45

According to the confusion matrix, the model accurately classifies all test samples. Specifically, it correctly predicts 22 non-cyberbullying posts as non-cyberbullying and accurately identifies 23 cyberbullying posts as cyberbullying. No misclassifications are present, indicating the absence of both false positives and false negatives. Consequently, the model achieves an accuracy of 100%, with precision, recall, and F1-score values of 1.00 for both classes. These findings demonstrate that the Stage 1 model possesses a robust capability in distinguishing between cyberbullying and non-cyberbullying content.

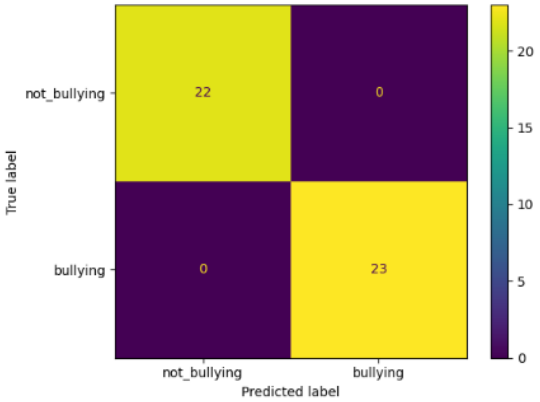

Fig 4. Confusion Matrix of Stage 1

The second stage of the proposed framework is dedicated to categorizing posts identified as cyberbullying in Stage 1 into three distinct categories: harassment, denigration, and exclusion. This stage functions conditionally, meaning that only posts predicted as cyberbullying are forwarded to the second classifier. The evaluation of Stage 2 is performed on 24 cyberbullying instances derived from the test set. The classification results indicate that the proposed model achieves an overall accuracy of 92%, demonstrating strong efficacy in differentiating between various types of cyberbullying.

TABLE III. EVALUATION METRICS ON STAGE 2

	Precision	Recall	F1-score	Support
Harrasment	0.80	1.00	0.89	4
Denigration	1.00	1.00	1.00	12
Exclusion	1.00	0.86	0.92	7
Accuracy			0.96	23
Macro Avg	0.93	0.95	0.94	23
Weighted Avg	0.97	0.96	0.96	23

According to the classification report, the harassment category exhibits robust performance, with a precision of 0.88, recall of 1.00, and an F1-score of 0.93. This indicates that the model is highly effective in identifying direct and explicit abusive expressions, with no instances of harassment overlooked during evaluation. The denigration category achieves a precision of 0.89, recall of 0.89, and an F1-score of 0.89, suggesting a balanced performance in detecting derogatory or defamatory content. Nevertheless, certain instances may still be misclassified as other forms of bullying due to overlapping linguistic features across categories. The exclusion category exhibits a precision of 1.00, a recall of 0.88, and an F1-score of 0.93, suggesting that while predictions for exclusion-related expressions are highly precise, a small number of exclusion instances remain undetected. Overall, the model attains an accuracy of 92%, with both macro-average and weighted-average F1-scores of 0.92, indicating robust and consistent performance in classifying various types of cyberbullying in the second stage.

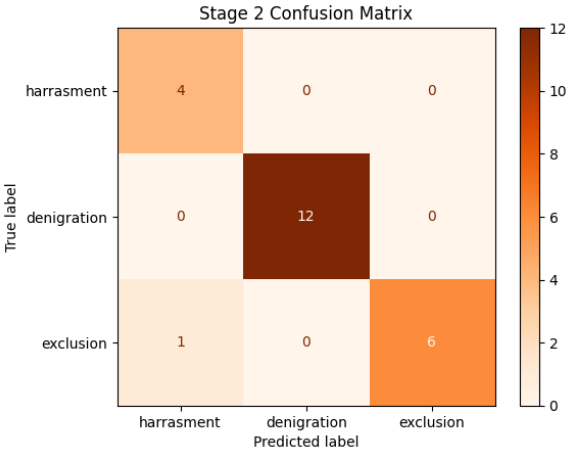

Fig 5. Confusion Matrix of Stage 2

Analysis of the confusion matrix provides additional insight into the classification behavior of the Stage 2 model. As shown in Fig. X, all harassment and denigration instances were correctly classified, indicating that these categories possess distinctive linguistic characteristics that can be effectively captured by the model. The only misclassification occurred when an exclusion instance was predicted as harassment. This confusion may arise because exclusion related messages can sometimes contain direct negative expressions or rejection statements that resemble explicit harassment. In contrast, denigration achieved perfect classification performance, suggesting

that statements targeting an individual's reputation or character exhibit more distinguishable semantic patterns. Overall, the confusion matrix indicates that the proposed model effectively separates the three cyberbullying categories, with classification errors occurring only in cases where category boundaries are linguistically subtle.

To further assess the practical applicability of the proposed two-stage framework, an inference experiment was conducted on previously unseen social media posts that were not part of the training, validation, or testing datasets. This experiment aims to simulate a real-world deployment scenario in which the model processes new incoming data. Each input text was initially analyzed by the Stage 1 classifier to ascertain whether it contains cyberbullying content. Only posts identified as cyberbullying were subsequently forwarded to the Stage 2 classifier for fine-grained categorization. This hierarchical inference procedure ensures that cyberbullying type classification is performed only when relevant.

TABLE IV. INFERENTIAL TESTING

New Data	Result: Stage 1	Result: Stage 2
Gobl*k bgtt gini aja gapaham	bullying	harrasment
Hmmm menarik sih idenya	not_bullying	none
Gausah diajak deh ganggu bgtt wkwkkw	bullying	exclusion
Tukang kayu pembohong rakyat	bullying	denigration

The findings of this experiment indicate that the proposed framework effectively distinguishes between cyberbullying and non-cyberbullying content in previously unseen data. Moreover, for posts identified as cyberbullying, the Stage 2 model accurately assigns relevant cyberbullying categories, such as harassment, denigration, and exclusion. These qualitative results suggest that the model generalizes effectively beyond the evaluation dataset and maintains consistent performance during real-world inference.

B. Model Comparison

To evaluate the effectiveness of the proposed framework, its performance was compared with several traditional machine learning models, including Support Vector Machine (SVM), Decision Tree, Random Forest, and XGBoost. Tables V and VI present the results for Stage 1 binary classification and Stage 2 multi class classification, respectively.

For Stage 1, IndoBERT achieved the best overall performance, demonstrating its strong capability in distinguishing bullying from non bullying content. Among the traditional machine learning models, Logistic Regression performed the best, while Random Forest also achieved competitive results. In contrast, SVM, Decision Tree, and XGBoost showed noticeably lower performance, indicating greater difficulty in capturing the contextual characteristics of cyberbullying language.

TABLE V. STAGE 1 RESULT COMPARISSON

Model	Accuracy	Macro Precision	Macro Recall	Macro F1
SVM	0.73	0.83	0.71	0.70
DT	0.60	0.79	0.57	0.49
RF	0.82	0.88	0.81	0.81
XGB	0.60	0.79	0.57	0.49
Our	1.00	1.00	1.00	1.00

For Stage 2, Logistic Regression and IndoBERT achieved comparable performance and consistently outperformed the other models in classifying bullying categories. Random Forest produced moderate results, whereas SVM, Decision Tree, and XGBoost struggled to differentiate certain bullying types. These findings suggest that models capable of learning richer semantic representations are more effective for both binary cyberbullying detection and fine grained bullying type classification.

TABLE VI. STAGE 2 RESULT COMPARISSON

Model	Accuracy	Macro Precision	Macro Recall	Macro F1
SVM	0.54	0.43	0.57	0.48
Decision Tree	0.42	0.43	0.48	0.36
Random Forest	0.79	0.83	0.81	0.77
XGBoost	0.62	0.80	0.67	0.60
Hierarchical Two Stage BERT	0.96	0.93	0.95	0.94

C. Ablation Study

The ablation study was conducted to evaluate the effectiveness of the proposed hierarchical classification framework. Instead of using the two stage architecture, a direct multi class classification approach was implemented in which the model directly predicted four classes: not bullying, harassment, denigration, and exclusion. The performance comparison between the direct multi class model and the proposed hierarchical framework is presented in Table VII.

TABLE VII. PERFORMANCE OF DIRECT MULTI CLASS CLASSIFICATION MODEL

Model	Accuracy	Macro Precision	Macro Recall	Macro F1
Direct Multi Class IndoBERT	0.84	0.87	0.84	0.82
Hierarchical Two Stage BERT	0.96	0.93	0.95	0.94

As shown in Table VII, the hierarchical two stage BERT model consistently outperformed the direct multi class approach across all evaluation metrics. The proposed framework achieved an accuracy of 0.960 and a macro F1 score of 0.940, compared with 0.840 and 0.820, respectively, for the direct multi class model. Similar improvements were observed in macro precision and macro recall, indicating that the hierarchical framework provides more balanced classification performance across all classes.

These results demonstrate the benefit of decomposing the cyberbullying classification task into two sequential stages. By first identifying bullying content and then classifying the bullying type, the model can focus on more specialized objectives at each stage. This design reduces the complexity of the overall task and enables the model to learn more discriminative representations, resulting in superior performance compared with directly predicting all classes in a single step.

V. Limitations and Future Work

This study utilized a dataset of 300 tweets collected from X (formerly Twitter). During data collection, it was observed that the frequency of cyberbullying categories varied considerably. While some categories were relatively common, others appeared much less frequently, making it challenging to obtain a balanced dataset across all classes. As a result, the study was conducted using a limited number of samples while maintaining a reasonable distribution among the selected categories.

Future research may expand the dataset by employing a broader set of cyberbullying related keywords to capture more diverse linguistic expressions and contexts. Although a larger keyword set would require additional filtering and annotation efforts, it could increase data diversity and improve model generalization. In addition, future studies may explore larger datasets collected over longer time periods and from multiple social media platforms.

Furthermore, this study focused on three cyberbullying categories: harassment, denigration, and exclusion. Future work may extend the proposed framework to include other forms of cyberbullying, such as flaming, cyberstalking, outing, impersonation, and other abusive online behaviors. Such an extension would provide a more comprehensive cyberbullying detection framework and further evaluate the effectiveness of the proposed hierarchical classification approach.

VI. Conclusion

This study presents a two-stage framework for the detection of cyberbullying on Indonesian social media platforms, employing IndoBERTweet. The first stage involves a binary classification to distinguish between bullying and non-bullying content. Subsequently, the second stage categorizes instances of bullying into harassment/flaming, denigration, and exclusion. This hierarchical approach enables a more precise modeling of the diverse and context-dependent manifestations of cyberbullying.

Experimental results demonstrate that the proposed method consistently exhibits robust performance across both stages, characterized by high precision and recall. In contrast to previous studies that rely on traditional machine learning and handcrafted features, the utilization of a pretrained transformer model enhances the system's capacity to effectively capture contextual and implicit abusive expressions. Overall, the findings suggest that integrating IndoBERTweet with a two-stage classification strategy constitutes an effective approach for detecting cyberbullying in Indonesian social media.

References

[1] Nouvan, “Countries with the Most X (Twitter) Users in July 2025, Indonesia in Fourth Place,” https://dataloka.id/humaniora/4627/negara-dengan-pengguna-x-twitter-terbanyak-juli-2025-indonesia-urutan-keempat/.

[2] Romy Maranatha Ginting and Muhammad Arif Sahlepi, “The Impact Of Cyberbullying On Adolescents On Social Media,” *International Journal of Sociology and Law*, vol. 1, no. 2, pp. 95–105, May 2024, doi: 10.62951/ijsl.v1i2.53.

[3] A.-L. Camerini, L. Marciano, A. Carrara, and P. J. Schulz, “Cyberbullying perpetration and victimization among children and adolescents: A systematic review of longitudinal studies,” *Telematics and Informatics*, vol. 49, p. 101362, Jun. 2020, doi: 10.1016/j.tele.2020.101362.

[4] N. Agustiningsih, A. Yusuf, and Ahsan, “Types of Cyberbullying Experienced by Adolescents,” *Malaysian Journal of Medicine and Health Sciences"*, vol. 19, pp. 99–103, 2023.

[5] S. M. B. Bottino, C. M. C. Bottino, C. G. Regina, A. V. L. Correia, and W. S. Ribeiro, “Cyberbullying and adolescent mental health: systematic review,” *Cad. Saude Publica*, vol. 31, no. 3, pp. 463–475, Mar. 2015, doi: 10.1590/0102-311x00036114.

[6] K. Hardiyanti and Y. Indawati, “PERLINDUNGAN BAGI ANAK KORBAN CYBERBULLYING: STUDI DI KOMISI PERLINDUNGAN ANAK INDONESIA DAERAH (KPAID) JAWA TIMUR,” *SIBATIK JOURNAL: Jurnal Ilmiah Bidang Sosial, Ekonomi, Budaya, Teknologi, dan Pendidikan*, vol.

2, no. 4, pp. 1179–1198, Mar. 2023, doi: 10.54443/sibatik.v2i4.763.

[7] Y. S. Yusuf, "PENDIDIKAN DI ERA MEDIA SOSIAL SINERGI PENDIDIKAN DAN PEMBINAAN MORAL PADA REMAJA," *Akhlak: Journal of Education Behavior and Religious Ethics*, vol. 1, no. 2, Jul. 2025, doi: 10.30998/jebg.v1i2.4355.

[8] P. Ayu Dhana Reswar and S. Susiana, "Pelindungan Anak dari Ancaman Kekerasan di Dunia Digital," Feb. 2025.

[9] 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.

[10] F. R. Sayed, E. H. Elnashar, and F. A. Omara, "Cyberbullying detection in social media using natural language processing," *Sci. Afr.*, vol. 28, p. e02713, Jun. 2025, doi: 10.1016/j.sciaf.2025.e02713.

[11] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[12] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

[13] D. R. Cox, "The Regression Analysis of Binary Sequences," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 20, no. 2, pp. 215–242, 1958, [Online]. Available: http://www.jstor.org/stable/2983890

[14] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967, doi: 10.1109/TIT.1967.1053964.

[15] A. Perera and P. Fernando, "Cyberbullying Detection System on Social Media Using Supervised Machine Learning," *Procedia Comput. Sci.*, vol. 239, pp. 506–516, 2024, doi: 10.1016/j.procs.2024.06.200.

[16] P. A. Mufva, K. H. Chandra, K. F. Aji, I. A. Iswanto, and S. Joddy, "Performance comparison of deep learning approaches for Indonesian twitter hate speech detection using IndoBERTweet embedding," *Procedia Comput. Sci.*, vol. 269, pp. 1663–1671, 2025, doi: 10.1016/j.procs.2025.09.109.

[17] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[18] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997, doi: 10.1109/78.650093.

[19] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998, doi: 10.1109/5.726791.

[20] L. A. Pekandi, R. G. Widjaja, A. Ananta, J. Harefa, and K. Jingga, "Evaluating IndoBERT for Indonesian Hoax News Detection: A Comparative Study with Ensemble and CNN-LSTM Models," *Procedia Comput. Sci.*, vol. 269, pp. 1625–1633, 2025, doi: 10.1016/j.procs.2025.09.105.

[21] M. E. Syahputra, A. P. Kemala, and D. Ramdhan, "Clickbait Detection in Indonesia Headline News Using Indobert and Roberta," *Jurnal Riset Informatika*, vol. 5, no. 3, pp. 425–430, Jun. 2023, doi: 10.34288/jri.v5i4.237.

[22] Z. Liu, W. Lin, Y. Shi, and J. Zhao, "A Robustly Optimized BERT Pre-training Approach with Post-training," 2021, pp. 471–484. doi: 10.1007/978-3-030-84186-7_31.

[23] M. M. Islam, M. A. Uddin, L. Islam, A. Akter, S. Sharmin, and U. K. Acharjee, "Cyberbullying Detection on Social Networks Using Machine Learning Approaches," in *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, IEEE, Dec. 2020, pp. 1–6. doi: 10.1109/CSDE50874.2020.9411601.